A. An overview of commercial sensors for TCM

A.1 Introduction

Although many researchers have considered the complex problem of TCM in many different approaches, a need still exists for adequate sensors for TCM. Many authors state that force and vibration sensors yield the best results in a TCMS, although a number of different approaches proposed by others also yielded satisfactory results. This includes methods like ultrasonic sensing, laser focus methods, temperature measurements and optical methods and Acoustic Emission (AE) monitoring. However, most of these sensors seem to be operational only in laboratory setups. In the case of AE, many authors have had success in establishing a reliable AE methodology for TCM, overcoming the sensor problems. The use of AE for TCM has also found its way into industry. However, there is still a lack of understanding / interpreting the AE signals generated during machining. As was shown in Chapter 4, the trends in the AE signal due to tool wear is not very consistent, and it will not always be possible to utilise AE for any TCM problem.

As far as the other methods are concerned, they are not without their respective shortcomings. For instance, the presence of cutting fluid and chips breaking away from the workpiece have a very large impact on the performance of vision-based systems. Although tool wear and cutting temperature are closely related, it will not be wise to design a TCMS based only on this feature. Also, the presence of coolant will make this approach impossible. Figure A.1 shows a sensor selection table for TCM suggested by Montronix [245]. The table suggests force measurement for tool wear monitoring. However, the performance of any sensor is always process dependent, and therefore this table must not be used as the ultimate sensor reference guide.

The sensor problem in the case of vibration and cutting force still require further developments. Dimla [41] state that the monitoring of both vibration and cutting force is of indispensable value to successful TCM. In the case of the cutting force, the static and dynamic component of the force is of importance. The dynamic behaviour of the cutting process is embodied in the vibration and the dynamic cutting force. In industry only the static cutting force is used for TCM, but this is mainly due to the fact that the physical measurement, analysis and interpretation of the dynamic component is more complex.

Due to nature of the machining environment, very robust sensors are required to monitor tool vibrations and cutting forces. Usually, a need exists to monitor as close to the machining process as possible to minimise the effect of external influences on the sensor performance. However, cutting fluid and chips breaking away from the workpiece can damage a sensor beyond repair.

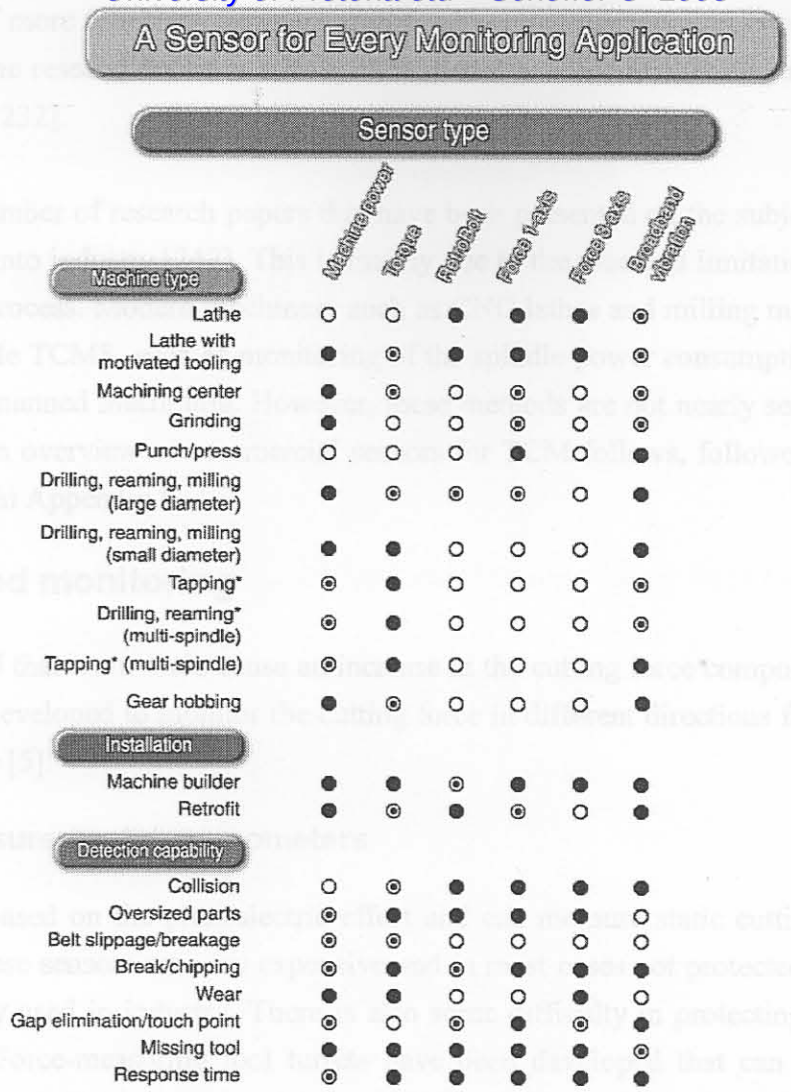


Figure A.1: Sensors for TCM, as suggested by Montronix

Key:

- ● Common, recommended application
- ⊙ Acceptable depending on application, but may not be preferred solution.
- ○ Unusual application - generally not recommended.

Also, in order to develop a TCMS for use in industry, sensors must have a realistic lifetime of at least a few years (also refer to sensor requirements listed in Chapter 3). Many modern sensors comply with some of these requirements, but a vibration or force sensor that complies with all of these requirements does not exist yet. This gives rise to another issue: Should a new sensor be developed specifically for TCMS, or should the designer of the TCMS attempt to get best results with available sensors?

The difficulty with the first question is the fact that there are so many different processes. Thus, a process specific sensor will probably not justify the costs involved. However, in the case of cutting force monitoring, Kistler [246] developed a range of sensors to monitor cutting forces for particular processes like turning, milling, drilling, grinding and even micro-machining. In the case of accelerometers, a need still exists for more robust and smaller sensors that can be integrated into the tool holder. A few sensors developed by Kistler and PCB are available that can be used for TCM applica-

tions. The design of more reliable sensors to continuously monitor machining processes is a topic for future research. Some researchers have already identified this niche market and are currently working in this area [32,103, 232].

Despite the large number of research papers that have been presented on the subject of TCM, very little found their way into industry [247]. This is mainly due to the practical limitations to obtain reliable sensor data on the process. Modern machinery such as CNC lathes and milling machines are equipped with relatively simple TCMS, such as monitoring of the spindle power consumption for detecting tool breakage during unmanned machining. However, these methods are not nearly sensitive enough to estimate tool wear. An overview of commercial sensors for TCM follows, followed by an overview of commercial TCMS in Appendix B.

A.2 Force-based monitoring

It is well recognized that worn tools cause an increase in the cutting force components. Many types of sensors have been developed to monitor the cutting force in different directions for a number of processes. These include [5]:

A.2.1 Direct measurement dynamometers

These sensors are based on the piezoelectric effect and can measure static cutting forces very accurately. However, these sensors are very expensive and in most cases not protected from overload, and therefore not widely used in industry. There is also some difficulty in protecting the sensors against cutting lubricants. Force-measuring tool turrets have been developed that can measure three force components, but these still have a very high cost. Kistler is the leader in supplying force dynamometers for machining processes.

Most academics active in the area of TCM have used the Kistler dynamometers for cutting force measurements. These are usually two or three axis quartz dynamometers. Quartz is a natural crystalline material with extreme ruggedness and ultrastability. The piezoelectric force measuring system differs fundamentally from most other methods. The forces acting on the quartz elements are directly converted into proportional electrical signals, and the resulting displacement amounts to a few thousands of a millimetre. Dynamometers are very accurate and can measure quasi-static forces. Their frequency range is unfortunately fairly low, allowing dynamic force measurements to about 1kHz. Many different types of dynamometers exist, developed for specific machining operations.

The Kistler Three Component Dynamometer 9121, shown in Figure A.2, is a quartz three-component dynamometer for measuring the three orthogonal components of a force. The force components are measured practically without any displacement. The sensors are fitted so that they are ground isolated. This largely excludes ground loop problems [246]. The dynamometer is corrosion-resistant and protected against penetration by spray water and cutting fluid. The control unit of the dynamometer is easy to operate and contains a power pack and a keyboard with status displays together with a connector for signal input. The output voltages are proportional to the forces occurring at the tool tip. Figure

A.3 shows a recently developed three-component dynamometer for hard turning [248].

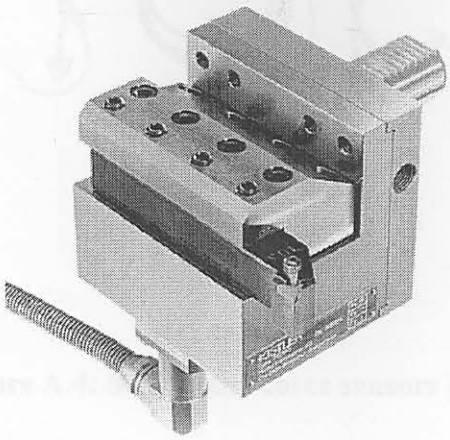


Figure A.2: Kistler Three Component Dynamometer 9121 [246]

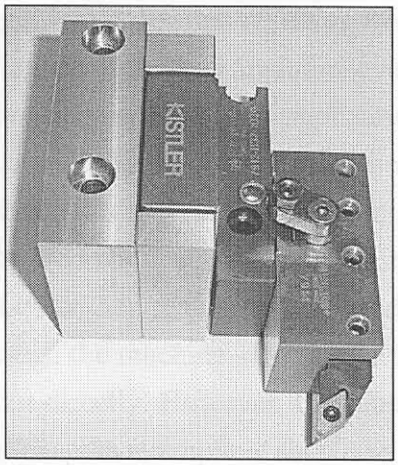


Figure A.3: Kistler 7914 dynamometer for hard turning [248]

The IntelliTool developed by Sandvik is a force sensor integrated tool holder for machining centres, developed around 1995. A spokesperson from Sandvik confirmed the existence of the IntelliTool, but stated that it is not in production anymore. Sandvik is currently not exploring further developments of the IntelliTool concept.

A.2.2 Sensors in plates and rings

Force-measuring plates can be fitted relatively easy on turning machines between the turret housing and the cross slide. These thin plates are fitted with piezoelectric force measuring sensors. This approach has some advantages, but is subject to many disturbing factors such as thermal expansion. Montronix provide a fairly wide range of force sensors for TCM, primarily used for fixed tool applications such as lathes, presses, and broaches. Typical sensors from Montronix are shown in Figure A.4.

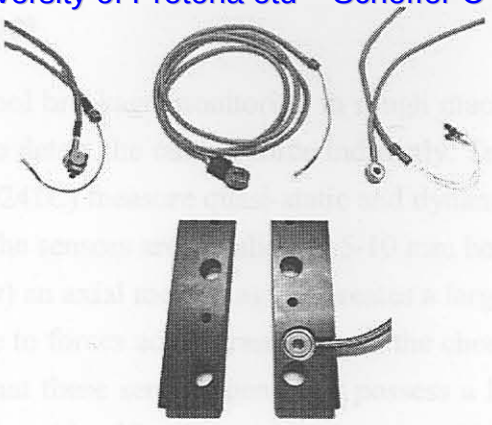


Figure A.4: Montronix force sensors [245]

The Montronix RetroBolt is a retrofit force sensor typically used on turret and tool block lathes. This single axis sensor is installed as a washer under a bolt head. The selected bolt must be in the load path. The sensor measures the small changes in bolt tension resulting from cutting forces exerted on the tool. Typical monitoring performance using the RetroBolt includes overload detection, machine protection and detection of catastrophic tool breakage in roughing applications. Examples of ring-shaped force transducers from Montronix that can be fitted on a CNC machine are shown Figure A. [245]. Kistler also offers a number of ring-shaped sensors that can be fitted in a similar way onto CNC machines. Some sensors offer the possibility for 3-component force monitoring. An advantage of using these sensors is that their frequency range is higher than dynamometers, and the cost is lower. The disadvantage is that force must be transmitted through the machine structure to the sensor, in contrast to dynamometers, which measure the cutting forces directly. However, the sensors are very robust and can be used for a wide range of machining operations. An example is shown in Figure A.6.

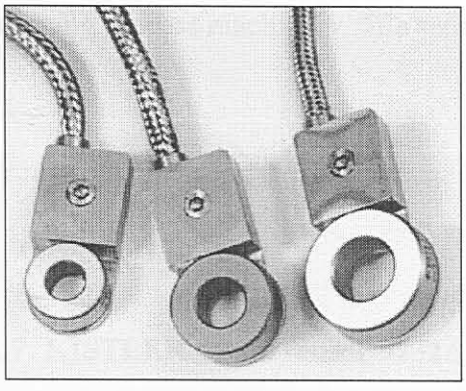


Figure A.5: Examples of ring-shaped force transducers for TCM from Montronix [245]

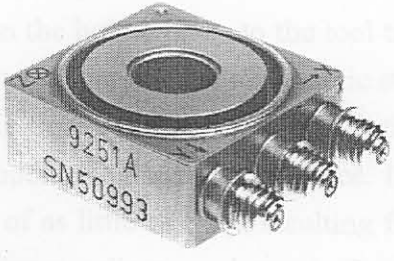


Figure A.6: Force sensor from Kistler [246]

A.2.3 Pins, extension sensors

These sensors are suitable for tool breakage monitoring in rough machining. They are fitted on force-carrying machine components to detect the cutting force indirectly. Transverse and longitudinal measuring pins (*e.g.* Kistler 9243A/9241C) measure quasi-static and dynamic strains in the structure of machine components or fixtures. The sensors are installed in 5-10 mm holes. Owing to the larger measuring range (length of pressure bar) an axial measuring pin creates a larger measuring signal than a radial pin. Both sensors are insensitive to forces acting transverse to the chosen axis. In view of the unavoidable interference and the fact that these sensors generally possess a low level of sensitivity, they are normally only suitable for breakage identification during rough machining and for press force monitoring. The identification of a suitable fitting position can only be determined experimentally, which is another disadvantage.

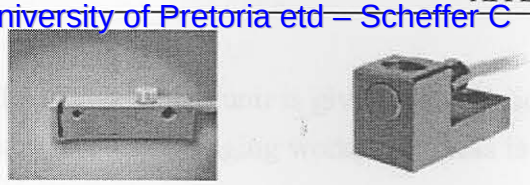
A.2.4 Measurement of displacement and strain

Non-contact sensors to measure the displacement or bending of tools can be mounted directly on the tool holder. However, these sensors are subjected to the high risk of damage and disturbances due to chips, dirt and cooling lubricant. It is relatively easy to retrofit various strain and displacement sensors. Their working principle is indirect force measurement. A quartz strain transducer can be mounted on a part of the machine where the mechanical stress is large and disturbances are low.

All force transmitting parts in machines and fixtures are deformed elastically by the forces acting on them. As a result of this deformation, a detectable displacement occurs on the surface of the force transmitting parts. Because of high machine rigidity, the elastic deformations are extremely small. A strain sensor development for measuring dynamic and quasi-static forces on stationary and moving machinery is the Kistler 9232A strain sensor. Its high sensitivity and acceleration-compensated design allows a sensor application on fast running process machinery. The sensor is shown in Figure A.7.

Figure A.7: KISTLER strain sensor 9232A [246]

The PCB dynamic strain sensor has also proved its worth in TCM applications [165]. These are very small sensors that can be adhered on the holder close to the tool tip. Another advantage of this sensor is the fact that it is extremely sensitive and can measure dynamic strain accurately in the range of 5Hz - 10 kHz. One disadvantage though is the fact that they cannot estimate static strains, which would be useful for estimating the static component of the cutting force. Induction-based distance sensors are capable of resolving displacements of as little as 5 nm resulting from forces acting on machine parts. Examples of such sensors from leading suppliers are shown in Figure A.8.



NORDMANN BDA-Kralle PROMETEC DSA 50

Figure A.8: Induction-based displacement sensors

Resistance strain gauges are widely used for stress analysis, but are not often found in TCM applications. A standard design of such a sensor consists of a 0.025 mm layer of polyimide film that covers the gauge except that portion of the tab necessary for lead attachment. The polyimide overlay provides protection of the sensing element during installation handling, and promotes better long-term stability with the foil grid protection from airborne contaminants or fingerprints. Figure A.9 shows how HBM resistance strain gages are applied to the shaft of milling tools [248,239].

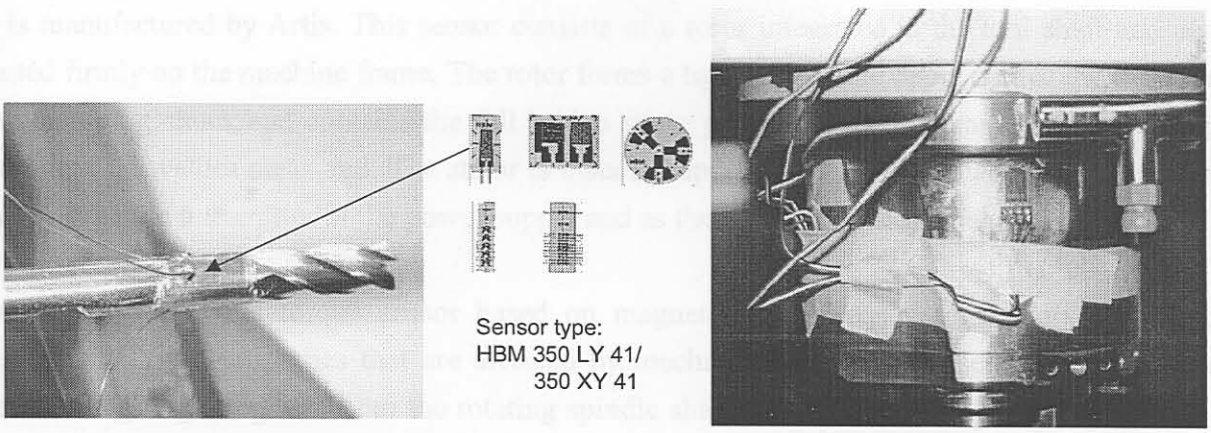


Figure A.9: HBM resistance strain gage application [248]

A.2.5 Force-measuring bearings

Bearings and bushes can be specially fitted with strain gauges in certain positions to measure cutting forces. Force-measuring bearings require a low-pass filter due to disturbances from the ball contact frequency, and therefore high frequency signal processing is not possible. The force-measuring bushes are only accepted in special cases because they reduce the rigidity of the machine.

A.2.6 Force and torque at spindles

These systems can be very complex because they have to monitor the torque of the spindle with high resolution and within the entire range of the motor. Furthermore, the signal must be transmitted on a non-contact basis. The installation of such a system is not possible on most machines due to the limited space available for sensor mounting. The Kistler two-component high-speed dynamometer type 9125A is a part of a rotating cutting force measuring system and allows measurements of both the axial feed force and torque on the rotating tool. The instrument consists of a two-component piezoelectric sensor, a built-in charge amplifier and a built-in two-channel telemetry system for wireless data transmission. The main advantages of such a machine tool integrated, rotating cutting force measuring system are:

- The force measurement is directly on the rotating tool holder, i.e. the signal is recorded close to

the tool tip.

- A constant mass of the measurement unit is given, i.e. a steady resonance frequency of the dynamometer is realised, since the changing workpiece mass is not connected to the dynamometer.
- Low and constant inertia forces are acting during the cutting process.
- Any orientation of the dynamometer is possible.

Much more accurate tool and process monitoring can be achieved by measuring mechanical torque directly instead of measuring the power consumed by the spindle motor. It is especially useful for tapping and a multi-spindle application where power monitoring is insufficient. The sensor can monitor tool wear, tool breakage, did-not-cut condition, thread depth, oversized or undersized pre-drilled holes, and damaged or missing threads on taps.

A torque tool sensor that uses strain gauges for torque measurement and non-contact signal transmission is manufactured by Artis. This sensor consists of a rotor integrated in the tool shaft and a stator mounted firmly on the machine frame. The rotor forms a tight ring in the upper part of the tool shaft or of the clamping chuck and contains the full bridge strain gages and the electronics for acquiring and transmitting the values measured. The stator is installed approximately 5 μm away from the rotor and serves both as the transmitter of the power supply and as the receiver of the signal.

Montronix introduced a torque sensor based on magneto-elastic sensing methods (Accu-Torque), which has magnetic properties that are affected by mechanical torque. The sensor includes a small torque-sensing ring integrated onto the rotating spindle shaft, and non-rotating pick-up. The ring converts mechanical shaft torque into a linearly proportional magnetic field. The pickup converts this field into a linearly proportional electrical signal and acts as a non-contact means of gathering shaft torque information.

Excellent rotating quartz four-component (Fx, Fy, Fz and Mz) dynamometers for measuring cutting forces and torque on a rotating tool spindle are also available from Kistler. However, because of their high cost they can be considered as laboratory tools only. Examples are shown in Figure A.10.

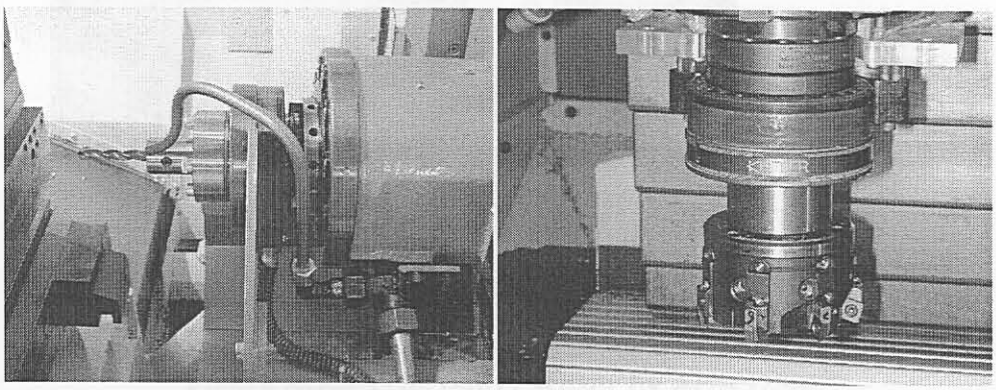


Figure A.10: KISTLER RCD 9123 sensor used in drilling and milling applications [246]

A.3 Measurement of motor current

The measurement of motor current is an easy alternative to the above-mentioned systems and can be installed without much difficulty. A wide range of sensors is available for this purpose. However, due to fluctuations in the signal due to friction, this approach is not accurate. Furthermore, tool breakage is not detected directly, but only after damage has occurred. Spindle power is also proportional to the cutting force in the primary motion, which is not always sensitive enough for TCM. The cutting process consumes only a small portion of the measured power of the spindle, which also makes monitoring difficult. However, monitoring systems based on the principle of spindle current can be successful when used with an appropriate process. Such sensors are often found in the automotive industry in drilling and tapping applications. A typical sensor for estimating the spindle power is shown in Figure A.11 [249].

