COMPUTATIONALLY EFFICIENT GABOR TRANSFORM AND ITS APPLICATION FOR EXTRACTING DOMINANT SPEECH SIGNAL HARMONICS

N. Venkateswaran, Kaushik Subramanian+*

*Director, WAran Research FoundaTion (WARFT), Chennai, India
+Research Trainee, WAran Research FoundaTion (WARFT), Chennai, India

ABSTRACT

In the conventional Gabor Transform, the Fast Fourier transform is used for computation of the biorthogonal function and its multiplication complexity is of $O(N^3)$, N being the number of samples. In this paper we define an efficient Gabor expansion algorithm employing the Arithmetic Fourier Transform (AFT). The proposed algorithm reduces the multiplication complexity to $O(N)$. The noisy speech signal input is first passed through a Kalman filter and the enhanced output is obtained by an iterative noise removal process. The proposed Gabor expansion is applied to the enhanced signal to extract the unique feature vector consisting of dominant harmonics and the associated phase. This is a novel proposal and it is shown that the extracted features help design computationally efficient speech recognition systems.

Index Terms— Speech recognition, Time-frequency analysis, Fourier transforms.

1. INTRODUCTION

The dynamic characteristics of the human speech have been commonly exploited in Automatic Speech Recognition (ASR) systems by using feature extraction techniques, temporal filtering and several short-term spectral representations. Most of the available techniques are aimed at extracting or improving very specific parameters of the speech signals, thus necessitating the need for a composite system designed to work under near real-time conditions. This paper describes an iterative approach for speech signal feature extraction aimed at speech recognition. The system comprises of a noise removal filter, an efficient Gabor representation followed by feature vector extraction.

The input speech signal is assumed to be corrupted by a noise factor inherent in the recording device and the transmitting channel. An iterative Kalman filter is used to initiate the noise removal process. The noise related statistical parameters are estimated and constantly updated by using the previous iterative estimate. With every iteration, the measurement noise covariance tends towards zero, producing the enhanced noiseless signal.

The Gabor representation transforms the input into a discrete set of shifted and modulated versions using the Fast Fourier Transform (FFT). To reduce the computational complexity, the Gabor coefficients are obtained through the Arithmetic Fourier Transform (AFT), which has a complexity of $O(N)$ multiplications and $O(N^2)$ additions. The enhanced speech is passed through the Gabor filter and a feature vector is determined, consisting of a set of dominant harmonics and associated phase components. In the tests conducted, the acquired feature vector was used to set a threshold in order to differentiate the words in a set of sentences containing similar sounding words (Homonyms).

This system has been one of the major research projects at the Waran Research FoundaTion. A database and associated biological neural system is being constructed to map the feature vector onto the Broca region of the brain to study the application of the proposed system for individuals with speech disabilities.

This paper is organized as follows. Section 2 describes the Kalman based noise removal filter. Section 3 deals with the Proposed Computationally Efficient Gabor transform. Section 4 presents the experimental results and Section 5 presents the conclusion.

2. KALMAN FILTER FOR NOISE REMOVAL

Let x_k and z_k denote the clean speech and noise respectively. They are represented by linear stochastic difference equations

$$x_(k+1) = Ax_k + w_k \quad (1)$$

$$z_(k+1) = Hx_k + v_k \quad (2)$$

The factors A and H represent the estimated linear coefficients, w_k and v_k are random variables and represent the process and measurement noise respectively. They are assumed to be independent of each other, white, and with normal probability distributions. In reality, there is often no a priori knowledge of the environment, i.e. whether it is white or colored noise. Under such conditions the Kalman filter provides a minimum mean-squared error estimate of the clean signal if the noise is a Gaussian process or a linear minimum mean-squared error estimate if the noise is non-Gaussian.

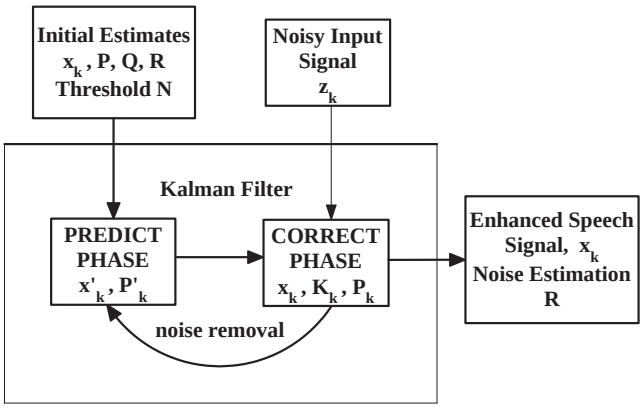


Fig.1 Kalman Filter for Noise Removal

The filter achieves good speech quality by recursively reducing the processing distortion embedded within the speech signals. To begin this iterative process an initial estimate of the statistical noise parameters is required. The noise removal filter operates by using the available noisy information z_k at the current time step to refine the previously acquired prediction $x'k$ and finally arrive at a new, more accurate state estimate of the clean speech signal, x_k, for the current time step as in Fig.1.

$$x_k = H^T.x'(k|z_k) \tag{3}$$

With each time step the filter is tuned to reduce the estimated value of P_k. The accuracy of the enhanced speech signal is a measure of the tendency of the Estimation Error Covariance, P_k towards zero.

$$P_k \rightarrow 0 \tag{4}$$

2.1. Tuning Parameters

In order to tune the system to function under real-time conditions, two parameters are calculated, an Estimation Threshold N and a Noise Covariance R.

$$P_k \rightarrow N \tag{5}$$

$$R = (K_k/P_k H^T) - HP_k H^T \tag{6}$$

K_k represents the Kalman Gain.

The threshold is preset during the initial estimation process and the filter iteration progresses and halts as this threshold is reached. Under ideal conditions $N = 0$. The new calculated value of R is a direct measure of the noise factor present within the recording devices and the transmitting channel. Use of this value during the next filter iteration increases the efficiency of the system and also helps model the noise parameters of the system related components. Selection of appropriate estimation parameters is vital in tuning the filter to perform efficiently in real time conditions.

3. PROPOSED GABOR EXPANSION FOR FEATURE VECTOR EXTRACTION

3.1. Gabor Transform using AFT

The enhanced speech signal x_k obtained from the filter is then processed by a Gabor Expansion Algorithm. The Gabor Transform was chosen due to its decomposition of the input signal into functions localized in both time and frequency, enabling us to study the signal modulation. To compute the spatio-temporal representation of the signal, a Gaussian window $h(k)$ is used. Given the input signal of length L, the expansion follows as

$$x(k) = (1/\sqrt{N}) \sum_{m=0}^{M-1} \sum_{n=0}^{N-1} a_{m,n} h(k - mN) e^{j\pi 2nk/N} \tag{7}$$

where $L = MN$

In the above equation, M and N represent the respective time and frequency domain shifts. A common method for computing the coefficients involves the multiplication of the input signal by a function which is biorthogonal to the Gaussian window. The biorthogonal function is obtained with the help of the Fourier Transform. Unfortunately, the biorthogonal function is nonlocal and the pre-multiplications involved is computationally expensive. To overcome this, the biorthogonal function is computed using the Arithmetic Fourier Transform shown in Fig.2. This efficiency can be attributed to use of the Number theoretic concept, the Mobius Inversion Formula.

$\mu(n) = 1$ if $n = 1$
$\mu(n) = (-1)^\alpha$ if $n = p_1 p_2 ... p^\alpha$
$\mu(n) = 0$ if $p_i^2 | n$ for any i

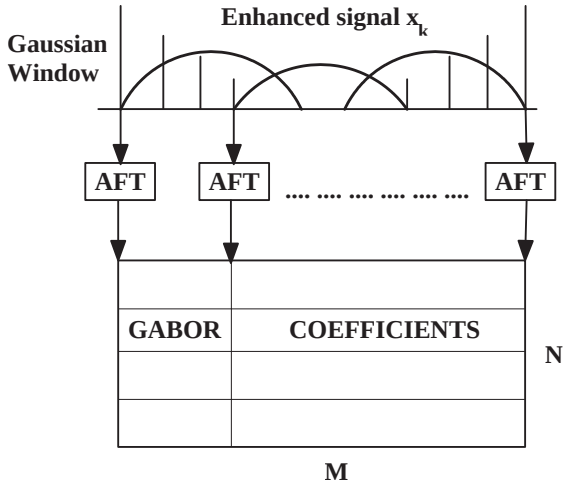


Fig.2 Gabor Coefficients computed using the AFT blocks

A Kronecker delta function is necessary for the computation of the biorthogonal matrix of the Gabor Expansion as shown below

$$\sum_{m=0}^{L-1} h(k+mN)\gamma'(k)e^{j\pi 2nk/N} = \delta_m.\delta_n \tag{8}$$

It is known that this function can be represented using the Mobius function. The Kronecker delta function is thus defined for positive integers m and n by the formula

$$\delta(m,n) = \sum_{d|(m/n)} \mu(d) = 1 if m/n = 1 \tag{9}$$

0 otherwise

The required Fourier Coefficients are computed as the E_a matrix. It is comprised of diagonal blocks, with each block computed as

$$E = \sum_{n=1}^{\lfloor L/K \rfloor} \mu(n).S(kn) \tag{10}$$

where $S(kn)$ represents the mean/averaged input signal. The 1D Gabor coefficients are computed using the Gaussian Window matrix H, the Arithmetic Fourier Transform matrix Ea and the clean speech input x_k.

