Journal of Basic and Applied Engineering Research
Print ISSN: 2350-0077; Online ISSN: 2350-0255; Volume 1, Number 10; October, 2014 pp. 95-99

http://www.krishisanskriti.org/jbaer.html

Spoken Document Retrieval System with Monosyllable Indexing

Shashi Kant Tiwari

IT FIMT, Kapashera New Delhi 110037

Abstract: **Developing a system that can understand natural language has been a continuing goal of Artificial Intelligence researchers. The system must have some general world knowledge as well as knowledge of what human knows and how they reason**

Keywords: **Phonological, Morphological, Syntactic, Semantic, Pragmatic.**

1. INTRODUCTION

Speech is the most natural and effective mode of communication. With the rapid development of the information technology and increasing popularity of digital computer, human-machine communication using natural speech has received wide attention from both academic and business communities. Developing a system that can understand natural language has been a continuing goal of Artificial Intelligence researchers.

2. WORKING FORMANT FREQUENCY

A predictor polynomial defined as the Fourier Transform of the corresponding second-order predictor is given by

$$A_k(e^{jw}) = 1 - \alpha_k e^{jw} - \beta_k e^{-j2w} \qquad (2.1)$$

The parameter β_k determines the bandwidth of the resonator defined as negative

logarithm of $(-\beta_k).[A_k(e^{jw})]^2$. The formant frequency is given by

$$\varphi_f = \arccos \left[\frac{-\alpha_k (1 - \beta_k)}{4 \beta_k} \right] \qquad (2.1)$$

The beginning point and the end point of segment k is denoted by ω_{k-1} and ω_k respectively. Using the predictor polynomial the prediction error is described as

$$E(\omega_{k-1}, \omega_k | \alpha_k, \beta_k) = \frac{1}{\pi} \int_{\omega_{k-1}}^{\omega} |S(e^{j\omega})|^2 \, |A_k(e^{j\omega})|^2 d\omega \qquad (2.2)$$

where $|S(e^{j\omega})|^2$ denotes the short-time power density spectrum of the speech signal. Using (3.2) the predictor error can be represented as

$$E(\omega_{k-1}, \omega_k | \alpha_k, \beta_k) = (1 + \alpha_k^2 + \beta_k^2) r_k(0) - 2\alpha_k (1 - \beta_k) r_k(1) - 2\beta_k r_k(2) \qquad (2.3)$$

where $r_k(v)$ are the autocorrelation coefficients for segment k for v =0,1, 2.

$$r_k(v) = r_{\omega_{k-1}, \omega_k}(v)$$

$$= \frac{1}{\pi} \int_{\omega_{k-1}}^{\omega_k} |S(e^{j\omega})|^2 \cos(v\,\omega) d\omega \qquad (2.4)$$

By minimizing the prediction error as given by (3.6) with respect to α_k and β_k, the optimal predictor coefficients are expressed as

$$\alpha_k^{opt} = \frac{r_k(0) r_k(1) - r_k(1) r_k(2)}{r_k(0)^2 - r_k(1)^2}$$

$$\beta_k^{opt} = \frac{r_k(0) r_k(2) - r_k(1)^2}{r_k(0)^2 - r_k(1)^2}$$

The value of the minimum predictor error is given by

$$E_{\min}(\omega_{k-1}, \omega_k) = \min_{\alpha_k, \beta_k} E(\omega_{k-1}, \omega_k | \alpha_k, \beta_k)$$

$$= r_k(0) - \alpha_k^{opt} r_k(1) - \beta_k^{opt} r_k(2) \qquad (2.5)$$

Thus, from the minimization requirement, α_k, β_k as follows:
$\beta_k + \alpha_k < 1$
$\beta_k - \alpha_k < 1$
$| \beta_k | < 1$

This requirement is fulfilled by the condition that

$$\alpha_k^2 + 4\beta_k < 0$$

which can be tightened further by combining the previous constraints to the new constraints