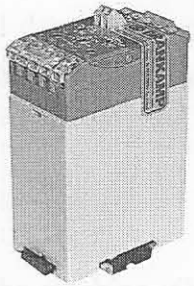


Figure A.11: Hall-effect sensor for measuring spindle power [249]

A power sensor measures the spindle or axis drive power for AC, DC or variable frequency motors (frequency range 0-200 kHz). In all commercial applications currently available, power is measured directly by means of a voltage and current measurement. The power sensor is installed directly in the electrical cabinet, and measures with 1-3 Hall effect sensors. A typical application is shown in Figure A.12.

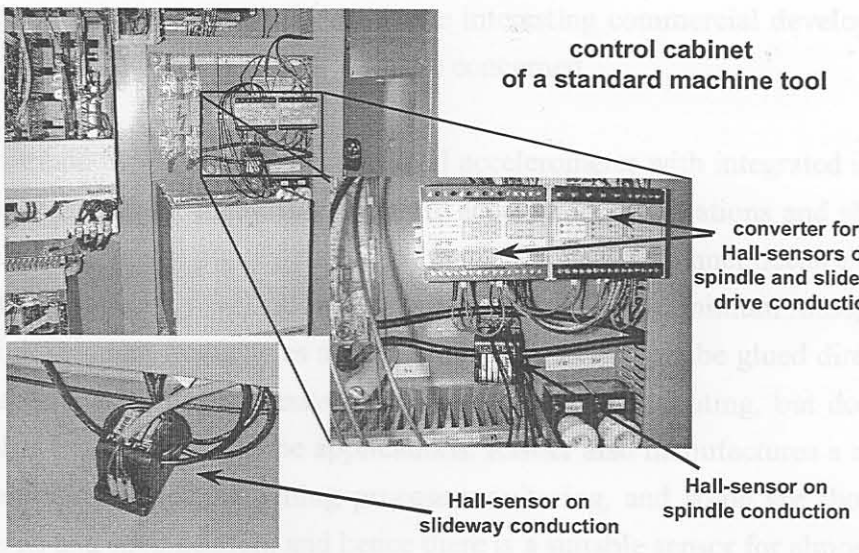


Figure A.12: True power sensor by Artis [250]

Use of three balanced Hall effect sensors eliminates large phase shift errors. When two sensors in two

phases are used, the third current quantity results from the two that are measured, making it possible to take the usual grid fluctuation between three phases into account. Effective power measurement has the advantage over simple current measurements in that the idle current, which provides no information about the motor load on the tool, is not measured. Power is linear, so that a change in motor load is a change in power. Current is not a sensitive indicator of power at low loads in three-phase motors. The sensor can provide an indication of a missing tool or tool wear in certain applications. Using an additional logarithmic signal amplifier, small tools can also be monitored. If the power-measuring curve is wavy, or displays ripples, peaks or even brief collapses (<1s), the power measuring value can be smoothed. The quality of effective power monitoring depends largely on the relationship between the cutting power and the nominal drive power of the motor. This means that small tools (such as drills with a diameter of less than 1mm) can be monitored only on tight fitting spindle drives. Owing to the inert masses, the output signal has a low-pass filter characteristic. Therefore, tool breakage is not detected directly, but only after consequential damage has occurred. In many cases, force, vibration or AE signals have to be used as complementary information to extend the capability of the monitoring system.

A.4 Acceleration measurement

Piezoelectric accelerometers can measure the machine vibration caused by oscillations of the cutting forces. It is known that vibration levels change with tool wear. Some accelerometers are inexpensive and can measure vibration levels within a very wide frequency range. For these reasons, accelerometers are often used for TCM [5,41]. Miniature accelerometers available on the market are not always robust enough to withstand cutting fluid and chips for very long. The ones that may be able to withstand the environment are either too big or cannot measure in the frequency range of interest. The sensor cables must also be watertight, and preferably inside a steel shielding. Hence, a dedicated accelerometer must be used for machining process monitoring. The development of an accelerometer integrated into a tool holder is definitely a worthwhile research topic. Such a system was recently proposed by Lago [32]. Lately, there have been some interesting commercial developments as far as accelerometers for machining process monitoring are concerned.

The Kistler Piezotran® 8694 is a miniature triaxial accelerometer with integrated impedance converter and low impedance output for measuring dynamic accelerations, vibrations and shocks. The sensor is shown Figure A.13. For power supply and signal processing each channel needs a Piezotron® coupler. The accelerometer is especially well suited for applications where minimum mass, small mounting dimensions and high resonant frequencies are essential. The sensor can be glued directly to the object or to mounting adapters. A contact adhesive is recommended for mounting, but double sided adhesive tape or wax is also acceptable for some applications. Kistler also manufactures a range of other accelerometers that are suitable for machining process monitoring, and some are shown in Figure A.14. These vary and size and specification, and hence there is a suitable sensor for almost any application.

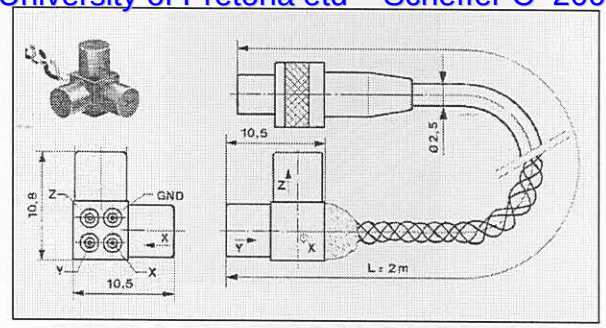


Figure A.13 KISTLER Piezotran® 8694 accelerometer [246]

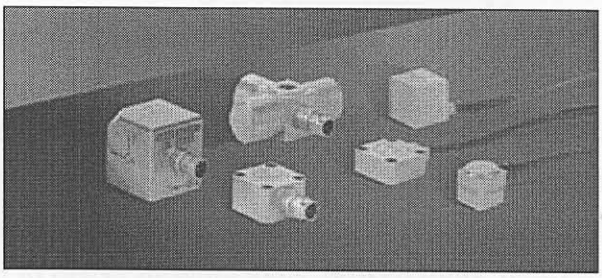


Figure A.14: Kistler three-component accelerometers for TCM [246]

The IMI 356A11 miniature triaxial ICP® accelerometer, depicted in Figure A.15, is a miniature, four-gram, hermetically sealed titanium accelerometer. The hermetic titanium case provides a rugged, reliable sensor for minimal mass loading during triaxial vibration measurements. The 10mV/g output from ceramic shear mode sensing elements provides a wide measurement range from 2 to 10 000 Hz (±5%) in the "Z" axis and 2 to 7 000 Hz (±5%) in the "X" and "Y" axes. The 10ft integral shielded cable terminates in a four-pin connector. This robust and small sensor would work well for TCM applications.

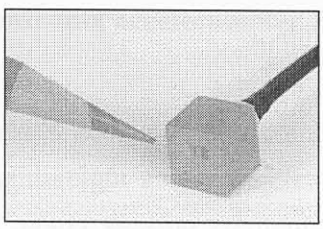


Figure A.15: IMI 356A11 Miniature Triaxial ICP® Accelerometer [251]

An accelerometer that can be used in a corrosive environment compatible with 304 Stainless Steel (SS), and survive long term submersion in water (including the cable) is ENTRAN's EGS Series. Its case is fabricated from 304SS with welded seals and its cable is sheathed with a 3.2mm diameter soft annealed SS tube, which can be easily bent or shaped for routing around corners. The sensor is shown in Figure A.16.

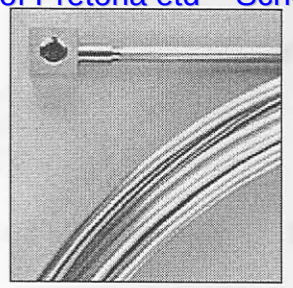


Figure A.16: Entran EGS series [252]

Montronix supplies the BV100 sensor designed specifically for application on machining centres that use a wide variety of tooling. The sensor is shown in Figure A.17. The sensor boasts an extremely broad frequency range for improved sensitivity to the condition of small and large diameter tools. It is easily to install: The 5 meter cable is rated to IP67 and NEMA 6 industrial standards and is shielded against electro-magnetic interferences, hence 100% inner and outer shielding. An inner polyurethane jacket protects against coolant damage and the stainless steel verbraid protects against hot chips and sharp corners. It can be used to detect missing tools, broken tools, out-of-tolerance parts, machine collision and severe process faults. It is also possible to monitor excessive vibration on bearings or spindles. The vibration sensor is easy to install on new or existing machines.

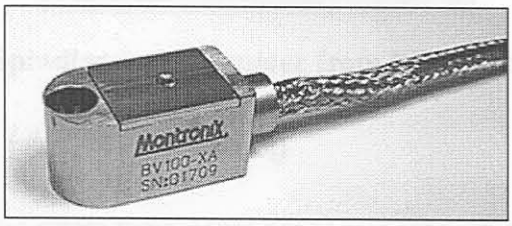


Figure A.17: Montronix accelerometer BV 100-XA [245]

The Kistler 8730A500 miniature K-Shear® accelerometer combines a reliable, easy to handle connector with an integral stud to create a lightweight accelerometer that fits into narrow areas [246]. The hermetically sealed titanium case and quartz sensing elements ensure proper mechanical protection. The sensor is shown in Figure A.18

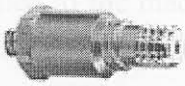


Figure A.18: Kistler 8730A500

Prometec supplies a number of sensors for TCM, optimised for use with their other TCM hardware. These sensors include: force, torque, displacement, strain, position, hydraulic pressure, power, AE, and vibration. All of these sensors are robust, and designed specifically for the purpose of TCM [253]. The force and vibration sensors from Prometec are shown in Figure A.19.

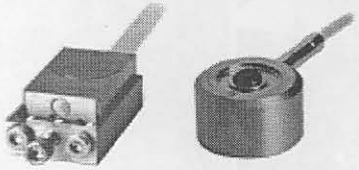


Figure A.19: Force and vibration sensors from Prometec [253]

The ICP compatible "Spindler" accelerometer from PCB, shown in Figure A.20, is especially designed for machining process monitoring. The accelerometer is provided with a stainless steel cable shielding that can withstand cutting fluids and chips. Furthermore, the sensor can swivel around its base, making it easy to install in limited space applications.

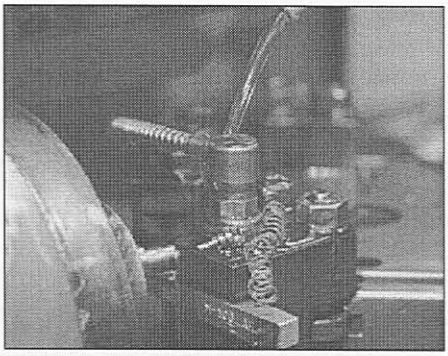


Figure A.20: Spindler accelerometer from PCB piezotronics [251]

A.5 Acoustic Emission (AE) monitoring

Cutting processes produce elastic stress waves that propagate through the machine structure. Different sources in the cutting process generate these stress waves known as Acoustic Emission (AE). AE is defined to occur in the range of 50kHz - 250kHz. An AE sensor measures the high-frequency energy signals produced by cutting processes. When a tool breaks, the sensor also measures the AE energy resulting from the fracture. An AE sensor is best suited to applications where the level of background AE is low compared to that of tool breakage. The sensor is easy to install on both new and existing machines. In combination with true power, it increases the reliability of breakage monitoring significantly. It is used especially with solid carbide tools, or very small tools on large machines and multi-spindles. Most AE sensors have to be attached to the machine tool surface (sometimes with different mounting variants, *e.g.* side, top, or bottom connection, and spring disk fixing). An example of mounting for a drilling application is shown in Figure A.21.

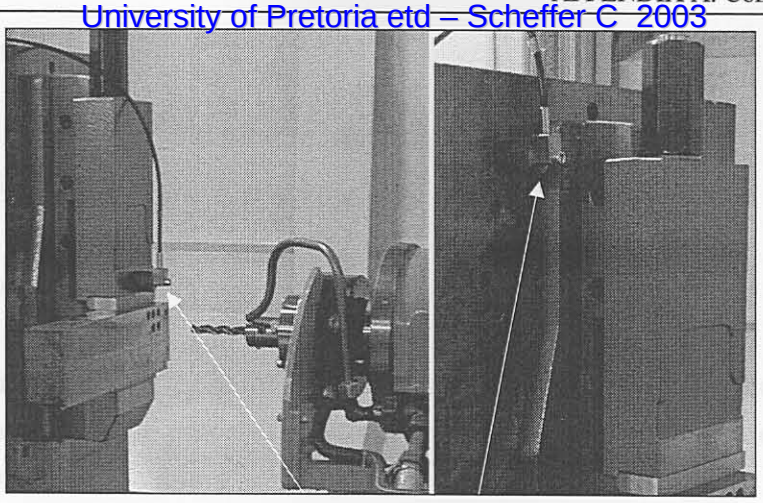


Figure A.21: Kistler AE 8152A1 sensor mounted

The primary and secondary cutting shear zones are important sources of AE. In the presence of flank wear, the interface between tools and work materials becomes an additional zone of AE generation due to intensive friction between the tool and the workpiece surface which move past one another at relative high velocities. Changes of cutting conditions also affect the behaviour of AE signals.

There are a number of ways through which AE waves are transmitted. The Prometec AEL 200 is a rotating, wireless AE sensor. It is suitable for applications where signals from moving parts have to be passed to a fixed receiver for analysis and monitoring. When applied in grinding, the sensor enables precise detection of any sparking, which enables optimisation of the depth setting. Another approach is the Prometec fluid sound sensor WAE100, an AE sensor that receives the acoustic waves via a jet of cooling lubricant, which can be connected directly to the tool (or workpiece). The sensor and an example of its mounting is shown in Figure A.22. Its advantage over the conventional AE sensors is that it measures close to the tool tip. The fluid AE sensor allows detection of high frequency stress waves from moving or rotating workpieces or components, or from materials with very rough surfaces. In some cases the acoustic waves in the coolant stream are damped so much by the air bubbles that the resulting measurement is too low. Particles of metal and dirt, on the other hand, do not cause problems.

Although announced a few years ago, a dual-mode sensor for the simultaneous measurement of AE and one to three orthogonal force components never reached the production line. The advantage of this dual-mode sensor would be to compliment the force measurement with AE.

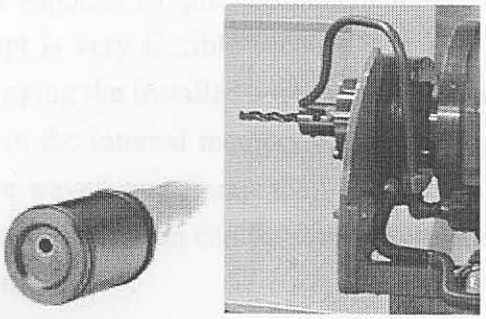


Figure A.22: Prometec WAE 100 [253]

A.6 Other methods

Other methods for indirect / direct tool wear monitoring are:

- non-contact capacitive sensors
- vision systems
- laser scatter methods
- stereo imaging

At the time of the review, none of these methods were available commercially, besides vision systems, which are available for machining process monitoring in general, but not dedicated to TCM. However, examples of these systems can be found in the literature (refer Chapter 3).

A.7 Latest developments

A.7.1 SeTac

The Sequoia Triaxial Acceleration Computer (SeTAC) is a device that includes three accelerometers on each Cartesian axis and a microprocessor for the signal analysis and data communication. SeTAC belongs to the category of smart sensors, which offer a direct connection between signal recording and processing [239,248]. Regarding the signal understanding process, the SeTAC sensor aims at integrating the filtering functionality into the sensor. The SeTac functionality is schematically depicted in Figure A.23.

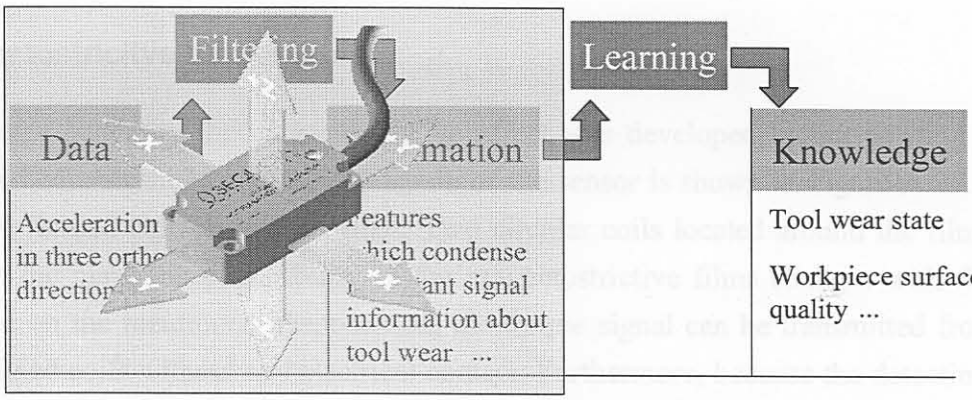


Figure A.23: SeTAC tri-axial accelerometer

The accelerometer signals are processed in the SeTAC by a built-in micro-controller that is programmable and fully customisable. It supplies simple digital output suitable for common devices, PLCs, actuators *etc*. The SeTAC concept is very flexible because it is able to monitor and evaluate a wide range of phenomena by only changing the installed software. A common Personal Computer (PC) can be used to upload the programs in the internal memory. The PC can even operate as a user interface where it is possible to display the waveform, change the filtering and set the thresholds. When the required behaviour is achieved the configuration can be saved in the memory and the SeTAC is able to operate without connection with the PC.

A.7.2 Intelligent tool coating

Various new developments are based on the notion to integrate sensors into the tool coating, *e.g.* German BMBF project IDEE or the EMO 2001 presentation of Kyocera Fineceramics Gmbh. Such a direct measurement would offer the advantage to exclude a large degree of external disturbances that are normally reflected in any indirect measured signal pattern. The tool coating has to be restructured by adding several layers of different materials without changing the quality of the coating. By integrating the sensor in the tool coating the conversion of signal data to process knowledge is extremely simplified. An example is shown in Figure A.24.

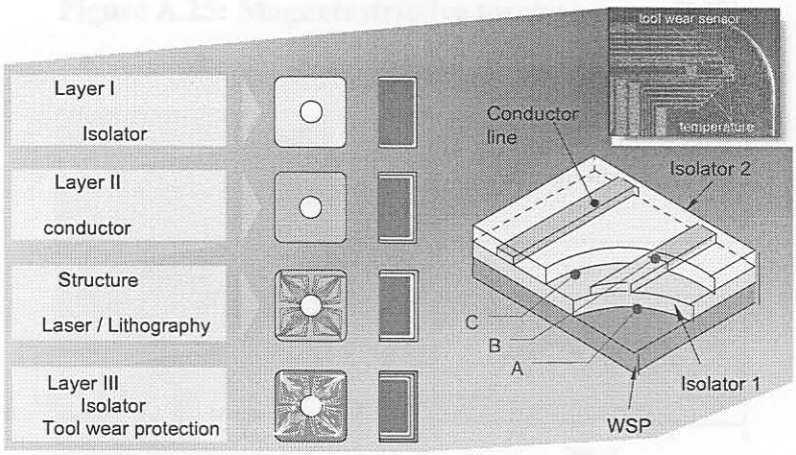


Figure A.24: Intelligent tool coating – direct tool wear and temperature measurement [254]

A.7.3 Magnetostricitve torque sensor

A torque sensor based on the magnetostrictive effect was developed by cooperation between Keio University and Mitsubishi Materials. A diagram of the sensor is shown in Figure A.25. Magnetostrictive films are formed onto the sensor shaft. Two circular coils located around the films measure the torque, since the magnetic permeability of the magnetostrictive films changes with the variation of torque. Based on the mentioned phenomenon, the torque signal can be transmitted from the rotating shaft to the fixed coils without any electrical contact. Furthermore, because the detecting elements are close to the cutting point of the tool, the magnetostrictive torque sensor provides a high quality cutting torque signal being independent of the workpiece location.

In order to simplify application of the sensor system, the torque sensor is applied to the tool holder. A prototype of the system together with example signals from a drilling application is shown in Figure A.26. The development of a magnetostrictive sensor that additionally sense a three component cutting force being applied on the cutting edge of a rotating tool is also under way.

