Appendix C

Signal Processing and Analysis Methods for Speech Recognition

**Sound.** Sound is the quickly varying pressure wave within a medium. Sound is produced when the air is disturbed in some way, for example by a vibrating object. We usually refer to audible sound, which is the sensation (as detected by the ear) of very small rapid changes in the air pressure above and below a static value. The static value is the atmospheric pressure. Associated with the sound pressure wave is a flow of energy. When the rapid variations in pressure occur between about 20 and 20000 times per second sound is potentially audible by the human ear. The rate at which the magnitude of a sound signal varies (the variation in sound pressure) is called frequency. We will refer to sound waveforms as sound signals or simply signals.

**Periodic Signals.** A signal with a magnitude that varies as: \( M(t) = M(t+kT) \) \((k=0,1,2,3,...,\ t:\ time)\) is called periodic with period \( T \). A typical periodic signal is \( A \sin(\omega t + \phi) \) \((\omega: \text{angular frequency}, \phi: \text{phase difference}, A: \text{magnitude})\). Note that \( \omega \) and \( T \) are related by \( \omega = \frac{2\pi}{T} \) and that \( f=1/T \) where \( f \) is frequency. Sinusoidal signals are key to the analysis of more complex signals.

**Fourier Transform Based Analysis.** Sounds are complex as they consist of a number of frequencies which occur simultaneously. In order to calculate the frequency content of a sound signal the Fourier Transform is used:

\[
F(\omega) = \int_{-\infty}^{\infty} f(t) \cdot e^{-j\omega t} dt
\]
where $f(t)$ is the sound signal, and $\omega$ is the frequency variable (angular frequency). $f(t)$ and $F(\omega)$ are called Fourier Transform pairs and they indicate where the signal energy is in the time domain and the frequency domain respectively.

The Fourier Transform of a pure sinusoid ($f(t) = \sin \omega_0 t$) is impulse functions at $\omega = \omega_0$ and $\omega = -\omega_0$. This means that all the energy of the sound signal is concentrated at that particular frequency in the frequency domain. At the other end of the spectrum is the white noise signal where the energy of the signal is spread uniformly over the entire frequency domain.

A group of frequencies is known as a frequency band and the difference between the largest and the smallest frequency in the group is the bandwidth. Sound can be filtered to emphasize a particular band of frequencies of interest and to cut off to a varying extent the rest. Depending on which frequencies are emphasized the filter is called low-pass, high-pass, band-pass and band-stop (figure E1).

**Discrete Fourier Transform.** The sampling of continuous signals at periodic intervals has become an important operation due to the introduction and the advancement of sophisticated digital systems. Also important to the sampling operation is the frequency limitations of the engineering system processing the signal. For example phone lines have an upper limit of 5 kHz that needs to be taken into account if a voice recognition system is designed for a telephone system. Another example is the Sona-graph that has a sampling frequency of 32 kHz.

The sound signal is sampled every $T$ seconds such that $f_{\text{discrete}}(k) = f_{\text{continuous}}(kT)$. The sampling theorem says that a continuous function can be completely reconstructed if its bandwidth is less than half of the sampling frequency (Nyquist criterion). If that is not the case a phenomenon called aliasing occurs. The phenomenon of aliasing can be likened to the wheels of a slowly forward moving vehicle which appear to be rotating backwards due to the human eye reaction time being about 1/10 of a second. Analogous to this theorem is the frequency sampling theorem which states that if a function (signal) is time limited, that is $f(t) = 0$ for $|t| > T_N$, then its Fourier Transform will exist at distinct sample frequency points $n\pi/T_N$. So, if a time function is sampled uniformly in time its Fourier Transform is a periodic function, and if a time function is periodic its Fourier Transform is a sampled function that exists only at distinct frequency intervals. In a practical application only a fixed number of data can be manipulated.
For the calculation of the Sound Spectrum through the means of digital computers the Discrete Time Fourier Transform (DFT) is utilized:

\[ F(n\Omega) = \sum_{k=0}^{N-1} f(kT) \cdot e^{-j\Omega nk} \]

where \( f(kT) \) is the sampled signal, using \( N \) number of sample values and \( T \) as the sampling time interval. There are only \( N \) distinct values of \( \Omega \) computable. Also note that \( F(n\Omega) \) is a complex function having a magnitude (amplitude) and a phase. The frequency resolution of the transform is related to the number of points used in the transform and the sampling frequency by \( f_{\text{res}} = f_{\text{sam}}/2N \).