$$a = (HE_a*)^{-1}.x_k \tag{11}$$

After the computation of the Fourier Transform Matrix, the coefficients are computed using the above equation.

3.2. Feature Vector Extraction

The feature vector set for the proposed system consists of locally dominant harmonics and associated phase components computed Fig.3.

$$F_v = [\delta, \psi, H(w)] \tag{12}$$

After the enhanced signal is obtained, a windowed spectrum of the speech signal is plotted and the corresponding dB levels are stored. An inverse Fourier Transform is performed on the stored data and the time-frequency expansion is initialized. The absolute value of the coefficients obtained are scaled and plotted. For a particular input speech signal, the plot reveals a unique sequence of spikes with each spike corresponding to a word as shown in the next Section.

$$[\delta] = \sum_{i=1}^{N} max(f_i) \tag{13}$$

where f_i represents the frequency component of each individual spike. The resulting dominant harmonic set was used to

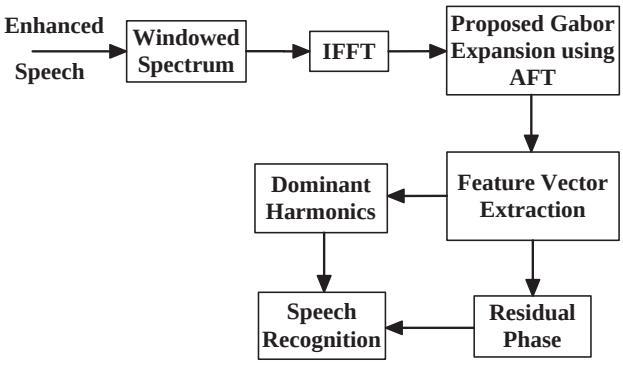


Fig.3 Gabor Feature Extraction

set a Threshold and found to be consistent in differentiating homonyms across a large number of sentences. The resulting Gabor signal spectrum can be decomposed into the short-time magnitude spectrum and the short-time phase spectrum.

$$X(t,w) = |X(t,w)|e^{j\psi(t,w)} \tag{14}$$

$\psi(t,w)$ = $\angle(X(t,w)$ represents the Short time phase spectrum which is unwrapped. A Linear Predictive approach similar to the Kalman Filter is tuned using the already estimated K_k and P_k values to obtain the principal phase spectrum. Let r(n) be the predicted sample, h(n) is the mean square value of the difference between the actual value and its predicted estimate.

$$sin\theta(n) = r(n)/h'(n) \tag{15}$$

The phase component obtained $\theta(n)$ is the estimated differential phase consisting of the required principle components. This feature is essential in pitch determination and formant extraction of the speech signal. The dominant harmonic set and phase estimations complete the feature vector set.

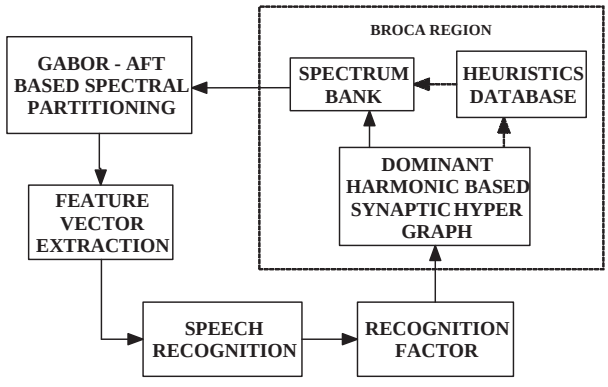
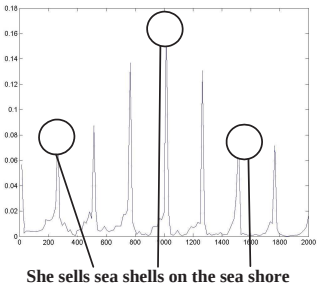
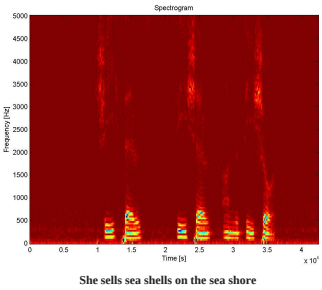


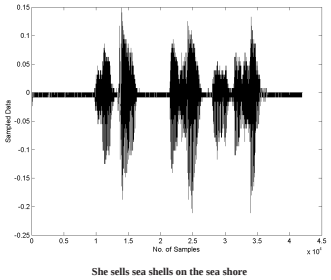
Fig.4 Database and Heuristics for Speech Recognition



(a)Observed Spike train



(b)Gabor Tuned Spectrogram



(c)Input Speech

Fig.5. Recognition results for the given sentence.

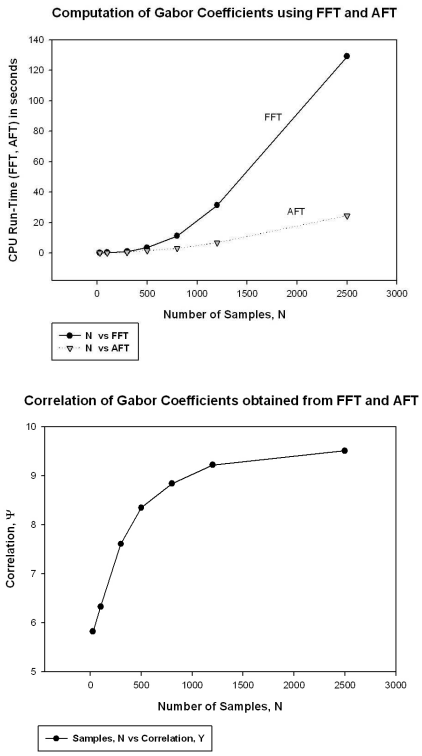


Fig.6 Comparitive Results using FFT and AFT

4. EXPERIMENTAL RESULTS

In the experiments conducted, a sampling frequency of 8 kHz, frame size of the speech signal 100 to 500 samples was selected. The data used for the experiments contained sentences with similar sounding words, speech signal with varying noise conditions. The sentence shown in Fig.5(a) is "She sells sea shells on the sea shore". The figure shows the spike train computed for the given input signal. The harmonics along with the associated phase was used to identify the individual components of the speech.

In Fig.5(b) the proposed Gabor algorithm is used to compute the Enhanced spectrogram. The components are localized in the time and frequency domain. It is seen that the harmonics and phase values were consistent for the same word sets under varying noise parameters, the Kalman filter estimates were fine tuned as shown in Section 1. and the desired output was obtained. The overall accuracy of the proposed system for speech recognition shows positive results.

5. CONCLUSION

The noisy input speech signal is passed through a Kalman filter and the estimated parameters were fined tuned using the proposed design in order to obtain the desired result. This process greatly reduces the time complexity of the recognition system. The Gabor coefficients computed using the Arithmetic Fourier Transform reduced the computational complexity from $O(N^3)$ to $O(N)$. The database under construction is framed based on the system flow shown in Fig.4. The computation of dominant harmonics and the associated phase values from the Gabor Transform is a novel approach and is found to be effective for speech recognition.

6. REFERENCES

[1]. N. Ma, M. Bouchard and R. A. Goubran, "Speech Enhancement Using a Masking Threshold Constrained Kalman Filter and Its Heuristic Implementations," Proc. IEEE Transactions on Audio, Speech, and Language Processing, Vol. 14, No. 1, pp. 19 - 32, Jan. 2006.

[2]. M. Bastiaans, "Gabors expansion of a signal into gaussian elementary signals," PhilipsJ. Opt. Eng., Vol. 20, no. 4, pp. 594-598, July 1981.

[3]. D. W. Tufts and G. Sadasiv, "The Arithmetic Fourier Transform," IEEE ASSP Mag., pp. 13-17, Jan. 1988.

[4]. I . S. Reed and D. W. Tufts et al, "Fourier analysis and signal processing by use of the Mobius inversion formula," IEEE Transactions on Acoustics, Speech and Signal Processing, Vol. 38, pp. 458-469, Mar. 1990.

[5]. L. Rabiner and B. Juang, "Fundamentals of Speech Recognition," Englewood Cliffs, NJ: Prentice-Hall, 1993.