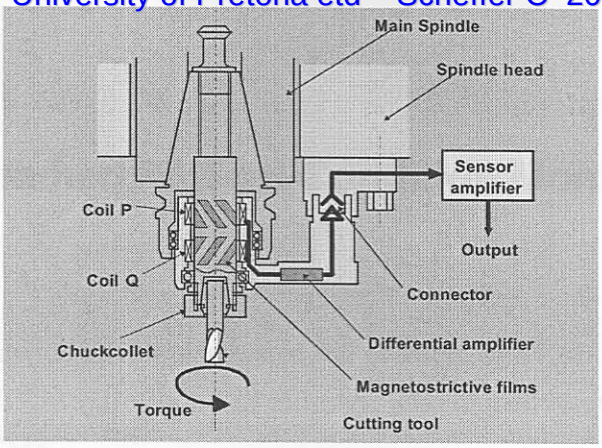


Figure A.25: Magnetostrictive torque sensor [248]

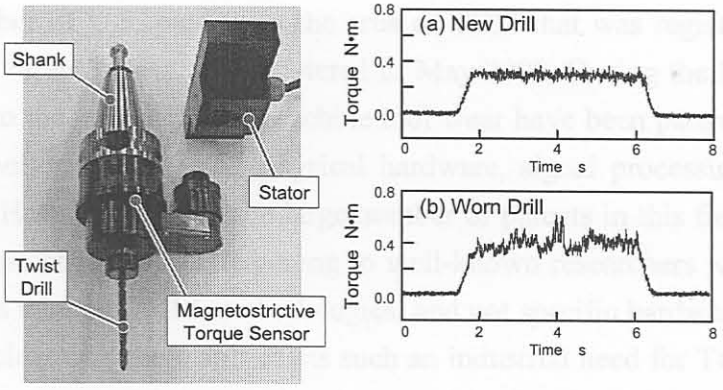


Figure A.26: Magnetostrictive torque sensor prototype [248]

B. Overview of commercial Tool Condition Monitoring Systems

B.1 Introduction

Numerous researchers worldwide have investigated the problem of TCM. However, due to many practical limitations, not many TCMS are found in industry. The objective of this appendix is to give an overview of commercial TCMS.

B.2 U.S.A. Patents

Table B.1 lists a number of U.S. patents in the area of TCM that was registered within the last ten years [255]. The most recent patent was registered in May 2000. During the last 15 years, more than 50 inventions related to the monitoring of machine tool wear have been patented in the U.S.A. These inventions include monitoring strategies, physical hardware, signal processing techniques and even mathematical models. However, despite the large number of patents in this field, only a few commercial versions exist. Most of these patents belong to well-known researchers within the field of TCM, and most of the patents relate to TCM methodologies, and not specific hardware. When looking at this list of patents, it is unclear why there still exists such an industrial need for TCMSs. The only reasonable explanation is that the patents were never developed into commercial products, either due to the costs involved or limited accuracy or applicability in industry. It is also often unclear if these methods were tested under conditions resembling a realistic industrial environment. If the systems were not tested under typical shop floor conditions, it remains a question whether they will be effective in industry.

Table B.1: U.S.A. patents related to TCM in the last 10 years

no.	*Description*
6059494	Tool bit monitoring system for machine tools **(2000)**
6055484	Tool monitor and assembly qualifier
5904457	Detecting tool wear by thermal monitoring of workpiece
5773949	Cutter fracture detecting system
5587931	Tool condition monitoring system
5542304	Magnetostrictive torque sensor magnetostrictive torque measuring apparatus and condition-monitoring apparatus for a cutting tool using the same
5414632	System and method for predicting failure in machine tool
5361308	3-D measurement of cutting tool wear
5266929	Metal cutting tool wear indicator method and system
5251144	System and method utilising a real time expert system for tool life prediction and tool wear diagnosis
5212645	Flexible real-time multi-tasking architecture for tool condition monitoring
5076102	Tool monitor **(1991)**

B.3 Commercial TCMS

B.3.1 Montronix

Montronix supply a fairly wide range of TCM products that can be used with most CNC machines. However, the signal processing methods proposed by Montronix do not employ the state-of-the-art methods available today. A comparison between the on-line and a reference signal is generally used as a tool state classification strategy. Furthermore, the positioning of the sensors on machines not always justified properly. Although the systems supplied by Montronix seem to be the best available on the market, it lacks the use of modern signal processing techniques that will enhance the capabilities of the system.

The systems supplied by Montronix are either stand-alone systems, or integrate with the CNC machine. A diagrammatic representation of a Montronix system is shown in Figure B.1. In Figure B.2, some of the hardware for process monitoring is shown. The integrated process monitor works with open architecture or PC-based machines where the Montronix software directly loads into the operator interface. A monitoring unit without graphical display is available for other machines (without open architecture), or an LCD display unit can be used. These systems are mounted on top of the machining centres. A remote display unit is also available [245].

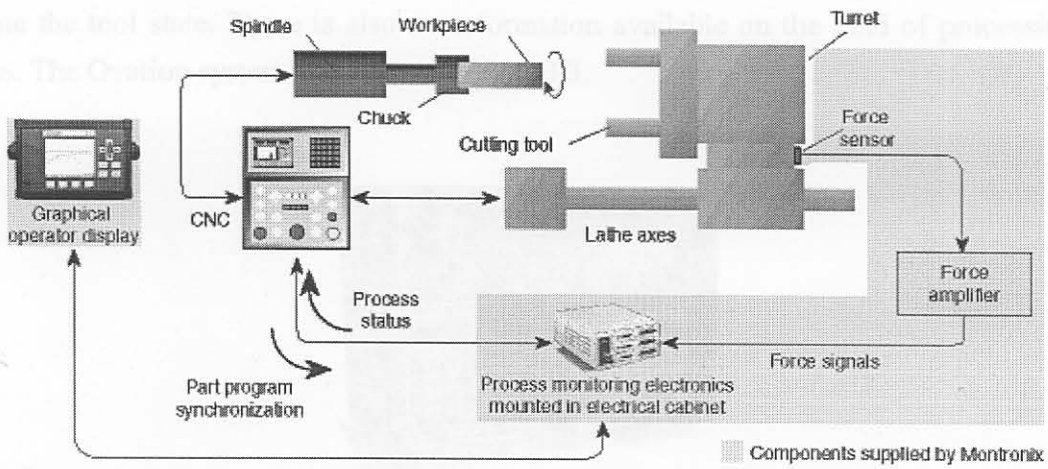


Figure B.1: Montronix TCM scheme for lathes [245]

B.3.2 Ovation

Ovation engineering developed a CNC tool monitor that can monitor the state of a range of tools that must be pre-programmed. The features of the Ovation CNC Tool Monitor includes [256]:

- Easy to use and program.
- Very fast response time (less than 10msec).
- Upload and download of all limits via RS232.
- Saves the last 300 alarms with date and time stamp.
- Extreme, wear and undercut limits for every tool and section.
- Up to 99 tools and 5 sections per tool.
- Parallel or serial machine interface.

- Monitoring of coolant flow and spindle speed.
- GE Specification Power monitoring.

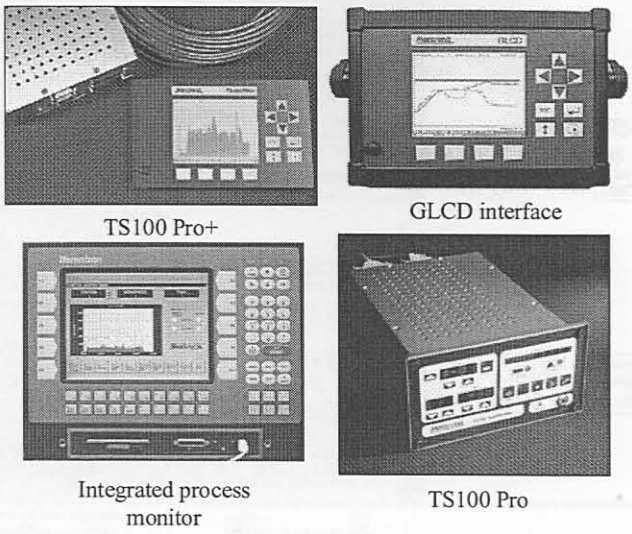


Figure B.2: Some of the TCM products from Montronix [245]

Additional transducers can be added to allow a different range to be used for smaller tools or for live tooling on lathes, and adaptive control can be added. The system only uses spindle power monitoring to determine the tool state. There is also no information available on the kind of processing that the system uses. The Ovation system is shown in Figure B.3.

Figure B.3: Ovation Engineering CNC Tool Monitor [256]

B.3.3 Artis Systems Inc.

Artis Systems specialises in tool and process monitoring. Their ranges of equipment for monitoring and diagnosis technology are used in the automotive industry worldwide. Artis supplies sensors, monitoring modules and data handling hardware. In Figure B.4, the range of hardware supplied by Artis is shown. The data visualisation can take place either through the CNC (open architecture) or a PC [250].

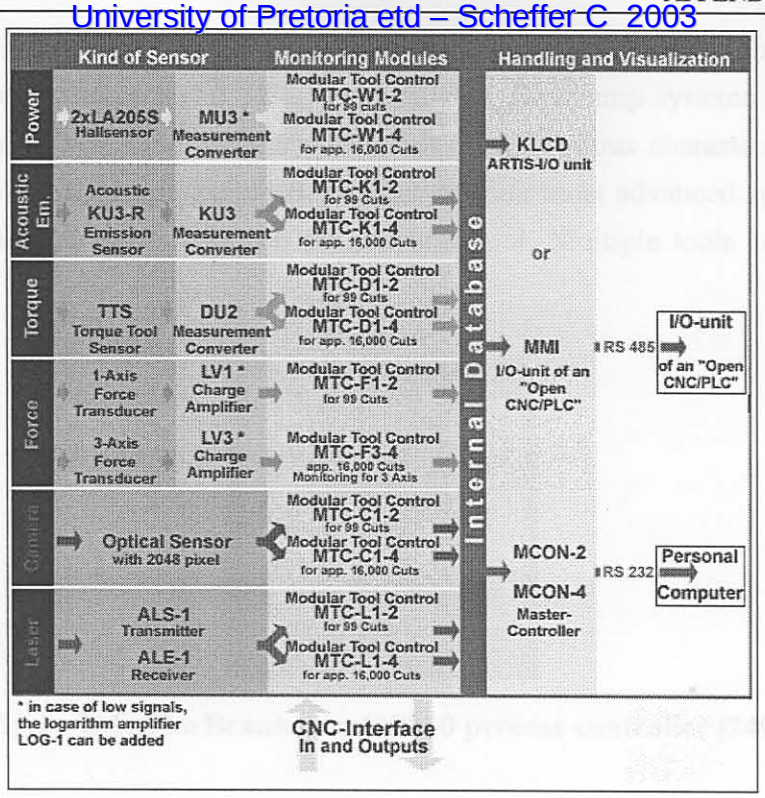


Figure B.4: Artis hardware for TCM [250]

Figure B.5 shows a diagrammatical layout of the Artis approach to TCM. The supported sensors are power, torque, force, AE, cameras and lasers. The tool control modules can be programmed for certain cuts.

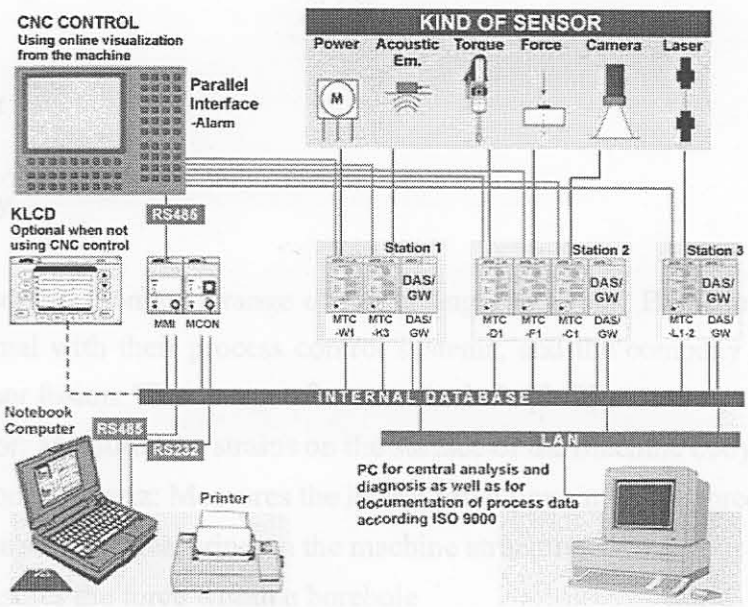


Figure B.5: Schematic representation of Artis TCM methodology [250]

B.3.4 Brankamp

Brankamp is a well-established German company that focuses on process control and monitoring. They offer a wide range of hardware for product quality, computer aided quality control software, and monitoring units mostly with large LCDs to allow process observation. The units are designed for all

cutting processes: presses, stamping machines, cold forming and general assembly and manufacturing equipment. A range of hardware for TCM is also available. Brankamp systems are either open architecture type of controllers, or stand-alone systems that boasts various characteristics. The Brankamp C8060 process controller, shown in Figure B.6, is one of their most advanced systems. The LCD display can show a variety of information on the machine tool. Multiple tools can also be monitored [249].

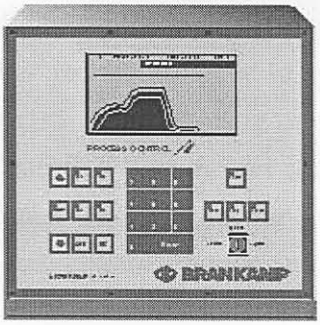


Figure B.6: The Brankamp C 8060 process controller [249]

This monitoring system has four sensors or monitoring channels. The system presents a lot of different process visualisations, for instance the envelope curve on multistage production machines. The whole process is wrapped up and monitored against the envelope limits. The following screens can be seen on the LCD display:

- Actual process values
- Process memory
- Tool life counter
- Productive counter
- Stop/Go-diagram
- Stop-code-memory

Brankamp offers sensors to monitor a range of machining operations. Brankamp sensors are also developed to work optimal with their process control systems, and the company attempts to emphasise the advantages of sensor fusion. Their range of sensors include [249]:

- Piezoelectric sensor: Measures the strains on the surface of the machine body
- One/Three- component-quartz: Measures the lateral force flow in one or three directions
- Magnetic inductive sensor: Measuring on the machine structure
- Vario sensor: Measures the force within a borehole
- Acoustic emission sensor: Measures high frequency noises
- Brankamp power sensor: Measures the power of motors on machine tools
- Micro sensor: Measures the strains on the surface of the machine body

B.3.5 Prometec

Prometec is a German company specialising in the fields of tool and machine monitoring, as well as process analysis for turning, drilling, milling, and grinding. Their systems are widely used in the auto-

motive industry. Their tool monitoring systems are also complemented by new systems for the detection of oscillations and imbalance. Prometec supplies a number of sensors for TCM, optimised for use with their other TCM hardware. These sensors include: force, torque, displacement, strain, position, hydraulic pressure, effective power, AE, structure-borne sound and vibration. All of these sensors are robust, and designed specifically for the purpose of TCM [253].

The TCMS supplied by Prometec consist of operator panel modules, machine interface modules and monitoring modules. The processing of the sensor signals are conducted by the monitoring modules, which work independently of one another and are assigned to a machining station. Various monitoring modules are available for differing extents of monitoring or numbers of cycles (tools or cuts). All monitoring modules are supplied in a compact housing for mounting in switch cabinets. Single dual channel monitoring modules are available. This allows, for example, a cost-effective solution for a twin-turret drilling machine by using a dual-channel monitoring module.

The monitoring module's notification of breakage, wear *etc.* are transmitted via a machine interface module. The display of the changes in the signals occurs over time, and an operator panel module is necessary for data visualisation. This allows for the adjustment of the monitoring parameters to the process at hand. This is done first of all during the teach-in where the progression of the signals is learned and the limits then defined. The characteristics of these systems are shown in Figure B.7. Figure B.8 shows the monitoring module, the machine interface module, the operator panel module and the PROVIS visualisation tool supplied by Prometec [253].

Operator panel module	OPM12	OPM 20 with PROVIS	Open CNC with PROVIS
Display	LCD, 4 x 20 characters	VGA, color	PROVIS software in the operator unit of the open NC. Connection to monitor module via serial interface. For DOS, Windows 95 and Windows NT 4.0.
Program-dependent data storage	—	x	
RS 232 terminal for notebook, modem etc.	x	x	
Printer terminal	—	x	
Terminal for PC keyboard	—	x	
External connector module for the terminals listed above	Option	Option	
Protective housing	Option	Option	

Machine interface module	MINI	MIDI	BAPSI	PROFI-BAPSI INTER-BAPSI
1-channel				As with BAPSI, but serial operation in accordance with PROFIBUS-DP/ INTERBUS-S protocol
Cycles	8	—	250	
Output messages	3	—	7	
2-channel				
Cycles per channel	2	30	250	
Output messages per channel	1 Plus supplementary, joint output for collision message	3	7 Two BAPSI interface modules are to be used for 2-channel operation	
Connecion to machine interface module	Ribbon cable	2-wire cable	2-wire cable	2-wire cable

Figure B.7: Prometec TCMS [253]

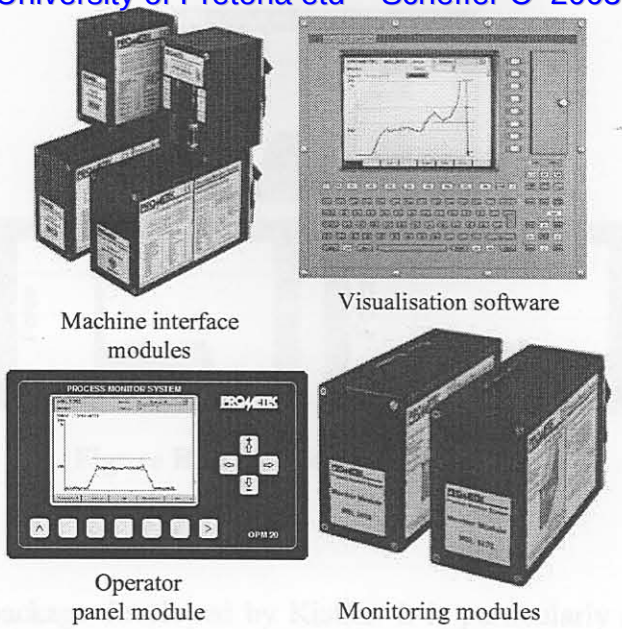


Figure B.8: TCM modules by Prometec [253]

B.4 TCM software

B.4.1 Introduction

A number of companies supply Windows supported software for TCM. Most of these software packages only enable relatively simple data processing. The software programs are usually graphical interfaces displaying the sensor signals and a statistical summary of events. Normally, these packages are used to view trends in sensor signals over a period of time, and to perform comparative TCM studies. Table B.2 summarise the visualisation methods supplied by various companies (from Jemielnaik [247]). Some are discussed in more detail.

Table B.2: Visualisation methods applied in TCM systems [247]

Visualisation method	Supplier					
	Artis	Brankamp	Kistler	Montronix	Nordmann	Prometec
Simple digital / bar display	●	●		●	●	●
Graphic display	●	●	●		●	●
Via open CNC control system	●			●	●	●
Via PC / Notebook computer	●	●	●	●		●

B.4.2 Montronix

The M-View package from Montronix was developed to enhance the performance of the Montronix range of TCM products. M-View is a typical visualisation tool for trends and comparative TCM studies. Typical windows from the M-View package are shown in Figure B.9.

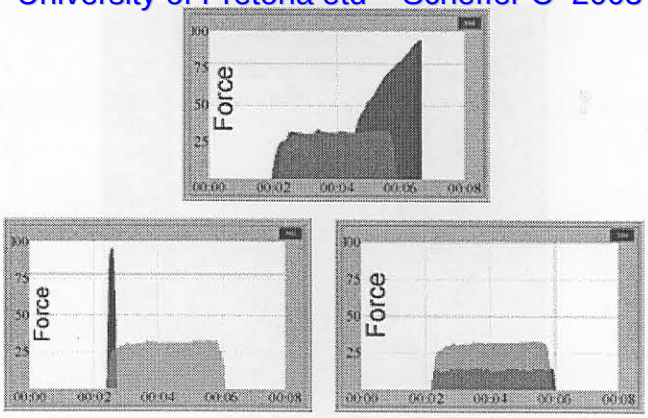


Figure B.9: M-View windows [245]

B.4.3 Kistler

DynoWare is a software package developed by Kistler. It is particularly suitable for force measurements with Kistler dynamometers and force sensors. DynoWare may be used to analyse, optimise and document cutting force measurements in a simple and efficient way. Programmed in the Windows environment, DynoWare is a powerful data acquisition system that can be used universally. Eight free selectable measuring signals may be displayed and analysed in several windows. DynoWare contains an easy to use integrated database containing default set up values for the most common factors. (sensitivity, scale, *etc.*). DynoWare also features [246]:

- 8 measuring channels with 12-bit resolution each.
- Measuring range per channel: ±10 V.
- Max. sampling rate of A/D card, divided by number of channels selected.
- 100 kHz/ number of channels.
- Remote control of Kistler electronics via RS-232C or IEEE-488 interface.
- Several measuring cycles can be recorded during a trial.
- Import / Export of measured data in ASCII format (Labview, Dia Dago, Excel).
- Trigger with keyboard, analogue, digital.

B.4.4 Prometec

The PROVIS software by Prometec offers facilities for visualisation of the process signal to enable the diagnosis and optimisation of machining operations. PROVIS software are available in Windows or DOS versions for all of the leading control manufacturers including Siemens, Indramat, GE-Fanuc, Schleicher *etc.* Figure B.10 shows an example of a PROVIS window [253].

B.5 Monitoring strategy

Monitoring strategies applied by different suppliers can be grouped into six approaches, summarised in Table B.3 (after Jemielniak [247]). Most of these are based on static limits. Static parameters are process parameters that remain fixed during the processing of the current workpiece. The basic monitoring strategies will be discussed in more detail.