Another issue that needs to be addressed is what is termed leakage. This is inherent in the discrete Fourier Transform because of the required time domain truncation. The truncation of a periodic function at other than a multiple of the period results in a sharp discontinuity in the time domain, or equivalently results in side-lobes in the frequency domain. These side-lobes are responsible for the additional frequency components which are called leakage. To solve this problem a window \( W(k) \) is introduced such that the new time function is now \( f^w(k) = f(k) \cdot W(k) \). An example of this is the Hamming window that was used for this study.

The Fast Fourier Transform (FFT) is a computational technique developed to reduce the number of mathematical operations in the evaluation of the DFT. These are based on \( N = 2^m \) data points and this converts the Discrete Fourier Transform to a useful form. The Sande-Tukey or decimation in Frequency algorithm and the Cooley-Tukey or decimation in time algorithm are the most notable.

**Signal Processing and Analysis Methods.** There are two fundamental signal processing approaches to speech spectral analysis: filter bank and linear predictive methods. The sonagraph uses the first of the two methods and is based on Fourier Transforms. The LP methods are based on Auto-Regressive models of the sound signal.

**Sonagraph Analysis - Filter Bank Method.** The Filter Bank method employs an array of Band-pass filters that are centered on different frequencies and they cover the entire spectrum of interest for the analysis. So the sound signal is converted to \( n \) different signals with each one
containing the frequencies of the signal that are in the frequency band of the particular Band-pass filter n. A less expensive method for implementing the Filter Bank Method uses the concept of the Short-Time Fourier Transform. The Short-Time Fourier Transform is the Fourier Transform of Short-Time segment of the total sound signal. If the length of the time segment is large (relative to the signal periodicity - frequency content) then as was described above the STFT gives good frequency resolution. If the length is small then it gives poor frequency resolution but a good representation of the gross spectral estimate.

So, the way the sonagraph operates is:

a. It blocks the signal into frames (short time segments) of equal size that are partially overlapping.

b. It performs a Short-term Fourier Transform on each frame. Each frame is about 30 milliseconds long in theory.

c. The signal is passed through a filter Bank type structure to clean the frequency information

d. The variation on the frequency spectrum is analyzed to determine characteristic patterns evident in the sound.

**LPC Analysis.** Linear Predictive Analysis is based on Auto Regressive models of the sound signal. LPC is the auto-regression of the auto-correlation of a signal where the waveform is analyzed in its entirety taking all of its properties into account as opposed to analyzing characteristics such as frequency or time spectra independently. As the resolution of frequency/time analyses is not fine enough to separate individuals the LPC technique permits components of the signal other than frequency to be analyzed more closely.

The value of the signal at time k is related to its past n values with the equation:

\[ s(k) = a_{k-1}s(k-1) + a_{k-2}s(k-2) + \ldots + a_{k-n}s(k-n) \]

The coefficients \( a_k \) are called the Linear Predictive Coding Coefficients or Parameters and the number of coefficients for this particular example is n.

The LPC processor for sound consists of the following steps:

- The signal is smoothed through the use of a low order digital filter. This process is called Preemphasis.
• It is blocked into frames (time segments) with partial overlap between frames.
• A window is applied to each frame to minimize signal discontinuities at the beginning and end of each frame.
• Each frame of windowed signal is autocorrelated:
  \[
  r_{ij}(m) = \sum_{n=0}^{N-1-m} x_i(n)x_j(n+m)
  \]
  where \( m = 0,1,\ldots,p \)

where \( p \) is the order of the LPC analysis.
• Next is LPC Analysis which converts the \( p+1 \) autocorrelations to an LPC parameter set (using the Durbin algorithm). That includes the LPC coefficients, the reflection or (PARCOR) coefficients, the log area ratio coefficients, the cepstral coefficients or any desired transformation of these sets.
• The coefficients can be analyzed to provide insight to the nature of the signal. This analysis can be performed by various pattern recognition techniques. Examples of these techniques are the Vector Quantization method, Neural Networks, Genetic Algorithms and standard linear system identification techniques. The pattern recognition analysis for this thesis was performed using perhaps the best pattern recognizer known thus far: the human eye.

Appendix References: