METHODS FOR DIGITAL SIGNAL PROCESSING OF SPEECH SIGNALS

B. Jakimovski, D. Gligoroski

Institute of Informatics, Faculty of Natural Sciences and Mathematics,
Sts. Cyril and Methodius University,
Arhimedova bb, PO BOX 162, Skopje, Macedonia
boroj@ii.edu.mk

Abstract: The goal of this research is to investigate the ways of processing the speech signals. Speech signals represented as an array of samples undergo a series of DSP techniques for extracting the key measures form the signal. Further more this signals are used into the process of speech recognition as the first part of the process. Here we will give an overview of these techniques and demonstrate them using the MATLAB and SIMULINK package. The final goal is to build a parameters of the speech recognition engine for Macedonian language.

Keywords: speech recognition, dsp, speech signal, cepstral coefficients, linear prediction

1. Introduction

Parameterization of an analogue speech signal is the first step in the speech recognition process. Several signal analysis techniques are used. These algorithms are intended to produce a "perceptually meaningful" parametric representation of the speech signal: parameters that emulate some of the behavior observed in the human perceptual systems. Of course, and perhaps more importantly, these algorithms are also designed to maximize recognition performance.

In speaker independent speech recognition, main effort is placed on developing descriptions that are somewhat invariant to changes in the speaker. Parameters that represent characteristic spectral energies of the sound, rather than details of the particular speaker's voice, are desired. Therefore signal modeling is used which means converting sequences of speech samples to observation vectors representing events in a probability space.

Signal modeling can be subdivided into four basic operations: spectral shaping, spectral analysis, parametric transformation and statistical modeling each depending on the previous ones results. The first three operations are straightfor-

ward problems in digital signal processing. The last task, however, is often divided between the signal modeling system and the speech recognition system.

The *spectral shaping phase* includes the pre-emphasis filtering of the input signal from the A/D converter in order to boost the signal spectrum towards the higher frequencies. The *spectral analysis phase* includes various dsp methods for extracting parameters from the signal. We will dedicate the biggest part from this paper into this phase. The *parametric transform phase* adds extra parameters to the parametric vector. The *statistical modeling phase* deals with modifications of the raw parameter vector giving better quality representation of the signal characteristics.

2. Spectral Analysis

Before going into the spectral analysis we have to define few other fundamental concepts.

2.1 Signal power

A power of a signal or more precisely a measure of the power of the system is defined with the following function:

$$P(n) = \frac{1}{N_s} \sum_{m=0}^{N_s-1} \left(w(n) s\left(n - \frac{N_s}{2} + m \right) \right)^2 \quad (1)$$

where Ns is the number of samples used to compute the power, s(n) denotes the signal, w(n) denotes a weighting function and n denotes the sample index (discrete time) of the center of the window. Rather than using power directly, many speech recognition systems use the logarithm of the power multiplied by 10, defined as the power in dB, in an effort to emulate the logarithmic response of the human auditory system.

The weighting function in Equation 4 is referred to as a window function. There are many types of windows including rectangular, Hamming, Hanning, Blackman, Bartlett, and Kaiser. Today, in speech recognition, the Hamming window is almost exclusively used.

The Hamming window is a specific case of the Hanning window for aw = 0.54. A generalized Hanning window is defined as:

$$w(n) = \frac{a_w - (1 - a_w)\cos(2\pi n/(N_s - 1))}{\beta_w} \text{ for } 0 \le n \le Ns \text{ else } w(n) = 0 \quad (2)$$

The purpose of the window is to weight, or favor, samples towards the center of the window. This characteristic, coupled with the overlapping analysis discussed

next, performs an important function in obtaining smoothly varying parametric estimates.

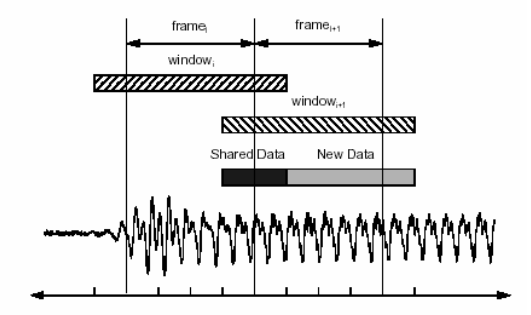

Figure 1: Window and frame representation

It can be seen from the calculation of the power that it is calculated on window basis (only the samples from the window are included into the calculation). Also every other parameter in the future will be calculated on window basis. The length of the window is N_s. Even though parameters will be computed over the window, they will be representing the frame characteristics. Frame duration, T_f, is defined as the length of time (in seconds) over which a set of parameters are valid. Frame duration typically ranges between 20 ms to 10 ms and window length 20ms to 30ms. Window and frame meaning can be seen in Fig. 1.

2.2 Spectral analysis

There are six major classes of spectral analysis algorithms used in speech recognition systems today. The procedures for generating these analyses are summarized in Fig. 2. It can be seen that they are grouped in three conceptual fields.

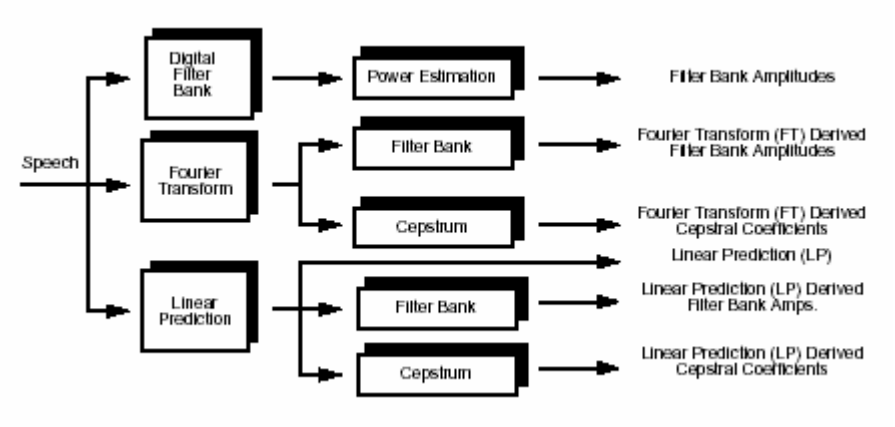

Figure 2: Classes of spectral analysis algorithms

2.2.1 Digital Filter Bank

The digital filter bank is one of the most fundamental concepts in speech processing. Experiments in human perception have shown that frequencies of a complex sound within a certain bandwidth of some nominal frequency cannot be individually identified. When one of the components of this sound falls outside this bandwidth, it can be individually distinguished. We refer to this bandwidth as the critical bandwidth. We can define a mapping of acoustic frequency, *f*, to a "perceptual" frequency scale, as follows:

$$Bark = 13\arctan\left(\frac{0.76f}{1000}\right) + 3.5\arctan\left(\frac{f^2}{7500^2}\right) \quad (3)$$

Another similar and more common in speech recognition is the *mel* scale generated by:

$$mel\ frequency = 2595\ \log 10(1 + f/700,0) \quad (4)$$

The critical bandwidth for Bark and mel scales are calculated using the following function:

$$BW_{critical} = 25 + 75\ [1 + 1.4\ (f/1000)^2]^{0.69} \quad (5)$$

Both the *Bark* scale and the *mel* scale can be regarded as a transformation of the frequency scale into a perceptually meaningful scale that is linear. The combination of these two theories gave rise to an analysis technique known as the critical band filter bank. A *critical band filter bank* is simply a bank of linear phase FIR bandpass filters that are arranged linearly along the *Bark (or mel)* scale. The bandwidths are chosen to be equal to a critical bandwidth for the corresponding center frequency.

The output of this analysis is a vector of power values (or power/frequency pairs) for each frame of data. These are usually combined with other parameters, such as total power, to form the signal measurement vector.