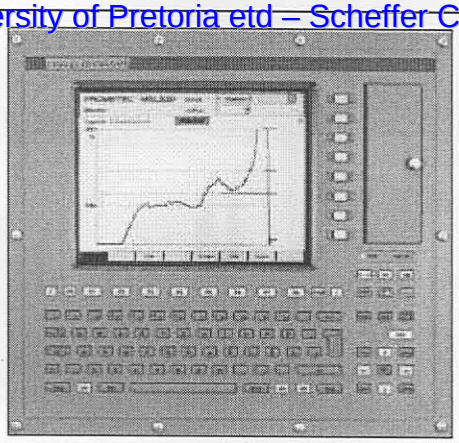


Figure B.10: PROVIS software by Prometec, in operator display panel [253]

Table B.3: Strategies applied in TCMS [247]

Strategy	Supplier					
	Artis	Brankamp	Kistler	Montronix	Nordmann	Prometec
Simple fixed limits				●	●	●
Time defined limits			●		●	●
Part signature	●	●	●	●	●	
Pattern recognition				●		
Wear estimator				●		
Dynamic limits						●

B.5.1 Simple fixed limits

This simple strategy is based on fixed limits that apply to the signals from the sensors (after simple pre-processing). During machining of the first workpiece, the monitor reference signal is constructed, by normalising it between 0 and 100 % for each cut and each tool. The positions of limits relative to the set standards are defined (manually or automatically). For example, three limits can be set, one for a worn tool, one for a broken tool, and one for collision. Instead of fixed limits, floating limits can also be used which track cycle-to-cycle trends by using information from the current cycle to adjust the limits of the next cycle automatically.

B.5.2 Time defined limits

Fixed limits for a machining process can also be time displaced, which means that limits for certain sub-areas of the signal are set. This can enhance the performance of the fixed limit strategy to detect sudden tool wear, a premature cut, incorrect adjustments or sudden tool fracture.

2.5.3 Part signature

The cutting of a certain part is divided into segments using stepped limits based on time and / or position. This can create a "part signature" that can more closely track a complex cutting cycle than a single set of fixed limits. The part signature is like a combination of floating and time defined limits. Whenever the current signal has higher or lower values of the reference signature, an alarm goes off. Figure B.11 shows an example of the part signature concept from Montronix [245].

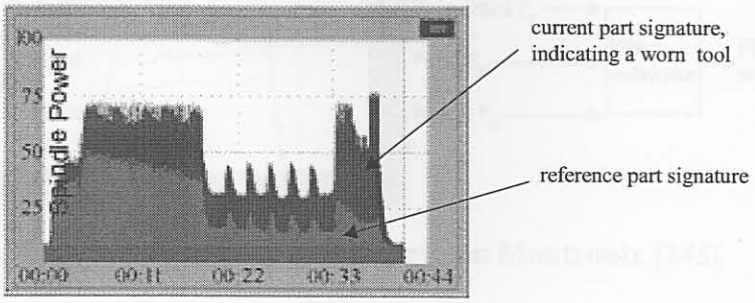


Figure B.11: Part signature concept from Montronix [245]

B.5.4 Pattern recognition

Montronix developed a tool breakage monitoring strategy based on pattern recognition for turning. The system stores a number of reference signals indicative of tool breakage, such as breakage of a carbide or ceramic tool. The system continuously monitors the current signal for one of the breakage patterns. When the pattern occurs, the machine stops. This method has the advantage that the pattern is independent of the process signal value, and therefore it can be applied to a number of different cuts. It is therefore also independent of the tool condition and other machining parameters. Figure B.12 displays an example of the pattern recognition process by Montronix [245].

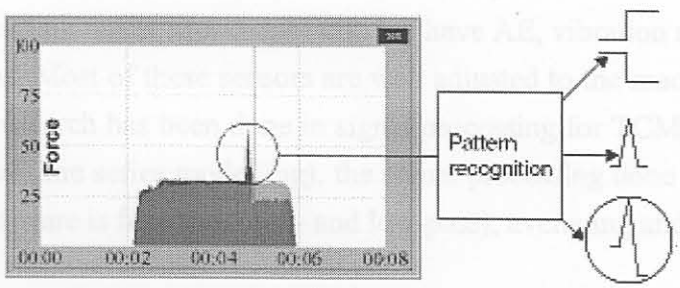


Figure B.12: Pattern recognition for tool breakage detection in turning form Montronix [245]

B.5.5 Wear estimator

The wear estimator is a proprietary technique developed by Montronix used for turning tool flank wear estimation. The method uses the relationship between all three cutting force components, and requires a three-component force sensor. The technique can distinguish between normal tool wear and variations in process parameters. These variations, such as workpiece hardness or runout, can adversely affect other techniques such as part signature analysis. Figure B.13 shows the wear estimator concept from Montronix [245].

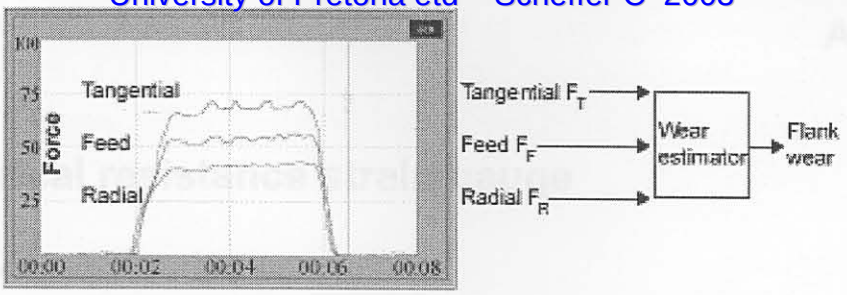


Figure B.13: Wear estimator from Montronix [245]

B.5.6 Dynamic limits

The PROMOS system for Prometec incorporates the use of dynamic limits for tool breakage detection. Two dynamic limits, above and below the monitoring signal, follow the signal continuously for every load level. In the case of an extremely fast crossing of one of the two dynamic limits, the tool state is determined by visual comparison of the monitoring signal. Another version of the dynamic limit by Prometec, is when a special feature is generated from the force signal. The monitoring strategy then combines the feature generation with dynamic limits, to enhance the performance of the TCMS to automatically adjust over a range of sensor signals in different machining situations.

B.6 Conclusion

The most commonly used sensors for TCM in commercial systems are sensors measuring force or quantities related to the cutting force. Most suppliers also have AE, vibration and optical sensors available for some applications. Most of these sensors are well adjusted to the machining environment. Despite the fact that much research has been done in signal processing for TCM, (such as spectral analysis, statistical analysis and time series modelling), the signal processing done by commercial suppliers of TCM hardware and software is filtering (high- and low-pass), averaging and rms calculations.

Since constant limit methods only work when all restrictions remain constant, the use of self-adjusting limits is more appropriate. Only Prometec with its "dynamic limit" method and Montronix with the pattern recognition and wear estimator use more sophisticated signal processing, although these are not nearly the state-of-the-art methods in modern signal processing and data classification. Monitoring strategies developed by researchers nowadays almost always incorporate more sophisticated signal processing where features are extracted from sensor signals. In commercial systems, the one-sensor, one process approach dominates. The processing and fault detection techniques on the sensor signals in commercial TCMS primarily remain independent. Only the Montronix wear estimator uses more than one signal to classify the tool condition.

C. The electrical resistance strain gauge

C.1 Introduction

The most common type of strain transducer is the electrical resistance strain gauge. Strain gauges are extremely versatile due to their small size, low cost and the fact that they are available in a wide range of configurations for different applications. A diagram of the single element strain gauge is shown in Figure C.1.

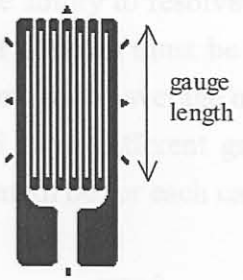


Figure C.1: Typical single element strain gauge

Strain gauges are usually connected in a Wheatstone bridge configuration. A basic Wheatstone bridge configuration is shown in Figure C.2.

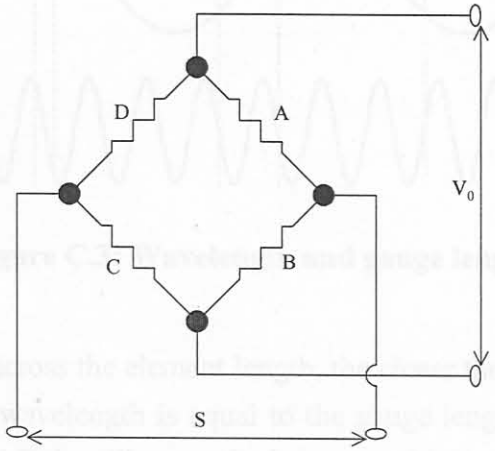


Figure C.2: Strain gauge bridge configuration

In Figure C.2, four strain gauges are represented by A, B, C and D, and their resistances would typically be 120Ω each. A voltage V_0 is supplied across the bridge, and this is typically 10V. The signal S is then measured across the other ends of the bridge. A change in resistance of the gauge is caused by the base strain (elongation or compression) of the surface it is adhered to. If all four gauges (A, B, C and D) are adhered to the surface it is called a full-bridge configuration. If two gauges are attached to the surface, the other two are "dummy" gauges, and this is called a half-bridge configuration. If only one is adhered to the surface it is called a quarter bridge. Many other configurations are possible and

are used for particular applications. The types of strain gauge configurations and their applications are beyond the scope of this text. However, there are at least three significant considerations when using strain gauges for measuring dynamic signals. The following limitations should be considered before choosing a strain gauge:

- Frequency resolution of the gauge
- Spatial resolution of the gauge
- The so-called rise time of the gauge

C.2 Frequency resolution

The ability to resolve a wave spatially becomes limited for pulses near the order of the gauge length. As the gauge length increases, so does the ability to resolve the spatial location of the pulse decrease. Therefore the total length of the pulse of concern must be much greater than the gauge length. The output of a strain gauge tends to give an integrated average of the strains imposed over its length. Consider the three different wavelengths and three different gauge lengths, shown in Figure C.3. What would the average strain over the gauge length be for each case?

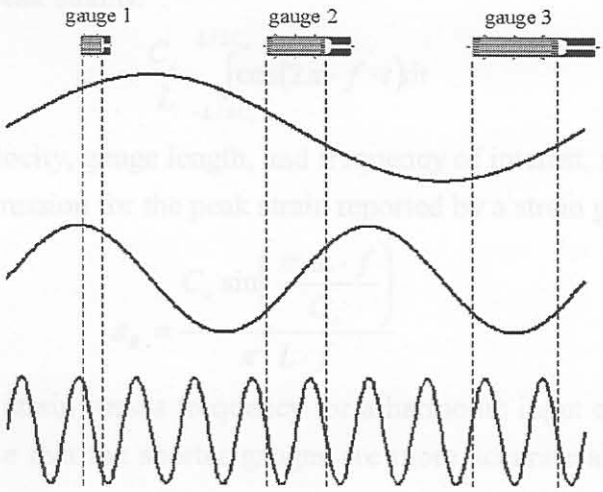


Figure C.3: Wavelength and gauge length

The smaller the strain gradient across the element length, the closer the output will be to the true strain. For the special case where the wavelength is equal to the gauge length, the average output would be zero, but the actual strain is not! This will occur for integer multiples of the wavelength. As the wavelength decreases, so too does the ability to resolve the peak strain due to the averaging effect over the gauge length. The peak strain is consequentially always estimated low. The concept is illustrated in Figure C.4 - the steeper the gradient (i.e. short wavelength), the less the peak strain can be resolved [257].

C.3 Spatial Resolution

For harmonic waves propagating through a strain gauge, the peak output from the gauge will always occur when the peak strain is centred in the gauge. To determine the closed form solution for the peak

strain reported by a strain gauge, the harmonic wave must be integrated over the gauge length.

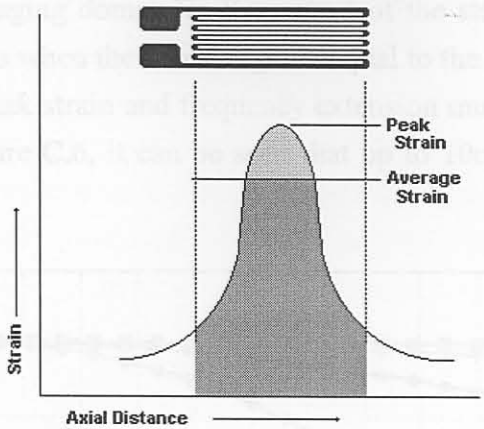


Figure C.4: Averaging effect on peak strain [258]

By recognizing that the peak of a cosine wave coincides with the centre of the gauge, the following expression yields the peak strain reported by the strain gauge (remember the peak output of the gauge is always less than the true peak strain):

$$\varepsilon_R = \frac{C_o}{L}\int_{-L/2C_o}^{L/2C_o}\cos(2\pi \cdot f \cdot t)\mathrm{d}t \tag{C.1}$$

C_o, L, and f are the wave velocity, gauge length, and frequency of interest, respectively. The evaluation of this integral yields an expression for the peak strain reported by a strain gauge:

$$\varepsilon_R = \frac{C_o \sin\left(\frac{\pi \cdot L \cdot f}{C_o}\right)}{\pi \cdot L \cdot f} \tag{C.2}$$

A plot of the reported peak strain versus frequency for a harmonic input of amplitude one for steel is shown in Figure C.5. Notice that the shorter gauges are more accurate at measuring high frequency strains.

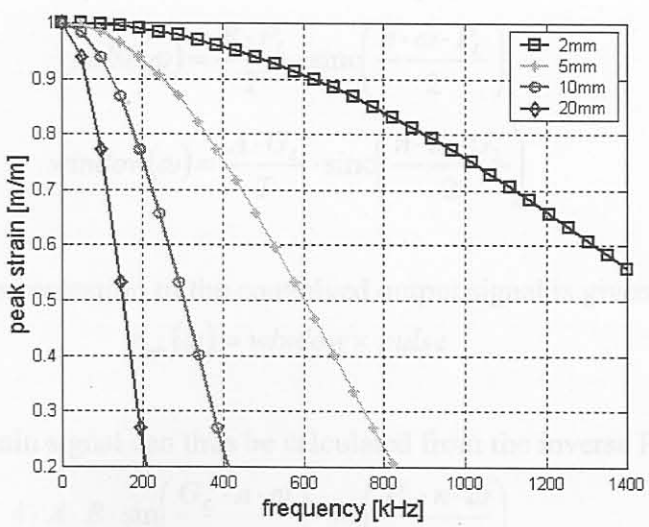


Figure C.5: Reported peak strain from various gauge length strain gauges

Also note that accurate static (DC) strain values are reported. As frequency increases, however, the reported strain becomes drastically attenuated. As was expected, shorter gauges offer better frequency response. The effect of the averaging dominates the output of the strain gauge for high frequencies. The output from the gauge is zero when the wavelength is equal to the gauge length. To choose a strain gauge, the desired accuracy of peak strain and frequency extension must be considered. Looking at the 0-50kHz region of the plot, Figure C.6, it can be seen that up to 10mm gauges can be used on steel with reasonable accuracy.

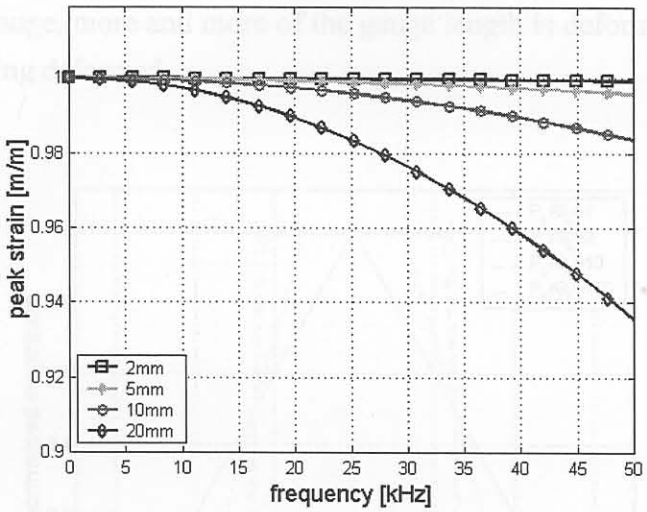


Figure C.6: Reported peak strain from various gauge length strain gauges

Another condition (related to the physics of the strain gauge) exists that distorts the output of strain gauges. This is a windowing effect that distorts the actual strain pulse. The effective gauge length represents a rectangular window that in effect changes the shape of the true strain pulse. To calculate the effect of windowing requires convolution. Recall that convolution in time is equal to multiplication in frequency. Hence, the inverse Fourier transform of the product of two functions of frequency is equal to the convolution of the two same functions in the time domain. Consider a square pulse of magnitude B and length P_L / 2, measured by a rectangular window of magnitude A and length G_L / 2, and both pulses having period T. The Fourier series of the pulses and window are:

$$pulse(\omega) = \frac{B \cdot P_L}{T} \cdot \mathrm{sinc}\left(\frac{n \cdot \omega \cdot P_L}{2}\right) \tag{C.3}$$

$$window(\omega) = \frac{A \cdot G_L}{T} \cdot \mathrm{sinc}\left(\frac{n \cdot \omega \cdot G_L}{2}\right) \tag{C.4}$$

The frequency domain representation of the convolved output signal is given by:

$$\varepsilon_{out}(\omega) = window \times pulse \tag{C.5}$$

The convoluted time domain signal can thus be calculated from the inverse Fourier transformation:

$$\varepsilon_{out}(t) = \sum_{-\infty}^{\infty} \frac{4 \cdot A \cdot B \cdot \sin\left(\frac{G_L \cdot n \cdot \omega}{2}\right) \cdot \sin\left(\frac{P_L \cdot n \cdot \omega}{2}\right)}{T^2 \cdot n^2 \cdot \omega^2} \exp(-i \cdot n \cdot \omega \cdot t) \tag{C.6}$$

C.4 Rise time of strain gauge

By plotting this rather complex equation, the windowing effects of the strain gauge can be seen, as shown in Figure C.7. This assists to predict gauge output for many configurations of gauge lengths (G_L) and pulse lengths (P_L). Figure C.7 is a plot of the convolved output of various pulse lengths to gauge length ratios. Notice that all of the convolved pulses exhibit increased rise time, not characteristic of the square input pulse. This rise time is due to the period of time when the pulse just begins and just exits the strain gauge. Just before entering the gauge, the strain gauge output is zero. As the pulse progresses through the gauge, more and more of the gauge length is deformed, until the point when all of the gauge length is being deformed.

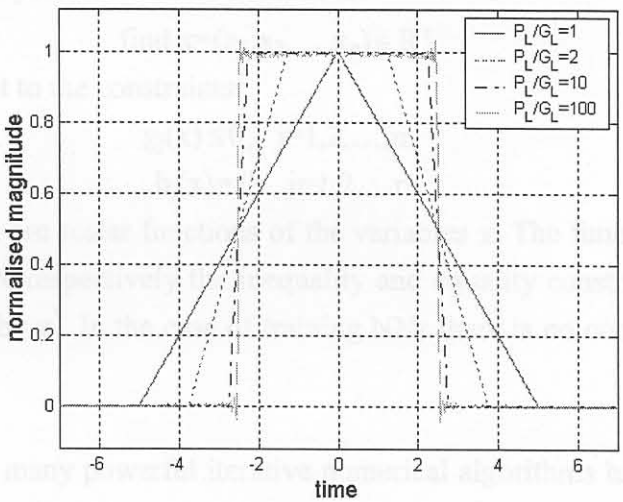


Figure C.7: Convolved output of strain gauges

In the limiting case where the gauge length equals the pulse length, the maximum value is only reached only once because there exists only one time when the gauge window and pulse window perfectly overlap. It can be seen that the windowing effect can be minimized by selecting a sufficiently large P_L / G_L ratio. However, with the measured strain signals in hand, corrections for the windowing effect of the gauge can also be employed by dividing the Fourier transform of the strain signal by the Fourier transform of the gauge measurement window [257].

C.5 Conclusion

With reference to the strain gauges selected for measurements as described in Chapter 4, these considerations were taken into account when selecting the gauges and also when analysing the signals obtained from them. In essence, very small gauge lengths were used compared to the fairly low frequency range that was investigated. Thus, the effects of rise-time and frequency distortion were minimal. Furthermore, it should be mentioned that the selected strain gauges were self-temperature-compensated, which means that the traditional use of dummy gauges was not necessary.

D. Mathematical Optimisation

D.1 General formulation

Training neural networks is in actual fact nothing but solving an unconstrained optimisation problem. Many algorithms exist whereby Neural Networks (NNs) can be trained. During the course of this research, different algorithms were investigated for their feasibility to train the dynamic NNs. For on-line implementation, the algorithm must be robust, quick and stable. The general formulation of the mathematical optimisation problem with constraints is:

$$\text{find } \mathbf{x}=(x_1,x_2,\ldots,x_n)\in R^n \tag{D.1}$$

that minimizes $f(\mathbf{x})$ subject to the constraints:

$$g_j(\mathbf{x})\leq 0,\quad j=1,2,\ldots,m \tag{D.2}$$

$$h_j(\mathbf{x})= 0,\quad j=1,2,\ldots r \tag{D.3}$$

where $f(\mathbf{x})$, $g_j(\mathbf{x})$ and $h_j(\mathbf{x})$ are scalar functions of the variables $\mathbf{x}$. The function f is called the objective function and g_j and h_j are respectively the inequality and equality constraint functions. A local optimum solution is denoted by $\mathbf{x}^*$. In the case of training NNs there is no constraints and thus no g and h functions.

Over the past thirty years many powerful iterative numerical algorithms have been developed to solve the above-mentioned general problem. It is however true to say that no single algorithm dominates in being superior to all others when applied to different subclasses of the general problem. Depending on the degree of non-linearity, the presence of noise or discontinuities in the functions, the number of variables involved and the time required to evaluate the functions, some methods will be preferred above others depending on the efficiency, accuracy or reliability required. Difficulties were experienced with conventional methods, and for this reason new methods were investigated for training the dynamic NNs.

D.2 Algorithms

D.2.1 SQSD

The Spherical Quadratic Steepest Descent (SQSD) Method [259] was employed to train the dynamic network. The method is a gradient-only algorithm for unconstrained minimisation, which applies the steepest descent method to successive spherical quadratic approximations of the objective function. No explicit line searches are preformed which makes the method computationally very efficient. It has also been shown that method performs extremely well with ill-conditioned problems, which is sometimes the case with neural networks. The gradient functions of the neural networks were calculated with finite differences. The selection of the step size when using this technique requires some care, especially if the functions are not analytical (*e.g.* NNs). The algorithm provided excellent results in less time than conventional training procedures.