It is thus evident that $| cos\ \omega | < 1$, thus, the value of α_k and β_k can be defined as

$$\alpha_k < 2$$

$$-1 < \beta_k < \frac{-\alpha_k^{\,2}}{4}$$

$$\cos\vartheta = \frac{\alpha_k}{2\sqrt{(-\beta_k)}} \qquad (2.6)$$

The resonance frequency for the k^{th} segment is given by the equation

$$\cos\varphi_k = -\frac{\alpha_k(1-\beta_k)}{4\beta_k} \qquad (2.7)$$

From the inequality $|\cos\varphi_k| < 1$, we can obtain the following constraints for α_k and β_k

$$| \alpha_k | < 2$$

$$-1 < \beta_k < -\frac{| \alpha_k |}{4 - | \alpha_k |}$$

Plotting the corresponding boundary line in the (α,β) plane it is evident that these constraints are tighter than the constraints for a pole solution.

3. SPECTROGRAM

The Fourier Transform of the windowed speech waveform, i.e., STFT is given by

$$X(\omega,\tau) = \sum_{n=-\infty}^{n=\infty} x[n,\tau]\exp[-j\omega n] \qquad (3.1)$$

where $x\ [\ n,\ \tau] = w[n,\ \tau]x[n]$ represents the windowed speech segments as a function of the window centre at time τ. The spectrogram is a graphical display of the magnitude of time-varying spectral characteristics and is given by

$$S(\omega,\tau) = | x(\omega,\tau) |^2 \qquad (3.2)$$

$$p[n] = \sum_{k=-\infty}^{\infty} \delta\ [n - kP] \qquad (3.3)$$

In the windowed speech waveform the result can be expressed as

$$x[n,\tau] = w[n,\tau]\{(p(n) * g(n)) * h(n)\} \qquad (3.4)$$

where the glottal waveform over a cycle and vocal tract impulse response are lumped into $\bar{h}[n] = g[n] * h[n]$.

Using Multiplication and Convolution theorem, the Fourier Transform of the speech segment is given by

$$X(\omega,\tau) = \frac{1}{P} W(\omega,\tau) \otimes \left[H(\omega)G(\omega) \sum_{k=-\infty}^{k=\infty} \delta(\omega - \omega_k) \right]$$

where $\bar{H}(\omega_k) = H(\omega_k)G(\omega_k)$ **and** $\omega_k = \frac{2\pi k}{P}$ **and** $\frac{2\pi}{P}$ **is the fundamental frequency.**

Therefore, the spectrogram of x [n] can be expressed as

$$S(\omega,\tau) = \frac{1}{P^2} \left| \sum_{k=-\infty}^{\infty} \bar{H}(\omega_k) W(\omega - \omega_k, \tau) \right|^2 \qquad \textbf{(3.5)}$$

4. TEMPORAL ENERGY

Temporal energy of the spectra due to a phoneme can be taken as an important representation of the characteristics of the phoneme. The energy of the sequence $x\ (n)$ can be written as

$$\varepsilon_n = \sum_{-\infty}^{\infty} | x(n) |^2 \qquad (4.1)$$

$$= \int_0^{\pi} \frac{| X(e^{jw}) |^2}{\pi} dw \qquad (4.2)$$

for real sequence using even symmetry. From equation (3.25) the energy density spectrum of $x(n)$ can be expressed as:

$$\Phi_x(\omega) \underset{=}{\Delta} \frac{\left|X(e^{j\omega})\right|^2}{\pi} \quad (4.3)$$

Then the energy of $x(n)$ in the $[\omega_1, \omega_2]$ band is given by

$$\int_{\omega_1}^{\omega_2} \Phi_x(\omega) d\omega, \quad 0 < \omega_1 < \omega_2 < \pi \quad (4.4)$$

LPC Cepstral Coefficient

The method of computing the LPC coefficients are based on the assumption that a speech sample at time n, $s(n)$ can be approximated by a linear combination of the past p speech samples as given by equation .

$$S(n) \approx a_1 s(n-1) + a_2 s(n-2) + \ldots\ldots\ldots + a_p s(n-p)$$

Where a_1, a_2, ….., a_p are constant coefficients.