2.2.2 Fourier Transform Filter Bank

We have previously discussed the advantages in using non-uniformly spaced frequency samples. One of the easiest and most efficient ways to compute a non-uniformly spaced filter bank model of the signal is to simply perform a Fourier transform on the signal, and sample the transform output at the desired frequencies. The *Discrete Fourier Transform* (DFT) of a signal is defined as:

$$S(f) = \sum s(n)e^{-j\left(\frac{2\pi f}{f_s}\right)n} \quad (6)$$

where f denotes the frequency in Hz, f_s denotes the signal sampling frequency and N_s denotes the window duration in samples.

In order to discover the power magnitude of each of filter banks, we have to sample the spectrum at the frequencies given by *Bark* transformation. But since usually the spectrum is over sampled the magnitude is an average of specific elements of the spectrum falling in that bank. The calculation of the power magnitude is done using the following function.

$$S_{avg}(f) = \frac{1}{N_{os}} \sum_{n=0}^{N_{os}} w_{FB}(n) A(f + \delta f(f, n)) \tag{7}$$

In noisy environments, noise often disproportionately degrades our estimates of the low amplitude areas of the spectrum. Stated another way, we are more confident of the reliability (and repeatability) of our estimates of the high amplitude areas of the spectrum. For this reason, we often impose a limit on the dynamic range of the spectrum. This is depicted in Fig. 3. We refer to this lower limit as the *dynamic range threshold*.

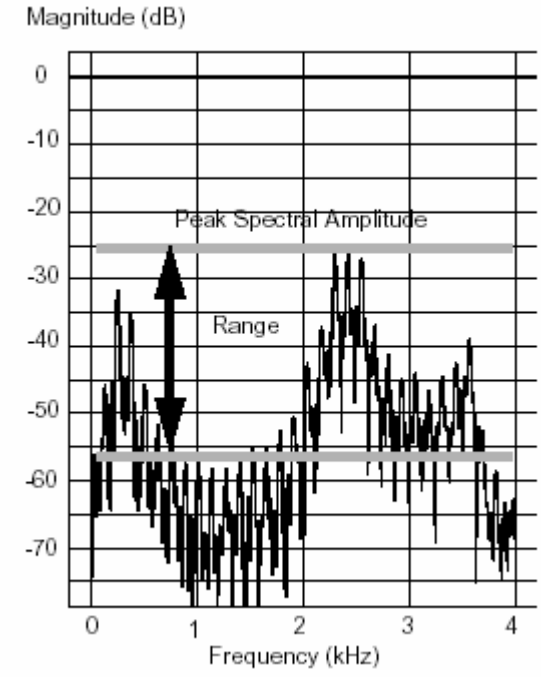

Figure 3: Dynamic range threshold

Recall that since the spectrum of the speech signal inherently drops per decade, a threshold based on low frequency energies, where the peak to valley spectral amplitude difference is large, can easily remove useful signal energy at higher frequencies. Later, we will discuss more sophisticated methods for implementing threshold of the spectrum based on parametric modeling techniques.

2.2.3 Cepstral Coefficients

The composite speech spectrum, as measured by a Fourier transform, consists of the excitation signal filtered by a time-varying linear filter representing the vocal tract shape. The process of separating the two components, often referred to as deconvolution. The frequency domain representation of this process is:

$$S(f) = G(f)\, V(f) \tag{8}$$

where $G(f)$ represents the spectrum of the excitation signal and $V(f)$ the spectrum of the vocal tract filter.

If we take the logarithm from the previous equation, we have:

$\text{Log}(\, S(f)\,) = \text{Log}(\, G(f)\,) + \text{Log}(\, V(f)\,)$

Hence in logarithm domain the excitation and the vocal tract shape are superimposed, so it is easy to separate them.

The cepstrum is defined signal derived from the log spectral magnitudes. This yealds:

$$c(n) = \frac{1}{N_s}\sum_{k=0}^{N_s-1} \log_{10}\left|S_{avg}(k)\right| e^{j\frac{2\pi}{N_s}kn}, 0 \le n \le N_s - 1 \tag{9}$$

We refer to this coefficients as *cepstral coefficients* computed via Fourier Transformation.