The algorithm can be described as follows by specifying the convergence criteria $\varepsilon_g, \varepsilon_x$, steplimit $d > 0$ and selecting a starting point $\mathbf{x}^0$. Set $c_0 := \left\|\nabla f\left(\mathbf{x}^0\right)\right\| / d$. Let $k := 1$ and proceed to (1).

(1) If $\left\|\nabla f\left(\mathbf{x}^{k-1}\right)\right\| < \varepsilon_g$, then $\mathbf{x}^* \cong \mathbf{x}^c = \mathbf{x}^{k-1}$ and stop; otherwise set:

$$\mathbf{x}^k := \mathbf{x}^{k-1} - \frac{\nabla f(\mathbf{x}^{k-1})}{c_{k-1}} \tag{D.4}$$

2) If $\left\|\mathbf{x}^k - \mathbf{x}^{k-1}\right\| > d$, then set

$$\mathbf{x}^k := \mathbf{x}^{k-1} - d\frac{\nabla f(\mathbf{x}^{k-1})}{\left\|\nabla f(\mathbf{x}^{k-1})\right\|} \tag{D.5}$$

if $\left\|\mathbf{x}^k - \mathbf{x}^{k-1}\right\| < \varepsilon_x$, then $\mathbf{x}^* \cong \mathbf{x}^c = \mathbf{x}^{k-1}$ and stop.

3) Set

$$c_k := \frac{2\left[f\left(\mathbf{x}^{k-1}\right) - f\left(\mathbf{x}^k\right) - \nabla^T f\left(\mathbf{x}^k\right)\left(\mathbf{x}^{k-1} - \mathbf{x}^k\right)\right]}{\left\|\mathbf{x}^{k-1} - \mathbf{x}^k\right\|^2} \tag{D.6}$$

if $c_k < 0$, then set $c_k := 10^{-60}$.

4) Let $k := k + 1$ and then go to step 1 for next iteration.

The algorithm was programmed in Matlab for training the neural networks.

D.2.2 ETOPC

The ETOPC algorithm can solve unconstrained and constrained problems by means of the conjugate gradient method. Constrained problems are solved via a penalty function formulation. The specific implementation used is that of Snyman [260] which is unique in the sense that it employs no explicit line searches. With the ETOPC conjugate gradient program the user may select either a Fletcher-Reeves or a Polak-Ribiére implementation. The ETOP (unconstrained) algorithm is formulated as follows [261]:

Assume x_0, Δt given and set $v_0 \leftarrow -\nabla F_0 \Delta t = s_0 \Delta t$. Do for $k = 0,1,2,....$ until convergence:

1) Set

$$x_{k+1} \leftarrow x_k + v_k \Delta t = x_k + s_k \Delta t^2 \tag{D.7}$$

2) Determine:

$$\Delta F_k = -v_k^T a_k' \Delta t \tag{D.8}$$

where:

$$a_k' = (a_k + a_{k+1})/2 \tag{D.9}$$

$$a_k = -\nabla F(x_k) \tag{D.10}$$

$$\theta = -\rho / (\Delta F_k - \rho) \tag{D.11}$$

$$x_{k+1}^* = x_k + \theta \Delta x_k / 2 \tag{D.12}$$

3) Set:

$$x_{k+1} \leftarrow x^*_{k+1} \tag{D.13}$$

4) Determine:

$$v_{k+1} \leftarrow \left(-\nabla F_{k+1} + \beta_{k+1} v_k / \Delta t\right)\Delta t = s_{k+1}\Delta t \tag{D.14}$$

with the Polak-Ribiere formula:

$$\beta_i = \frac{\left(\nabla F_i - \nabla F_{i-1}\right)^T \nabla F_i}{\left\|\nabla F_{i-1}\right\|^2} \text{ for } i = 1,2,... \tag{D.15}$$

The algorithm was implemented in Matlab for training the dynamic networks.

D.2.3 LFOP

This algorithm is the LFOPC algorithm of Snyman [261-263]. It is a gradient method that generates a dynamic trajectory path from any given starting point towards a local optimum. This method differs conceptually from other gradient methods, such as SQP, in that no explicit line searches are performed.

The original leap-frog method [262-263] for unconstrained minimization problems seeks the minimum of a function of n variables by considering the associated dynamic problem of a particle of unit mass in a n-dimensional conservative force field, in which the potential energy of a particle at point $\mathbf{x}(t)$ at time t is taken to be the function $f(\mathbf{x})$ to be minimized. The reader is referred to [261-263] for a mathematical formulation of LFOP.

The optimisation is terminated when either of the convergence criteria becomes active:

$$\|\Delta \mathbf{x}\| < \varepsilon_x \text{ or } \|\nabla f(\mathbf{x})\| < \varepsilon_g. \tag{D.16}$$

It is recommended that the step size, δ, be of the same order of magnitude as the "diameter of the region of interest", hence:

$$\delta = \frac{1}{10}\sqrt{\sum_i^n (\text{range}_i)^2} \tag{D.17}$$

where range_i is the size of the region of interest in the i-th variable direction. In practice the choice of μ and μ_{max} should be coupled to the accuracy required.

D.2.4 PSOA

The PSOA simulates the physical movement of social creatures, for instance the movement of a flock of birds. Consider a flock of p particles of birds. For particle i, Kennedy and Eberhart [224] proposed that the position $\mathbf{x}^i$ be updated as:

$$\mathbf{x}^i_{k+1} = \mathbf{x}^i_k + \mathbf{v}^i_{k+1} \tag{D.18}$$

where $\mathbf{x}$ represents the position of each bird, $\mathbf{v}$ the velocity, and k a unit pseudo-time increment, and would in this case be an inner step. Shi and Eberhart [264] proposed that the velocity $\mathbf{v}^i$ be updated with:

$$\mathbf{v}_{k+1}^{i} = w\mathbf{v}_{v}^{i} + c_1 r_1 \left(\mathbf{p}_k^i - \mathbf{x}_k^i\right) + c_2 r_2 \left(\mathbf{p}_k^g - \mathbf{x}_k^i\right) \tag{D.19}$$

where $\mathbf{p}_k^i$ represent the best ever position of particle i at time k, and $\mathbf{p}_k^g$ represents the global best position in the swarm at time k. The numbers r_1 and r_2 are random uniform numbers between 0 and 1. The inertia term w, is chosen as 1 and then decreased linearly during optimisation. The cognitive and social scaling factors c_1 and c_2 are chosen as $c_1 = c_2 = 2$ in order to allow a mean of unity.

In the case of the training of the dynamic networks, the network biases and weights were taken as the position vector **x**. The algorithm is stopped when it reaches a suitable convergence tolerance on the objective function value. If the algorithm does not reach the tolerance value, the vector **x** is re-initialised. This is necessary for the case when an old tool is replaced with a new tool, because the swarm moved with the wear of the older tool, it will be too far away from the solution to converge within limited steps for the new tool. The number of particles that were used was 25, and the maximum number of steps 40. The random initialisation for bias and weight values was made between -1 and +1. The algorithm usually achieved convergence within less than about 15 inner steps in a matter of seconds on a PIII computer.

E. Custom hardware specifications

E.1 Introduction

As mentioned in the main text of the thesis, a special measurement system was developed in-house for cost-effective strain gauge measurements on the tool holder. The details of the system are described here. The system was developed to measure strain with three strain gauge rosettes, each connected in a half-bridge configuration. Note that 120Ω self-temperature compensated strain gauges should be used. The basic components of the system are:

- HBM clip strain gauge amplifiers type AE101
- HBM clip voltmeter type DA101
- HBM clip power supply type NT101
- variable "dummy" gauges
- low-pass filter with power supply
- user interface for Eagle A/D cards (PC-30U)

All of these components will be described in more detail, besides the particular details of the HBM hardware, which could be sourced from HBM directly. The user interface for the Eagle card is very simple, and only brief details are given. A special connection cable was made for the strain gauge measurement system, and only this cable should be used to connect the custom hardware to Eagle cards. The complete system was built into a robust steel enclosure, depicted in Figure E.1. A picture of the enclosure with the front panel open is shown in Figure E.2.

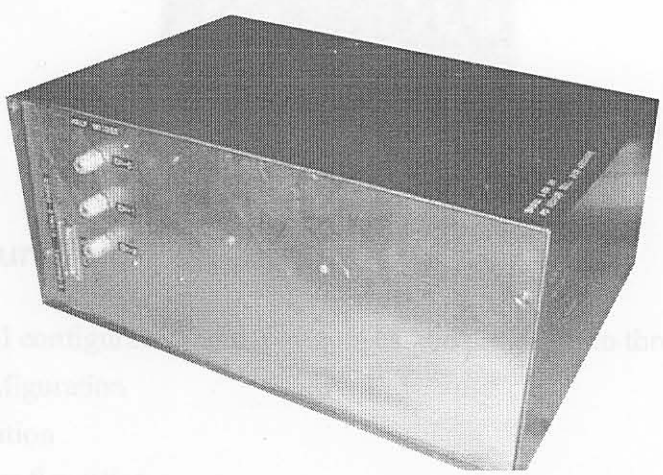


Figure E.1: Strain gauge measurement system

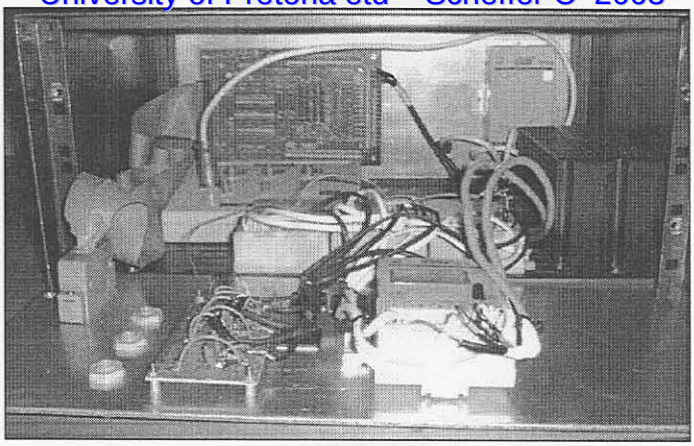


Figure E.2: Enclosure with front panel open

E.2 PC-30U

The PC-30U card is a simple analogue card that protects the PC-30 A/D card against high voltage peaks. A brief description follows:

- External adapter board to be used with any PC-30xx DAQ Card
- High degree of protection from high voltage spikes, accidental shorts and overloads.
- Internal PSU is assumed (thus PC-30xx card must be configured with ±12V and ±5V line enabled).
- Fuses F1, F2, F3 rated @ 125mA.

Only use the supplied cable to connect the strain gauge box to the PC-30 card. Contact Eagle Technologies [265] for additional information. The PC-30U is shown in Figure E.3.

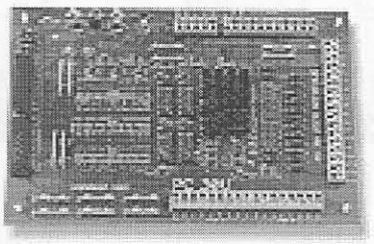


Figure E.3: PC-30U

E.3 Overall configuration

For simplicity, the overall configuration of the system can be divided into three parts:

- power supply configuration
- channel configuration
- cable screening configuration

E.3.1 Power supply configuration

The setup of the power supply is graphically depicted in Figure E.4. One normal South African mains plug is provided for the complete system, and it is shared with the HBM power supply and the power supply for the low-pass filter.

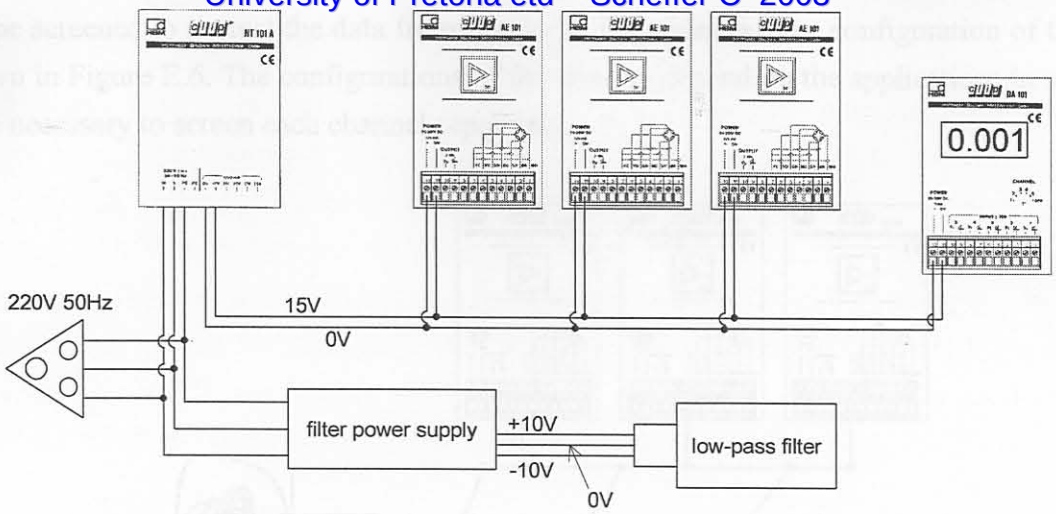


Figure E.4: Power supply configuration

E.3.2 Channel configuration

The three channels are formed from the output of the AE101 amplifiers. From this point, a single analogue ground and three channels are formed. Hence, the system is single-ended and the A/D card should be configured likewise. From the amplifiers the outputs pass through the low-pass filter, and then to the PC-30U. The output cable from the PC-30U is fastened on the front panel of the box and can be connected directly to an Eagle A/D card with the provided cable. The setup of the channels is shown schematically in Figure E.5.

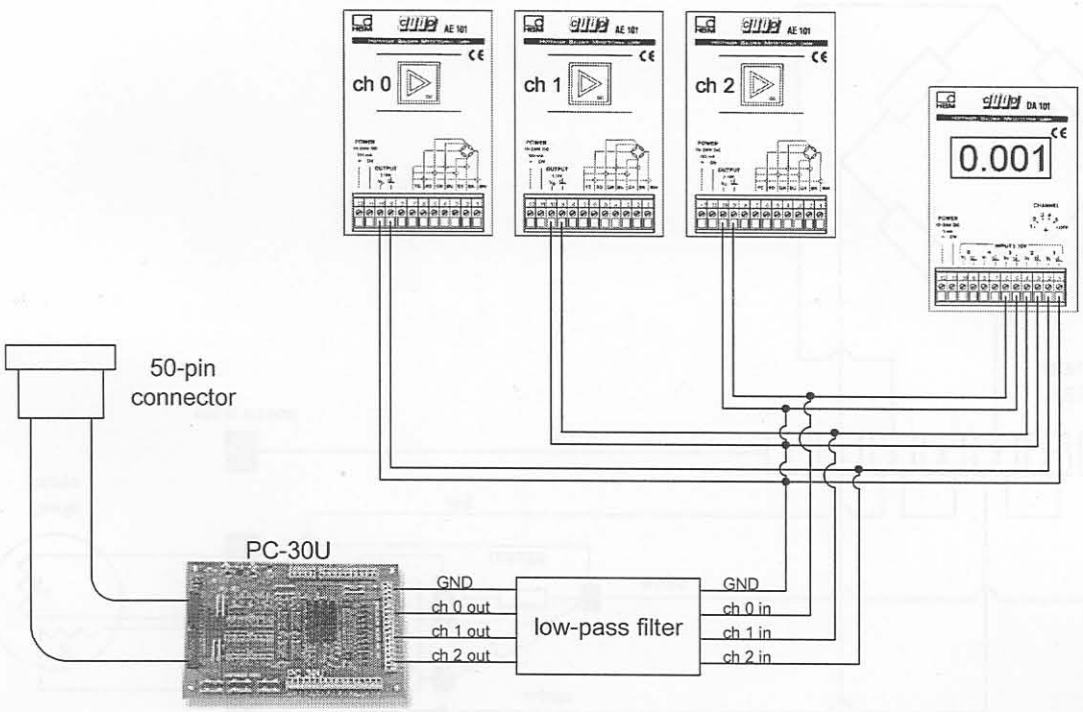


Figure E.5: Channel configuration

E.3.3 Cable screening configuration

All cables carrying very low voltage signals should be screened. The screenings can be connected to the appropriate pins on the AE101 amplifiers. Note that the actual strain gauges and the dummy resis-

tors must be screened to protect the data from electrical disturbances. The configuration of the screening is shown in Figure E.6. The configurations of the screens depend on the application: In some cases it might be necessary to screen each channel separately.

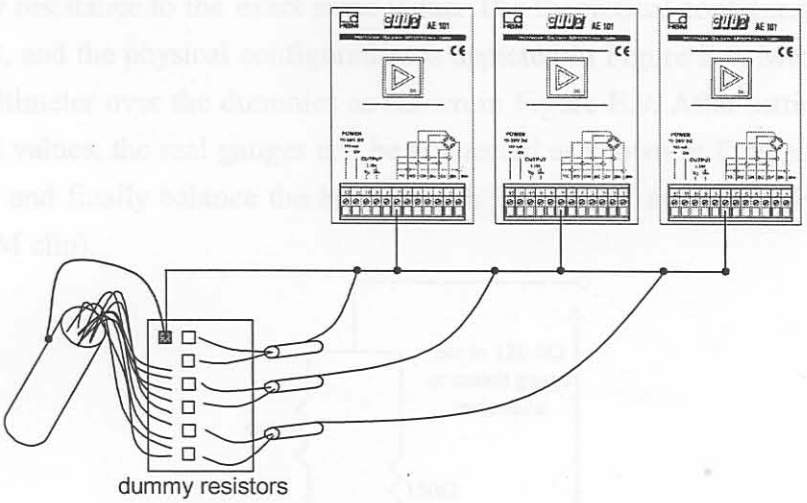


Figure E.6: Cable screening configuration

E.3.4 Per-channel configuration for half-bridges

The per-channel configuration for each half-bridge is graphically depicted in Figure E.7. Each half-bridge within the system is connected in the exact same way. The physical colours of the cables are also indicated on the figure.

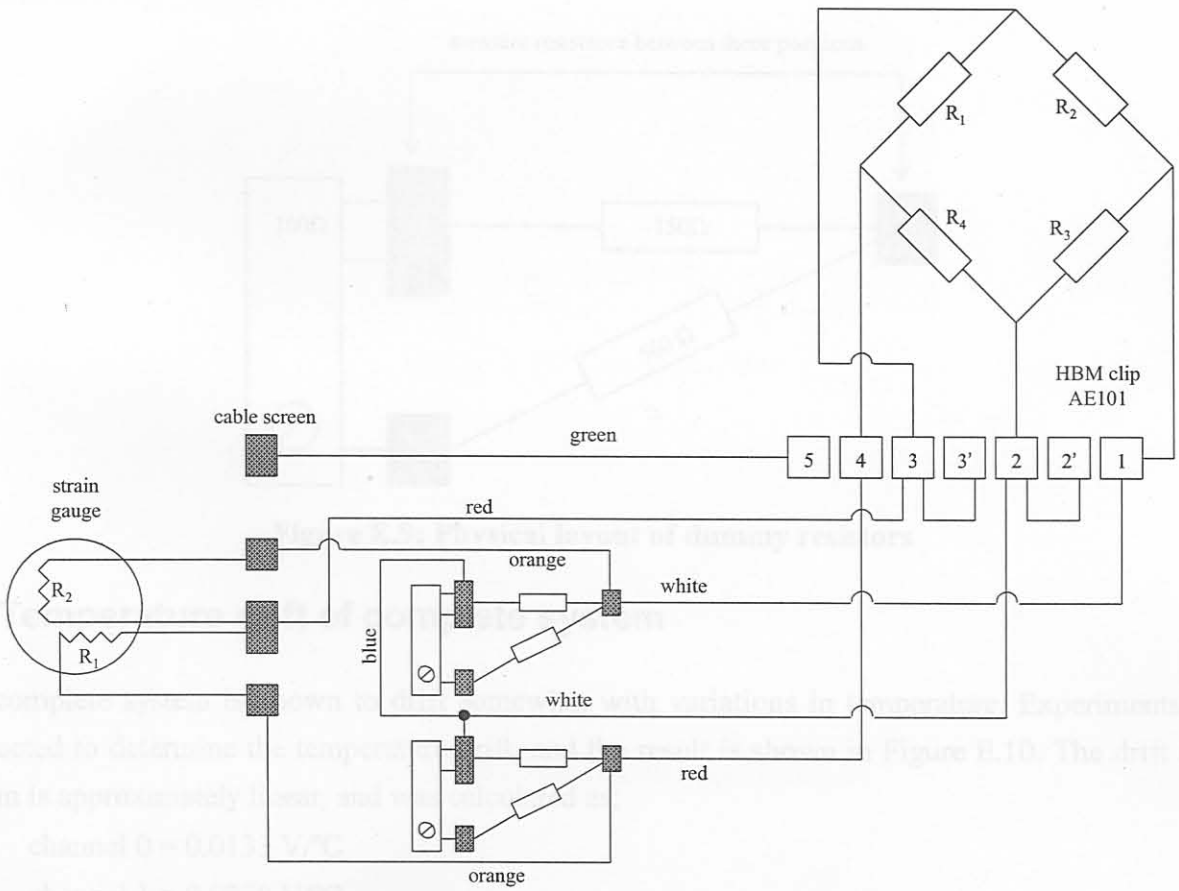


Figure E.7: Detailed per-channel bridge configuration

E.3.5 Dummy gauge configuration

The dummy gauges must be set to the ***exact*** same resistance as the strain gauges. The resistance of the strain gauges should be measured with their cables included. Use the multi-turn potentiometer to change each dummy resistance to the exact same value. The theoretical configuration of the system is shown in Figure E.8, and the physical configuration is depicted in Figure E.9. Measure the dummy resistances with a multimeter over the dummies as shown in Figure E.9. After setting the dummy resistances to the correct values, the real gauges can be connected as shown in Figure E.7. Switch the system main power on and finally balance the bridges with the AE101 amplifiers for each channel (see user manual for HBM clip).

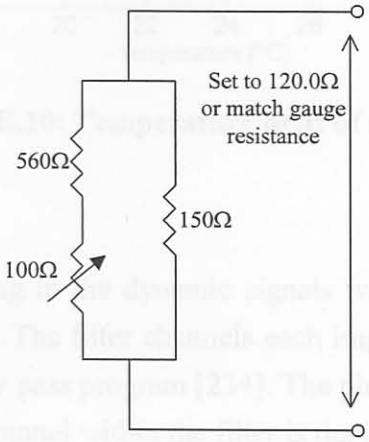


Figure E.8: Theoretical layout of dummy resistance

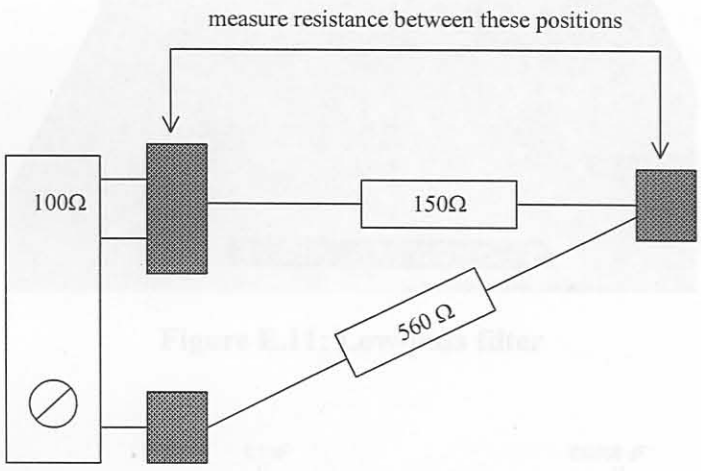


Figure E.9: Physical layout of dummy resistors

E.4 Temperature drift of complete system

The complete system is known to drift somewhat with variations in temperature. Experiments were conducted to determine the temperature drift, and the result is shown in Figure E.10. The drift of the system is approximately linear, and was calculated as:

- channel 0 = 0.0133 V/°C
- channel 1 = 0.0278 V/°C
- channel 2 = 0.0022 V/°C

The effects of temperature are compensated for (refer Chapter 5).

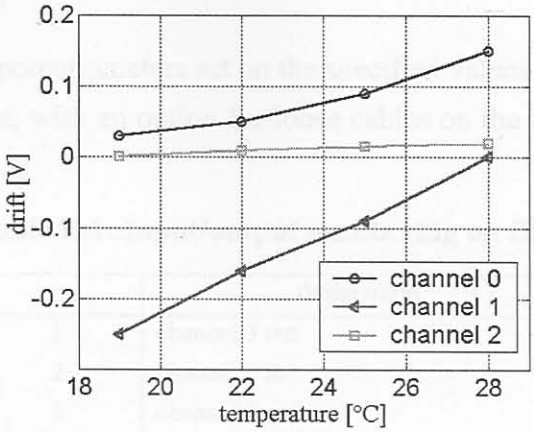


Figure E.10: Temperature drift of system

E.5 Low-pass filter

The low-pass filter to prevent aliasing in the dynamic signals was built in-house, and is part of the complete strain measurement system. The filter channels each implements a 4th order Chebyshev type filter designed with the FilterLab Low pass program [234]. The physical layout of the filter is shown in Figure E.11. The circuit of a single channel within the filter is depicted in Figure E.12:

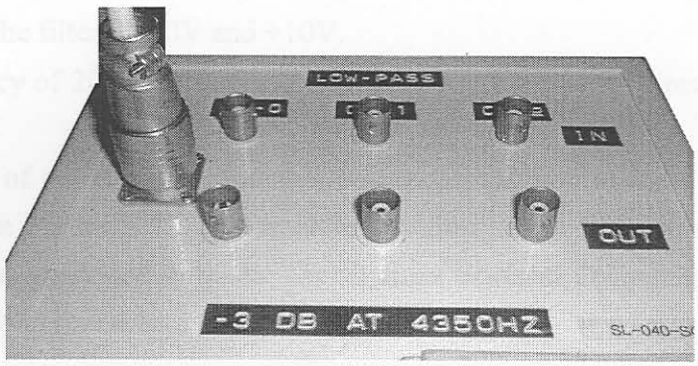


Figure E.11: Low-pass filter

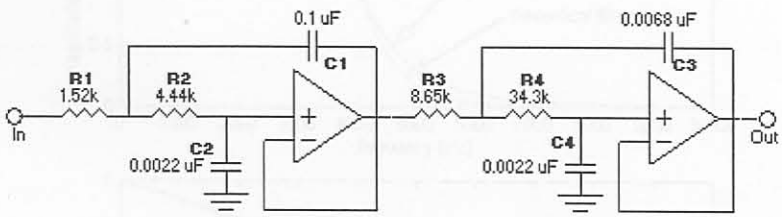


Figure E.12: Filter circuit (one channel)

Herewith some basic details of the filter:

- Three identical channels were built to pass DC - 4000Hz.
- The -3dB point is reached at approx. 4350Hz.

- All components are standard:
  LM324 op-amps
  Standard capacitors
  Resistors are standard potentiometers set on the specified values.
- BNC connectors in and out, with an option for loose cables on the right of the panel. The numbering is as shown in Table E.1:

Table E.1: Input/output numbering on filter

	designation
1	channel 3 out
2	channel 3 in
3	channel 2 out
4	channel 2 in
5	channel 1 out
6	channel 1 in
7	not used
8	analogue ground (0 V)

- A custom-built power supply is supplied with the unit, and is connected with the special 7-pin socket connector to the unit.
- The power supply supplies a 0 V, +12V and -12V DC to the unit from 220V 50 Hz AC power.
- Batteries can also power the unit.
- Maximum range of the filter is -10V and +10V.
- A sampling frequency of 20kHz is suggested when using the filter to prevent aliasing in the DC - 4000 Hz band.

The real characteristics of the channels, together with a comparison to the theoretic characteristic of the circuit, are shown in Figure E.13. The transfer function of the filter was constructed by using a wideband random signal generated with the Siglab from Spectral Dynamics. The transfer was also measured with the Siglab in real time.

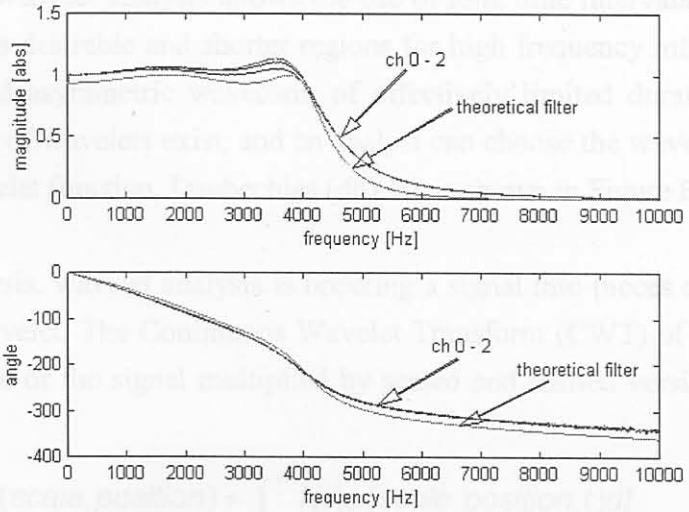


Figure E.13: Filter characteristics

F. Wavelet analysis

F.1 Introduction

The wavelet transform is a relatively new method of signal processing that has been applied to many engineering studies with great success. Recent studies also proved that wavelet analysis could be utilised for monitoring of machining processes [161-163]. The success of the wavelet transform is generally attributed to the natural shape of the wavelet, which is said to be more descriptive of natural phenomena than the sine and cosine functions used in Fourier analysis. Signals with sharp and sudden changes might be better analysed with an irregular wavelet than with a smooth sinusoid. Wavelet analysis is capable of revealing aspects of data that other signal analysis techniques miss, like trends, breakdown points, discontinuities in higher derivatives and self-similarity [266]. In this research, wavelet packet analysis was used to generate features that may show consistent trends towards tool wear. In this appendix, the basic principles of wavelet analysis are discussed.

F.2 Wavelet analysis background

Fourier analysis breaks a signal down into constituent sinusoids of different frequencies, which transforms our view of the signal from the time to the frequency domain. The drawback of Fourier analysis is that the time information is lost, which may be important if the signal contains non-stationary characteristics. This drawback may be overcome by the Short-Time Fourier Transform (STFT), but is of limited precision and not very flexible. These methods are based on windowing the signal and analysing each short time window separately. The signal can then be mapped onto a two-dimensional display of time and frequency [266].

To overcome the limitations of the STFT, wavelet analysis is based on a windowing technique with variable-sized regions. Wavelet analysis allows the use of long time intervals where more precise low frequency information is desirable and shorter regions for high frequency information. A wavelet is a sometimes irregular and asymmetric waveform of effectively limited duration that has an average value of zero. A variety of wavelets exist, and an analyst can choose the wavelet that suits his application best. A typical wavelet function, Daubechies (db) 10, is shown in Figure F.1.

Similar to Fourier analysis, wavelet analysis is breaking a signal into pieces of shifted and scaled versions of the original wavelet. The Continuous Wavelet Transform (CWT) of a function $f(t)$ is defined as the sum over all time of the signal multiplied by scaled and shifted versions of the wavelet function ψ [266]:

$$C(scale, position) = \int_{-\infty}^{\infty} f(t)\psi(scale, position, t)dt \qquad \text{(F.1)}$$

The result of the CWT is many wavelet coefficients C, which are functions of scale and position.

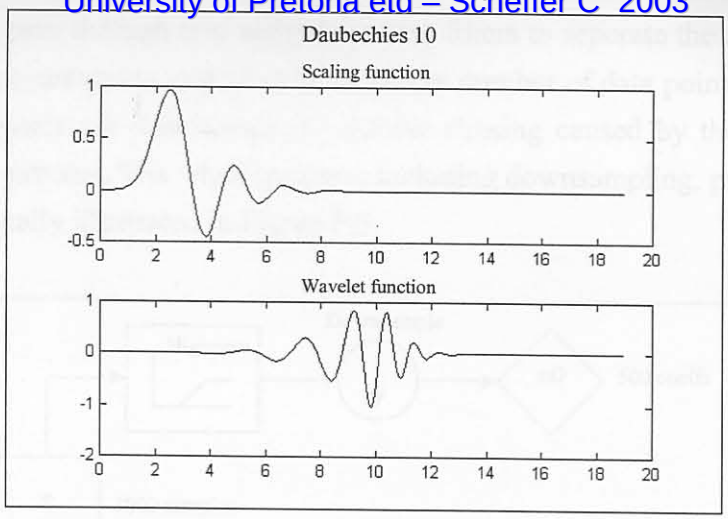


Figure F.1: Typical wavelet function, the db10

F.3 Scaling and shifting

Scaling a wavelet means stretching or compressing it, which is denoted by the scale factor a. The smaller the scale factor, the more "compressed" the wavelet, therefore the scale factor is related to the frequency of the signal. In wavelet analysis the low and high frequency contents of the signal are referred to as Approximations (A) and Details (D), respectively:

- Low scale / Compressed wavelet / High frequency / Details.
- High scale / Stretched wavelet / Low frequency / Approximations.

Shifting a wavelet simply means delaying or hastening its onset. For example, delaying wavelet function $\psi(t)$ by k will be represented by $\psi(t-k)$ [266].

F.4 The Discrete Wavelet Transform (DWT)

If the scales and positions are chosen based on powers of two, the analyses are much more efficient and just as accurate. This is called the Discrete Wavelet Transform (DWT). An efficient way to implement the DWT is by using filters. The filtering process is actually very complex, and will only be discussed in principle. At its most basic level, the filtering process can be illustrated as shown in Figure F.2.

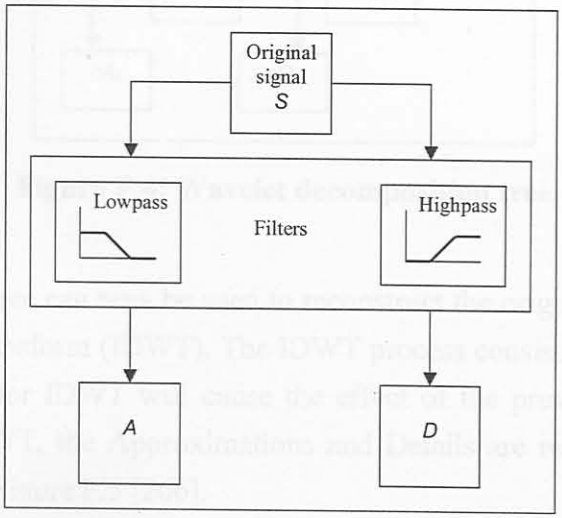


Figure F.2: DWT at its most basic level, using filters

The original signal, S, pass through two complementary filters to separate the approximations and the details. This causes computation to end up with twice the number of data points of the original signal. To correct this, the signals are downsampled, and the aliasing caused by the downsampling is accounted for later in the process. The whole process, including downsampling, produces the DWT coefficients as diagrammatically illustrated in Figure F.3.

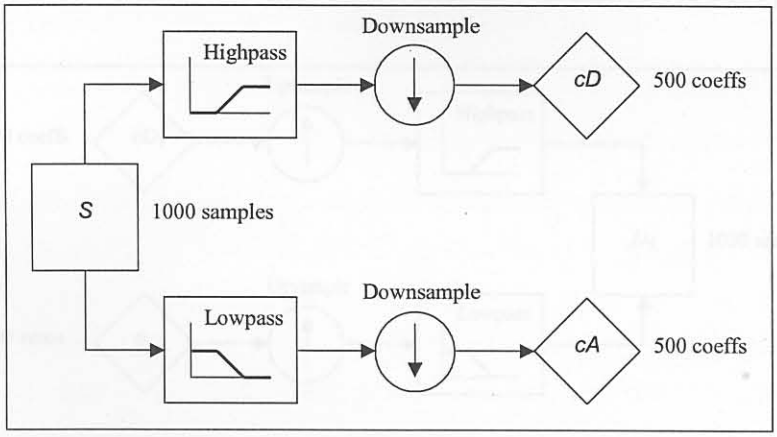


Figure F.3: DWT with downsampling

F.5 Multiple Level Decomposition and Reconstruction

The decomposition process can be iterated, with successive approximations being decomposed in turn, so that one signal is broken down into many lower-resolution components. This is called the wavelet decomposition tree, which in theory can be calculated to an infinite level, but in practice can only continue until the details consist of a single data point. In most cases, the optimal level of decomposition can be calculated based on an energy approach, such as entropy. The wavelet decomposition tree, as shown in Figure F.4, illustrates the multiple-level decomposition diagrammatically [266].

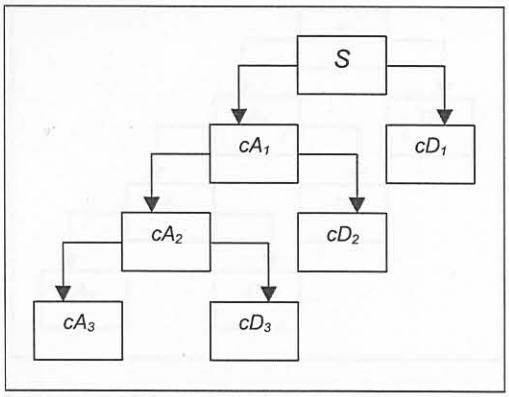


Figure F.4: Wavelet decomposition tree

The wavelet decomposition tree can now be used to reconstruct the original signal, which is called the Inverse Discrete Wavelet Transform (IDWT). The IDWT process consists of upsampling and filtering. A specific choice of filters for IDWT will cause the effect of the previously mentioned aliasing to 'cancel out'. During the IDWT, the Approximations and Details are reconstructed with the wavelet coefficients, as illustrated in Figure F.5 [266].

The original signal can now be reconstructed with a number of combinations from the Approximations and Details on the decomposition tree, for example:

$$\begin{aligned} S &= A_1 + D_1 \\ &= A_2 + D_2 + D_1 \\ &= A_3 + D_3 + D_2 + D_1 \end{aligned} \tag{F.2}$$

The wavelet tree with reconstructed components is illustrated in Figure F.6 [266].

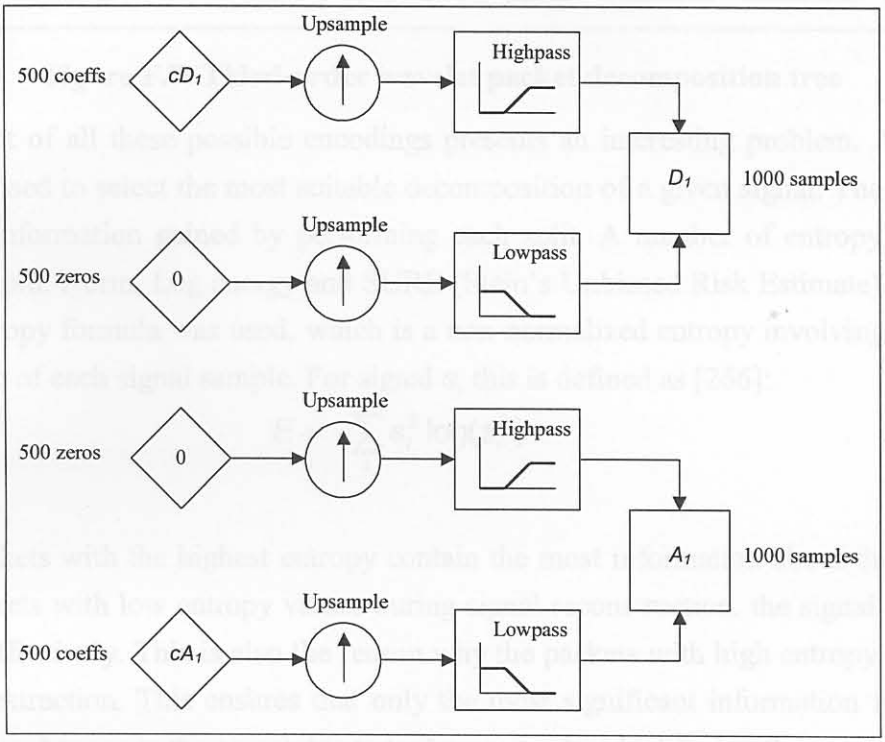


Figure F.5: Reconstruction of Approximations and Details

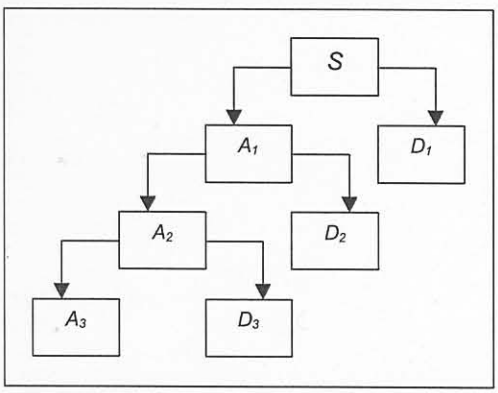


Figure F.6: Wavelet tree with reconstructed Approximations and Details

F.6 Wavelet Packet Analysis

The wavelet packet method offers a wider range of possibilities for signal analysis. With normal wavelet analysis, only the approximations are split in every step. In wavelet packet analysis, the details as well as the approximations are split in every step, illustrated in Figure F.7 [266]. The signal can be reconstructed using any number of packets on the wavelet packet decomposition tree, for example:

$$S = A_1 + AAD_3 + DAD_3 + DD_2 \tag{F.3}$$

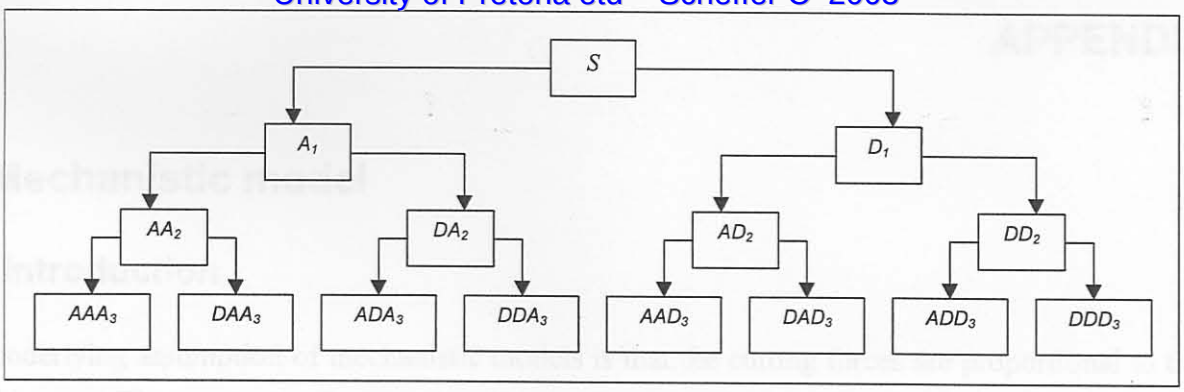


Figure F.7: Third-order wavelet packet decomposition tree

Choosing one out of all these possible encodings presents an interesting problem. An entropy-based criterion can be used to select the most suitable decomposition of a given signal. The entropy is an indication of the information gained by performing each split. A number of entropy types exist, like Shannon, Threshold, Norm, Log energy and SURE (Stein's Unbiased Risk Estimate). In this research, the Shannon entropy formula was used, which is a non-normalized entropy involving the logarithm of the squared value of each signal sample. For signal s, this is defined as [266]:

$$E = -\sum_i s_i^2 \log(s_i^2) \tag{F.4}$$

The wavelet packets with the highest entropy contain the most information about the process. By neglecting the packets with low entropy values during signal reconstruction, the signal can be de-noised or compressed effectively. This is also the reason why the packets with high entropy values were chosen for feature extraction. This ensures that only the most significant information from the signal is extracted, and any changes in the signal due to tool wear can be identified with ease.