The low order terms of the cepstrum correspond to short-term correlation in the speech signal (smooth spectral shape or vocal tract shape). The local maxima in the higher order terms demonstrate long term correlation, or periodicity, in the waveform (excitation information). The cepstrum of an unvoiced signal does not show any periodicity. In spectral analysis for speech recognition applications, normally only the low order terms ($n < 20$) are used.

2.3 Linear Prediction Coefficients

We now turn from Fourier Transform methods based on linear spectral analysis to a class of parametric modeling techniques that attempt to optimally model the spectrum as an autoregressive process. In this section, we will discuss computation of a parametric model based on least mean squared error theory. This technique is known as linear prediction (LP).

Given a signal, $s(n)$, we seek to model the signal as a linear combination of its previous samples. Let us define our signal model as:

$$s(n) = -\sum_{i=1}^{N_{LP}} a_{LP}(i)s(n-i) + e(n) \tag{10}$$

where N_{LP} represents the number of coefficients in the model (the order of the predictor), $\{a_{LP}\}$ are defined as the *linear prediction coefficients* (predictor coefficients), and $e(n)$ represents the error in the model (the difference between the predicted value and the actual measured value). The error term should tell us something about the quality of our model. It is also possible to show that a linear prediction model effectively models the spectrum of the signal as a smooth spectrum.

Under the constraint that we would like the mean-squared error to be as small as possible (seeking a solution that gives us the minimum error energy is reasonable), the coefficients (excluding $a_{LP}(0)$) of Eq. (10) can be obtained from the following matrix equation:

$$
\begin{aligned}
&\bar{a}_{LP} = \underline{\Phi}^{-1}\bar{\phi} \\
&\textit{where}, \\
&\bar{a}_{LP} = \left[a_{LP}(1) \quad \ldots \quad a_{LP}(N_{LP})\right]^T \\
&\underline{\Phi} = \begin{bmatrix} \phi_n(1,1) & \phi_n(1,2) & \ldots & \phi_n(1,N_{LP}) \\ \phi_n(2,1) & \phi_n(2,2) & \ldots & \phi_n(2,N_{LP}) \\ \ldots & \ldots & \ldots & \ldots \\ \phi_n(N_{LP},1) & \phi_n(N_{LP},2) & \ldots & \phi_n(N_{LP},N_{LP}) \end{bmatrix} \\
&\bar{\phi} = \left[\phi_n(1,0) \quad \ldots \quad \phi_n(N_{LP},0)\right]^T \\
&\textit{and} \\
&\varphi_n(j,k) = \frac{1}{N_s}\sum_{m=0}^{N_s-1} s(n+m-j)s(n+m-k) \\
&\phi_n(j,k) = \varphi_n(0,|\, j-k\,|)
\end{aligned}
\tag{11}
$$

This solution is known as Autocorrelation Method. The last function can be also written as a autocorrelation function

$$R_n(k) = \frac{1}{N_s}\sum_{m=0}^{N_s-1-k} s(n+m)s(n+m-k) \tag{12}$$

which used in the Levinson-Durbin [1] recursion can produce LP coefficients. Also by slight modification of Eq. 12 we can achieve a dynamic range threshold.

2.3.1 LP Derived filter bank amplitudes and cepstral coefficients

Using the LP model as a background we can derive the filter bank amplitudes and cepstral coefficients using the LP spectrum.

3. Conclusion

We have presented several popular signal analysis techniques used in speech recognition systems in a common framework. As it can be seen from the presented above the next step will be to process this raw vectors of information into more useful vectors and then to statistically analyze them and build a base point for building a speech recognition system.

4. References

1. Picone , J., "Signal Modeling Techniques In Speech Recognition", *Proc. of IEEE*, vol. 81, No. 9, Sept, 1993
2. Deller, J., Hansen, J., Proakis, J., *Discrete-Time Processing of Speech Signals*, IEEE Press, New York