G. Mechanistic model

G.1 Introduction

The underlying assumption of mechanistic models is that the cutting forces are proportional to the uncut chip area. The constants of proportionality are determined by the cutting geometry, material properties and machining conditions. The procedure for calculating the worn tool cutting forces is based on calculating the sharp tool forces and adding the forces caused by the worn tool. The complete model is built around a combination of models for estimating several unknowns, and also relies on a calibration procedure via selected experiments. The references for this entire appendix are the research papers by DeVor, Kapoor and their colleagues at UIUC [45,223,242,243,267,268].

G.2 Procedure for modelling the sharp tool forces in turning

G.2.1 Basic mechanistic relationship

The magnitudes of the normal and friction forces are modelled as being proportional to the process geometry characteristics that represent the amount of material being removed. These relations are:

$$F_n = K_n \cdot A_c \tag{G.1}$$

$$F_f = K_f \cdot A_c \tag{G.2}$$

where:

F_n is the normal cutting force

F_f is the frictional cutting force

K_n is the normal cutting pressure

K_f is the frictional cutting pressure

A_c is the area of chip load

G.2.2 Force transformation relations

For turning, an equivalent oblique cutting geometry can be obtained by taking the nose radius of the tool into account. The associated geometrical parameters can be obtained by integrating along the curved cutting edge. The local tool cutting forces can be determined by:

$$\begin{bmatrix} F_c \\ F_t \\ F_a \end{bmatrix} = [A] \cdot F_n + [B] \cdot F_f \tag{G.3}$$

where:

$$[A]=\begin{bmatrix}\cos(\alpha_{en})\cos(i_e)\\ \cos(\alpha_{en})\sin(i_e)\sin(\gamma_{eL})-\sin(\alpha_{en})\cos(\gamma_{eL})\\ -\cos(\alpha_{en})\sin(i_e)\cos(\gamma_{eL})-\sin(\alpha_{en})\sin(\gamma_{eL})\end{bmatrix} \quad (G.4)$$

$$[B]=\begin{bmatrix}\sin(\eta_c)\sin(i_e)+\cos(\eta_c)\sin(\alpha_{en})\sin(i_e)\\ \cos(\eta_c)\cos(\alpha_{en})\cos(\gamma_{eL})-[\sin(\eta_c)\cos(i_e)-\cos(\eta_c)\sin(\alpha_{en})\sin(i_e)]\sin(\gamma_{eL})\\ [\sin(\eta_c)\cos(i_e)-\cos(\eta_c)\sin(\alpha_{en})\sin(i_e)]\cos(\gamma_{eL})+\cos(\eta_c)\cos(\alpha_{en})\sin(\gamma_{eL})\end{bmatrix} \quad (G.5)$$

and:

γ_{eL} is the effective lead angle

i_e is the effective inclination angle

α_{en} is the effective normal rake angle

η_c is the chip-flow angle

For turning operations, these force components coincide with the external coordinate system:

$$\begin{bmatrix}F_x\\F_y\\F_z\end{bmatrix}=\begin{bmatrix}F_c\\F_t\\F_a\end{bmatrix} \quad (G.6)$$

G.2.3 Chip load model

To calculate the area of chip load, a numerical integration along the edge of the cutting tool intersecting with the workpiece is performed. The uncut chip thickness is also calculated during this procedure.

G.2.4 Chip flow model

The direction of chip flow is computed by dividing the round cutting edge into a large number of smaller straight edges and assuming oblique cutting on each (Stabler rule). The unified chip flow direction is then computed as the weighted average of all the individual chip flow directions, where the weighting for each is determined by the magnitude of the friction force at that edge. This process is done iteratively by initially guessing the friction force at each edge. The unified chip flow direction, measured with respect to a line on the rake face normal to the equivalent cutting edge is given by:

$$\eta_c=\frac{\sum\left|F_f(w)\cdot(\eta_c(w)+\gamma'(w))\right|}{\sum F_f(w)}-\gamma'_{eL} \quad (G.7)$$

G.2.5 Model calibration

Model calibration is achieved by conducting experiments over a range of machining conditions. Empirical equations are used to capture the effects of machining parameters on the normal and friction pressures:

$$\ln(K_n)=a_0+a_1\ln(t_c)+a_2\ln(V)+a_3\alpha_n+a_4\ln(V)\ln(t_c) \quad (G.8)$$

$$\ln(K_f)=b_0+b_1\ln(t_c)+b_2\ln(V)+b_3\alpha_n+b_4\ln(V)\ln(t_c) \quad (G.9)$$

where:

t_c is the uncut chip thickness

V is the cutting speed

α_n is the normal rake angle

G.3 Basic procedure for modelling the worn tool forces

G.3.1 Introduction

The basic assumption of this model is that for three-dimensional cutting operations, the plastic flow region grows linearly as the total wearland width of the tool increases, and this growth can be modelled independent of the cutting conditions. The width of the plastic flow region is of importance because the tool flank stresses (and consequently the cutting forces) depend on it. An integral part of determining the tool flank stresses is the development of a contact model for determining the normal stress at the tool tip. The three-dimensional cutting forces due to tool flank wear can then be determined by discretisising the rounded flank face into two-dimensional elements and adding the forces from all the elements.

G.3.2 Linear growth of plastic flow region

It has been shown that the plastic flow region on worn tools increase linearly with respect to the length of cut for a range of materials. Based on this, a piecewise linear function is proposed to model the growth of tool flank wear:

$$VB_p = c_0 + c_1 VB \text{ if } VB < VB_{cr} \text{ then } VB_p = 0 \tag{G.10}$$

where:

$$VB_{cr} = \frac{-c_0}{c_1} \tag{G.11}$$

VB is the total wearland width

VB_p is the width of the plastic flow region

c_0, c_1 are linear coefficients that can be obtained by fitting the function through experimental data. It has been shown that this model is independent of cutting conditions which means that only one wear experiment is required to obtain a model for the growth of the plastic flow region.

G.3.3 Two-dimensional stress on the tool flank

The cutting force and thrust force due to tool flank wear can be determined by integrating the normal and shearing tool flank stresses over the flank wear width:

$$F_{tw} = w \int_0^{VB} \sigma_w(x)\mathrm{d}x \tag{G.12}$$

$$F_{cw} = w \int_0^{VB} \tau_w(x)\mathrm{d}x \tag{G.13}$$

where:

F_{tw} is the thrust force due to tool flank wear

F_{cw} is the cutting force due to tool flank wear

σ_w is the normal tool flank stress

τ_w is the shearing tool flank stress

w is the width of cut

x is the distance from tool tip

In the plastic flow region at the tool tip, the stresses are equal to the tool tip stresses, σ_o and τ_o. By applying the linear growth of the plastic flow region, the following functions can be derived:

$$\text{if } x < VB_p \begin{cases} \sigma_w = \sigma_o \\ \tau_w = \tau_o \end{cases} \text{ Plastic flow} \tag{G.14}$$

$$\text{if } x > VB_p \begin{cases} \sigma_w = \sigma_o \cdot \left(\dfrac{VB - x}{VB - VB_p} \right)^2 \\ \tau_w = \tau_o \cdot \left(\dfrac{VB - x}{VB - VB_p} \right)^2 \end{cases} \text{ Elastic contact} \tag{G.15}$$

where:

σ_o is the tool tip normal stress

τ_o is the tool tip shear stress

G.3.4 Contact model to predict the normal tool tip stress

Since plastic deformation of the workpiece occurs under the flank of the tool, the maximum effective stress in the workpiece must be equal to the yield strength of the workpiece material. It is therefore assumed that the stresses in the tool flank are limited by the yield stress of the workpiece material. The contact model presented here calculates the maximum effective stress in the workpiece to be equal to the yield strength of the workpiece material. A plane strain assumption is followed because it has been shown that the tangential stress does not play an important role. The contact model is only applied once to a workpiece material, and the value of σ_o is assumed to remain constant for all values of VB. The stresses in the workpiece at an arbitrary point A due to the load can be found by integrating a set of concentrated loads over the loaded region. The following equations can be used to obtain the effective stress in the workpiece:

$$\sigma_x = -\frac{2z}{\pi} \int_{-a}^{a} \frac{\sigma_w(s)(x-s)^2 \, ds}{\left[(x-s)^2 + z^2\right]^2} \tag{G.16}$$

$$\sigma_z = -\frac{2z^3}{\pi} \int_{-a}^{a} \frac{\sigma_w(s) ds}{\left[(x-s)^2 + z^2\right]^2} \tag{G.17}$$

$$\tau_{xz} = -\frac{2z^2}{\pi} \int_{-a}^{a} \frac{\sigma_w(s)(x-s) ds}{\left[(x-s)^2 + z^2\right]^2} \tag{G.18}$$

where:

$$\sigma_w(s) = \sigma_o \left(\frac{s-a}{2a} \right)^2 \tag{G.19}$$

and s, x and z are defined in the Figure G.1.

To obtain the effective stress σ_{eff} the Von Mises criterion can be used:

$$\sigma_{eff} = \sqrt{(\sigma_1 - \sigma_2)^2 + \sigma_1^2 + \sigma_2^2} \tag{G.20}$$

where:

$$\sigma_{1,2} = \frac{\sigma_x + \sigma_z}{2} \pm \sqrt{\left(\frac{\sigma_x - \sigma_z}{2} \right)^2 + \tau_{xz}^2} \tag{G.21}$$

To find the maximum effective stress in the workpiece, the workpiece is divided into small square elements. The effective stress is calculated at all four corners of the squares and the maximum effective stress is the maximum of all these calculated values. To find the normal stress that causes the maximum effective stress to be equal to the workpiece yield strength, an iterative approach is used. A value for normal stress is chosen and the maximum effective stress is calculated. When the maximum effective stress is equal to the yield strength, the correct value for the normal stress is found.

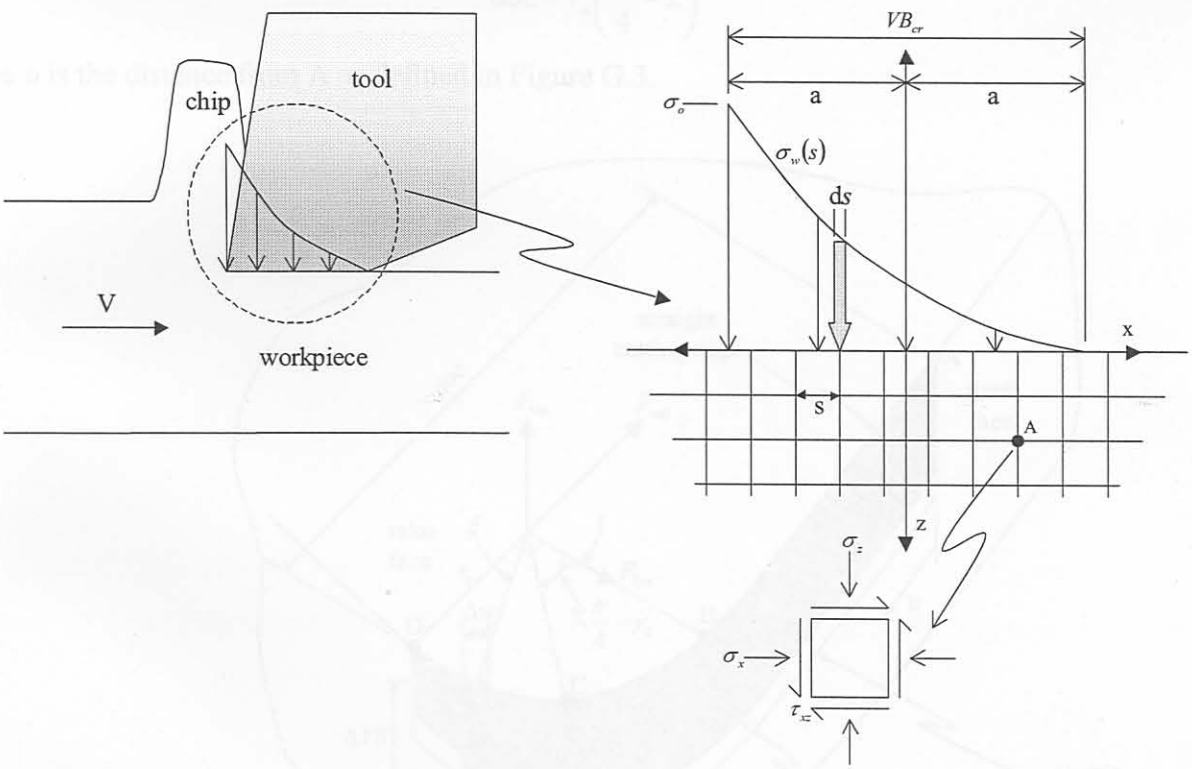


Figure G.1: Stress distribution on workpiece

G.3.5 Prediction of shearing tool tip stress

The shearing tool tip stress is found by assuming that the friction on the flank of the tool is adhesive in nature. Thus, the calculated value for the normal stress can be used to determine the shearing tool tip stress, using:

$$\tau_o = \frac{\sigma_o}{1+2\pi} \tag{G.22}$$

G.3.6 Calculating the three-dimensional forces

Once the tool tip stresses are known, the tool flank stresses can be found as described earlier. The three-dimensional forces are found by dividing the tool flank into small two-dimensional elements. For the straight edge section, the width of each two-dimensional element is found by choosing the number of elements, N :

$$\Delta w = \frac{doc - r_n}{N\cos(\gamma_1)} \tag{G.23}$$

where:

Δw is the width of the element

doc is the depth of cut

r_n is the tool nose radius

γ_1 is the lead angle of the tool

The individual element's VB values are given by:

$$\Delta VB = \frac{(u + 0.5\Delta w)(VB - 0.5 \cdot VB)}{doc + r_n\left(\frac{\pi}{4} - 2\right)} \tag{G.24}$$

where u is the distance from A as defined in Figure G.3.

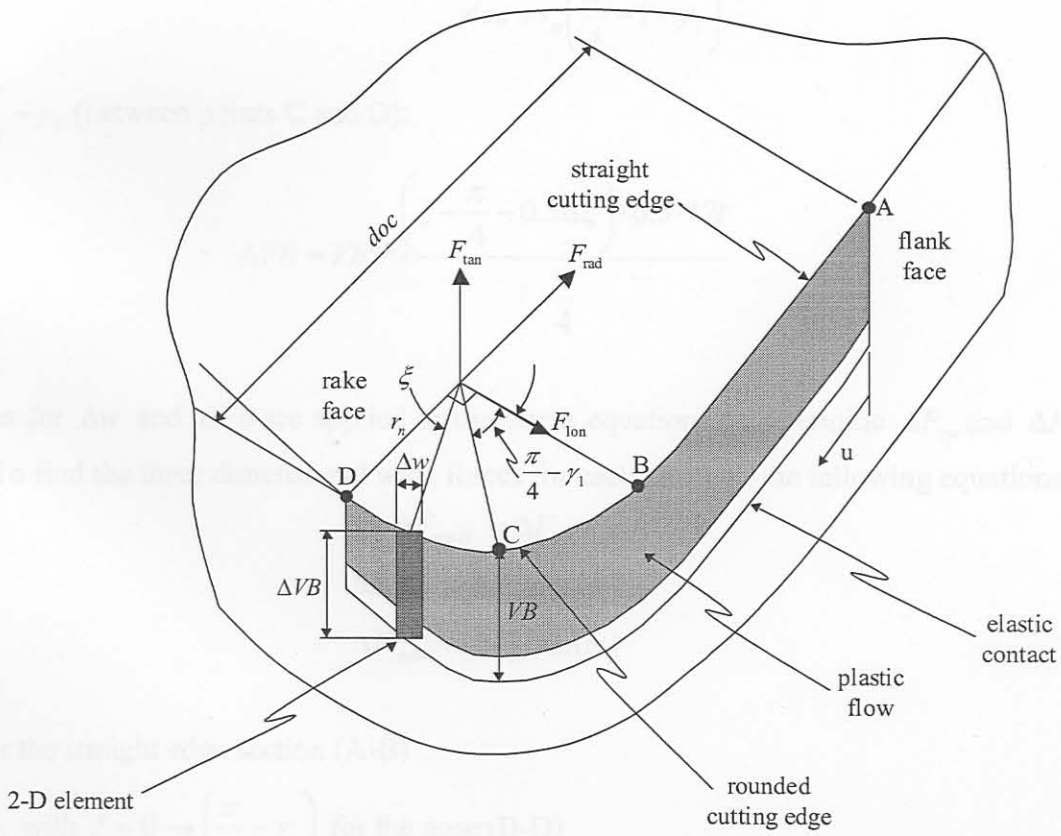


Figure G.2: Analysis of tool (1)

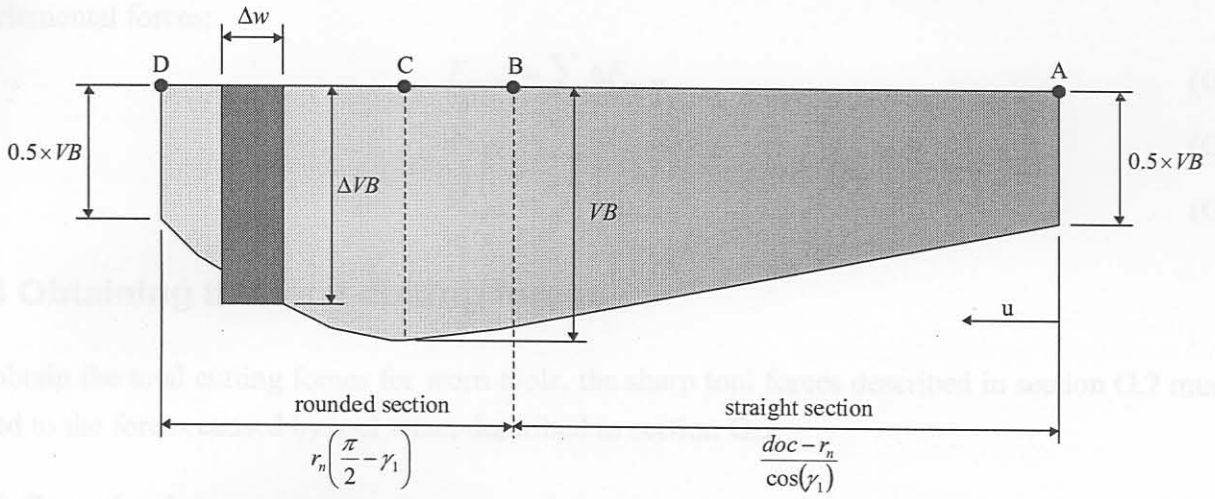


Figure G.3: Analysis of tool (2)

For the rounded nose section, Δw can be found by dividing the total included angle on the nose of the tool into $\Delta\xi$ sections:

$$\Delta\xi = \frac{\frac{\pi}{2}-\gamma_1}{N}, \Delta w = 2r_n \sin\left(\frac{\Delta\xi}{2}\right) \tag{G.25}$$

For $\xi \le \frac{\pi}{4}-\gamma_1$ (between points B and C):

$$\Delta VB = 0.5\cdot VB + \frac{doc + r_n\left(\xi - 0.5\Delta\xi\right)\cdot 0.5\cdot VB}{doc + r_n\left(\frac{\pi}{4}-1-\gamma_1\right)} \tag{G.26}$$

For $\xi > \frac{\pi}{4}-\gamma_1$ (between points C and D):

$$\Delta VB = VB - \frac{\left(\xi - \frac{\pi}{4} - 0.5\Delta\xi\right)\cdot 0.5\cdot VB}{\frac{\pi}{4}} \tag{G.27}$$

The values for Δw and ΔVB are applied in the stress equations to determine ΔF_{cw} and ΔF_{tw} for each element. To find the three dimensional wear forces for each element, the following equations apply:

$$\Delta F_{\tan W} = \Delta F_{cw} \tag{G.28}$$

$$\Delta F_{lonW} = \Delta F_{tw}\cos(\psi) \tag{G.29}$$

$$\Delta F_{radW} = \Delta F_{tw}\sin(\psi) \tag{G.30}$$

where:

$\psi = \gamma_1$ for the straight edge section (A-B)

$\psi = \xi + \gamma_1$ with $\xi = 0 \rightarrow \left(\frac{\pi}{2}-\gamma_1\right)$ for the nose (B-D)

Once these forces are known, the total cutting forces due to tool wear can be calculated by adding all

the elemental forces:

$$F_{\tan W} = \sum \Delta F_{\tan W} \tag{G.31}$$

$$F_{lonW} = \sum \Delta F_{lonW} \tag{G.32}$$

$$F_{radW} = \sum \Delta F_{radW} \tag{G.33}$$

G.4 Obtaining the total cutting forces

To obtain the total cutting forces for worn tools, the sharp tool forces described in section G.2 must be added to the forces caused by tool wear, described in section G.3.

G.5 Conclusion

Despite the fact that the model described in this appendix seems to be the most novel approach up to date for calculating worn tool forces using basic theoretical principles and combining experimental data, the implementation of the model proved unsuccessful due to reasons described in Chapter 6.

H. The Self-Organising Map (SOM)

H.1 Introduction

The Self-Organising Map (SOM), developed by Kohonen [212], is a fairly new and effective software tool for data analysis. The SOM has been implemented successfully in numerous applications in fields such as process analysis, machine perception, control and communication. The SOM implements the orderly mapping of high-dimensional data onto a two-dimensional grid of neurons. Thereby the SOM is able to identify hidden relationships between high-dimensional data with simple geometric relationships that can be displayed on a single figure [269]. The SOM can be described as a NN with self-organising capabilities. Most NNs require information and interaction from the user for classification. The training of the SOM is based on unsupervised learning, which means that the data is automatically arranged without output patterns. Although the SOM was originally intended as a data visualisation tool, it can be used for data classification as well. Cho [270] describes how the SOM can be used as a high-performance classifier for noisy data sets. The SOM automatically arranges the data on a two dimensional grid of neurons where similar observations are placed close to one another and dissimilar ones further away. If the classes of the training observations are known, certain regions on the grid could be allocated for these classes.

H.2 Computation of the SOM

H.2.1 Structure of the SOM

A SOM is formed of neurons on a 2-dimensional grid. Higher dimensional grids can also be used, but their visualisation is very problematic. Each neuron i of the SOM is represented by an n-dimensional weight or reference vector $\boldsymbol{m}_i = [m_{i1}\ m_{i2}\ \ldots\ m_i]$, where n is equal to the dimension of the input vectors. The map shape is usually rectangular, but other shapes have also been used successfully. The number of neurons is fixed before training. The number of neurons affects the accuracy and the generalisation capability of the SOM. As the size of the map increases, training becomes very time consuming.

H.2.2 Neighbourhood relation

The neurons on the map are connected to adjacent neurons by a neighbourhood relationship. Immediate neighbours belong to the 1-neighbourhood N_{i1} of the neuron i. In the 2-dimensional case the neurons of the map can be arranged either on a rectangular or a hexagonal lattice. Neighbourhoods of different sizes in rectangular and hexagonal lattices are illustrated in Figure H.1 [269].

H.2.3 Initialisation

Before the training phase initial values are given to the weight vectors. The SOM is robust with respect to initialisation, but a proper initialisation allows the algorithm to converge faster to a reliable solution.

One of the three following initialisation procedures can be used [269]:

- Random initialisation, whereby the initial values of the weight vectors are selected randomly.
- Sample initialisation, whereby the initial values of the weight vectors are selected based on samples from the training data.
- Linear initialisation, whereby the initial values of the weight vectors are generated linearly from the lowest to the highest value of the training data.

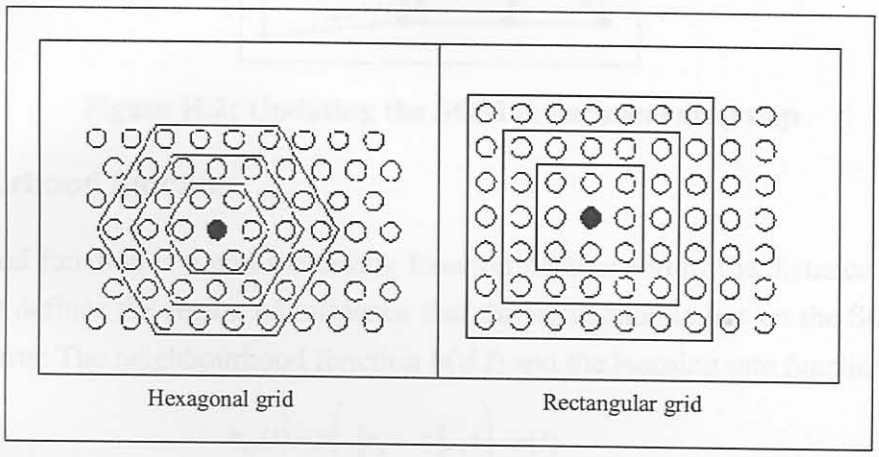


Figure H.1: Neighbourhood structures

H.2.4 Training the SOM

The computation of the SOM is a nonparametric, recursive regression process. In each training step, one sample vector $\boldsymbol{x}$ from the input data set is chosen randomly and a similarity measure is calculated between it and all the weight vectors of the map. The Best-Matching Unit (BMU), denoted as c, is the unit whose weight vector has the greatest similarity with the input sample $\boldsymbol{x}_i$. The similarity is usually defined by means of a distance measure, typically the Euclidian distance. Formally the BMU is defined as the neuron for which [221]:

$$\forall i,\ \|\boldsymbol{x}(t)-\boldsymbol{m}_c(t)\| \le \|\boldsymbol{x}(t)-\boldsymbol{m}_i(t)\| \tag{H.1}$$

which means that $\boldsymbol{m}_c(t)$ is the model that matches best with $\boldsymbol{x}(t)$. This is the BMU.

After finding the BMU, the weight vectors of the SOM are updated. The weight vectors of the BMU and its neighbours are moved closer to the input vector in the input space. This adaptation procedure stretches the BMU and its neighbours towards the sample vector. This is illustrated in Figure H.2, where the input vector given to the network is marked by an x [269]. The SOM update rule for the weight vector of the unit i is:

$$\boldsymbol{m}_i(t+1) = \boldsymbol{m}_i(t) + h_{c(x),i}(\boldsymbol{x}(t)-\boldsymbol{m}_i(t)) \tag{H.2}$$

where t is the index of the learning step, and learning is performed recursively for each presentation of a sample of $\boldsymbol{x}$, denoted $\boldsymbol{x}(t)$. The scalar multiplier $h_{c(x),i}$ is called the neighbourhood function, which causes similar observations to be placed in the same region on the map.

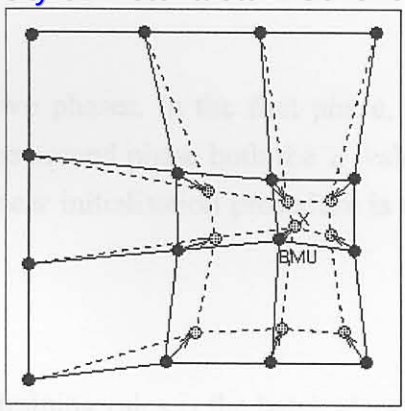


Figure H.2: Updating the SOM in each learning step

H.2.5 Neighbourhood function

The neighbourhood function is a non-increasing function of time and of the distance of unit *i* from the winning unit *c*. It defines the region of influence that the input sample has on the SOM. The function consists of two parts: The neighbourhood function $h(d,t)$ and the learning rate function $\alpha(t)$:

$$h_{ci}(t) = h\left(\|\mathbf{r}_c - \mathbf{r}_i\| , t \right) \alpha(t) \tag{H.3}$$

where $\mathbf{r}_i$ is the location of unit *i* on the map grid.

The simplest neighbourhood function is the Bubble: It is constant over the whole neighbourhood of the winner unit and zero elsewhere. Another is the Gaussian neighbourhood function. It gives slightly better results, but is computationally slower. Figure H.3 displays the two common neighbourhood functions [269].

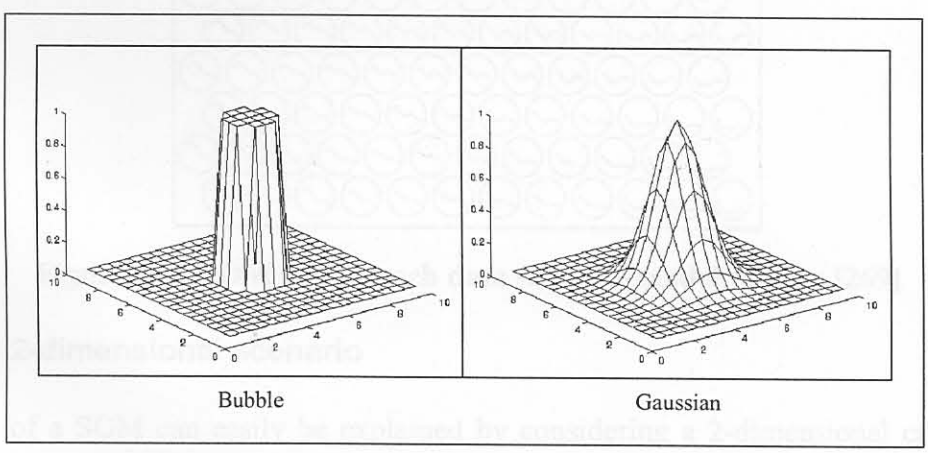


Figure H.3: Two common neighbourhood functions

H.2.6 Learning rate

The learning rate $\alpha(t)$ is a decreasing function of time. Two commonly used forms are a linear function and a function inversely proportional to time, such as:

$$\alpha(t) = \frac{A}{(t+B)} \tag{H.4}$$

where *A* and *B* are selected constants.

H.2.7 Training phases

Training is usually performed in two phases. In the first phase, relatively large initial α values and neighbourhood radii are used. In the second phase both the α values and the neighbourhood radii are small from the beginning. If the linear initialisation procedure is used, the first training phase can be skipped.

H.2.8 Batch-algorithm

Another variant of the basic SOM training rules is the batch algorithm. In this case the whole training set is used at once and after which the map is updated. The algorithm usually converges after a couple of iterations, and is much faster to calculate in MATLAB than the normal sequential algorithm.

H.3 Examples of SOMs

The SOM is best explained with simple examples. Two examples are discussed here in order to clarify the use of SOMs. For more information on SOMs, see references [221,198,199,269-274].

H.3.1 Speech data

The SOM in Figure H.4 was calculated with speech data (Finnish) [269]. The model vectors are shown on each neuron. Note that similar patterns are arranged close to one another on the map. This technique can be used for voice or word recognition.

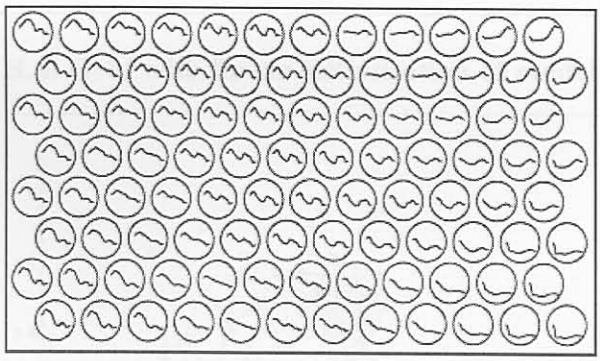


Figure H.4: SOM with speech data showing model vectors [269]

H.3.2 Simple 2-dimensional scenario

The principles of a SOM can easily be explained by considering a 2-dimensional case. Say, for instance, that a data set exists which clearly displays two clusters, as shown in Figure H.5. The one cluster may correspond to an error (*err*) situation, and the other to an acceptable (*OK*) region, for example:

$$\begin{aligned} W &= [x\ y] \\ W &= [0.3\ \ 0.3] - err \\ W &= [0.7\ \ 0.7] - OK \end{aligned} \tag{H.5}$$

Figure H.6 shows the SOM after linear initialisation of the model vectors, on a rectangular grid. Figure H.7 shows the SOM after training, which resembles a 'net' folding over the 'cloud' of data. The neurons move closer to one another where the data is dense.

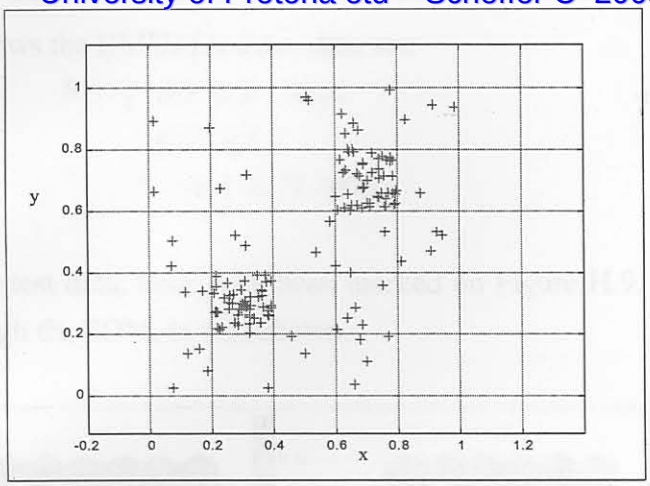


Figure H.5: Data with two clusters

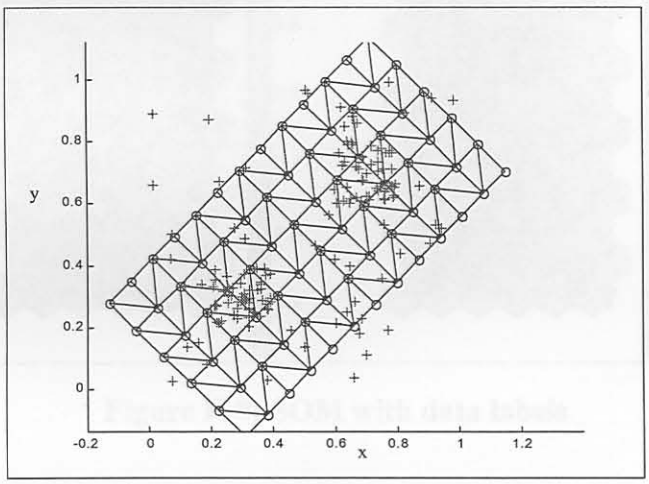


Figure H.6: Grid with linear initialisation of model vectors

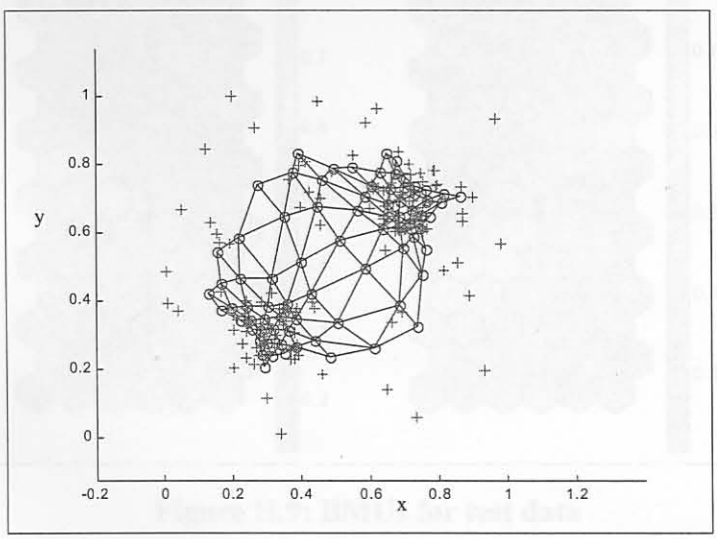


Figure H.7: Grid after training

Figure H.8 shows the SOM for variables *x* and *y*. Note that the SOM is only one entity, but a graph in the direction of each variable (dimension) can be shown. Also note that the colours of the neurons are an indication of the value of the variable, as shown on the colour map axis on the figure. The labels of the training data are also shown in Figure H.8, and it is clear that the *'err'* and *'OK'* regions have been

identified. Figure H.9 shows the BMUs for a test data set:

$$W = [\ 0.3\ \ 0.3\ \ (err)$$
$$0.5\ \ 0.5 \qquad (H.6)$$
$$0.7\ \ 0.7]\ \ (OK)$$

The three BMUs for this test data, have each been marked on Figure H.9. A trajectory, showing how the test data moves through the SOM, is also shown.

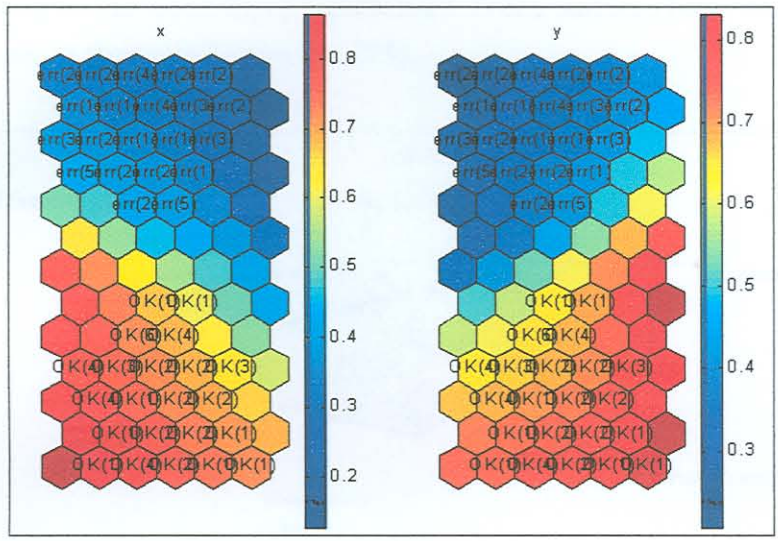


Figure H.8: SOM with data labels

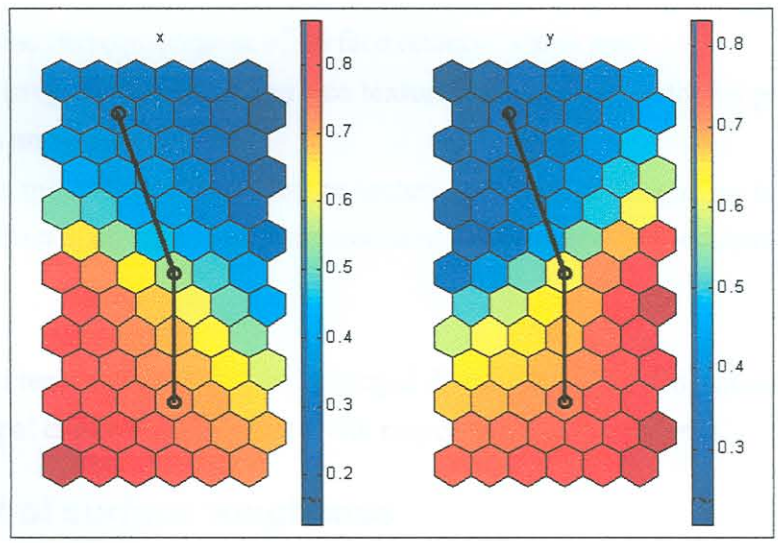


Figure H.9: BMUs for test data

I. Surface roughness analysis

I.1 Surface texture

A surface that is nominally smooth and flat will always reveal some roughness, which may vary from fine to coarse depending on the finishing operation used. Some surfaces have roughness and waviness, and may also be curved, as shown in Figure I.1 [275].

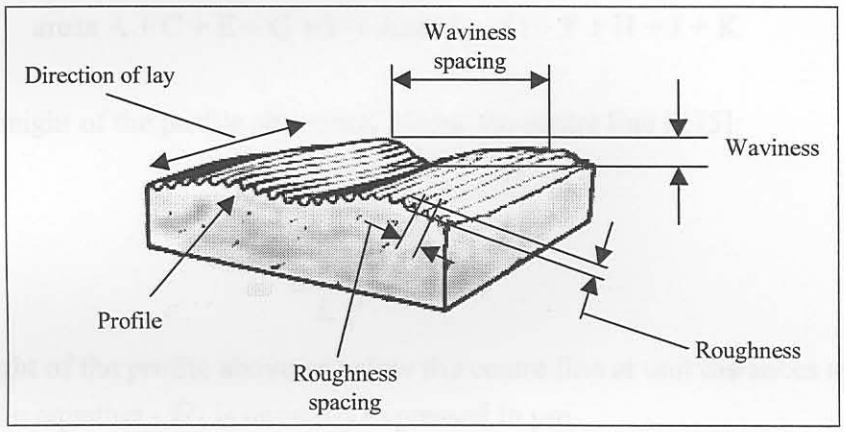


Figure I.1: Roughness and Waviness

Figure I.2 illustrates the two components of surface texture, which are:

- Roughness: The irregularities in the surface texture that is inherent to the production process, excluding waviness and errors of form.
- Waviness: This is the component of surface texture upon which roughness is superimposed. Waviness may result from such factors such as machine or workpiece deflections, vibrations, chatter or heat treatment.

Each pattern is characterised by the lay (the principal direction of the predominant surface pattern), the spacing of the principal crests, and in height (with respect to a reference line).

I.2 Assessment of surface roughness

The standard method for assessing surface texture is based on traversing a stylus across the surface to produce an electrical signal, which can generate the surface profile on a chart or an average reading on a meter. Roughness average, R_a, is defined as the arithmetical average of the profile above and below the reference line throughout the prescribed sampling length. This is illustrated in Figure I.2. Surface roughness values are normally assessed as mean results of several sampling lengths taken consecutively along the surface.

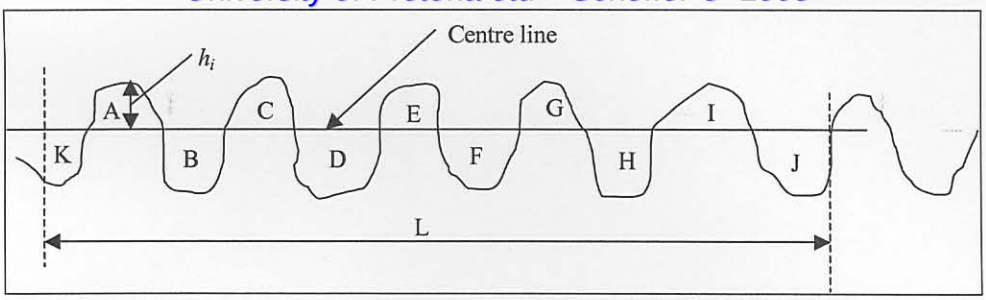


Figure I.2: Definition of the centre line

Over a length of surface L, the centre line is a line such that the sum of the areas embraced by the surface profile above the line is equal to the sum of those below the line, thus:

$$\text{areas A} + \text{C} + \text{E} + \text{G} + \text{I} = \text{areas B} + \text{D} + \text{F} + \text{H} + \text{J} + \text{K} \tag{I.1}$$

R_a is the average height of the profile above and below the centre line [275]:

$$\begin{aligned} R_a &= \frac{h_1 + h_2 + h_3 + \ldots h_n}{L} \\ &= \frac{1}{L}\int_0^L |h|\, dL \end{aligned} \tag{I.2}$$

where h is the height of the profile above or below the centre line at unit distances apart. The units of L is not added into the equation - R_a is normally expressed in μm.

R_z is the rms value for R_a:

$$R_z = \sqrt{\frac{1}{L}\int_0^L h^2(x)\,dx} \tag{I.3}$$

These are the basic parameters used for surface roughness assessment. The same type of parameters can also be derived to quantify the waviness of the surface profile.