The equation can be further transformed by including an excitation term $Gu(n)$ to:

$$S(n) = \sum_{i=1}^{p} a_i s(n-i) + Gu(n) \quad (4.5)$$

where G is the gain and $u(n)$ is the normalized excitation. The transformation of equation (4.3) to z-domain is given by equation (4,4)

$$S(z) = \sum_{i=1}^{p} a_i z^{-i} S(z) + GU(z) \quad (4.6)$$

and the corresponding transfer function H(z) is described as

$$H(z) = \frac{S(z)}{GU(z)}$$

$$= \frac{1}{1 - \sum_{i=1}^{p} a_i z^{-1}} = \frac{1}{A(z)} \quad (4.7)$$

with error transfer function

$$A(z) = \frac{E(z)}{S(z)} = 1 - \sum_{i=1}^{p} a_k z^{-k} \quad (4.8)$$

In is now required to obtain the set of coefficients a_k, that minimizes the prediction error in a short segment of speech. The mean short time predictor error per frame is given as:

$$E_n = \sum_m e_n^2(m) = \left[s_n(m) - \sum_{k=1}^{p} a_k s_n(m-k) \right]^2 \quad (4.9)$$

where $s_n(m)$ is the segment of speech selected in the neighborhood of a sample $s_n(m)=s(m+n)$. The value of the coefficient a_k that minimize the error E_n can be obtained considering $\frac{dE_n}{da_i} = 0, \quad i = 1,2,3....., p$, as given in equation (4.10).

$$\sum_m S_n(m-i) S_n(m) = \sum_{k=1}^{p} a'_k \sum_m s_n(m-i) s_n(m-k), 1 \le i \le p \quad (4.11)$$

where a'_k are the values of a_k that minimizes E_n. Defining $\Phi_n(i,k) = \sum_m s_n(m-i) s_n(m-k)$, the equation (4.11) can be rewritten as:

$$\sum_{k=1}^{p} a_k \Phi_n(i,k) = \Phi_n(i,0), \text{ i=1,2,….p} \quad \text{---(4.12)}$$

This is a system of p equations with p variables. The equations can be solved to find a_k coefficients for the segment $s_n(m)$. Thus, E_n can be represented as

$$E_n = \sum_m s_n^2(m) - \sum_{k=1}^{p} a_k \sum_m s_n(m) s_n(m-k) \quad (4.13)$$

and in compact form, E_n further reduced to equation (4.14)

$$E_n = \Phi_n(0,0) - \sum_{k=1}^{p} a_k \Phi_n(0,k) \text{ --- (4.14)}$$

Considering $s_n(m)$ as null outside the interval $1 \le m \le N$, the error function $\Phi_n(i,k)$, can be expanded as:

$$\Phi_n(i,k) = \sum_{m=1}^{N+p} s_n(m-i) s_n(m-k)$$

, $1 \le i \le p$, $1 \le k \le p$ (4.15)

which can be further rewrite as:

$$\Phi_n(i,k) = \sum_{m=1}^{N-(i-k)} s_n(m)s_n(m+i-k)$$

, 1≤i≤p, 1≤k≤p (4.16)

In this case, $\Phi_n(i,k)$ is related to the short-time autocorrelation function value $R_n(i,k)$ as:

$$\Phi_n(i,k) = R_n(i,k) \quad (4.17)$$

where $R_n(k) = \sum_{m=1}^{N-k} s_n(m)s_n(m+k)$, is a pair function. Thus,

$$\Phi_n(i,k) = R_n(\,|\, i+1-k \,|\,)$$

, 1≤i≤p, 1≤k≤p (4.18)

and therefore:

$$\sum_{k=1}^{p} a_k R_n(|i+1-k|) = R_n(i), \; 1 \le i \le p \quad (4.19)$$

By analogy, the square of the prediction error can be expanded as:

$$E_n = R_n(0) = \sum_{k=1}^{p} a_k R_n(k) \quad (4.20)$$

The Levinson-Durbin recursion algorithm is used to solve equation (4.20). The complete algorithm is described as follows:

Taking $E^{(1)} = R(1)$, the following sets of equation is solved recursively for *i*=2,3,…..p

$$k_i = \frac{R(i) - \sum_{j=1}^{i-1} a_j^{i-1} R(i-j)}{E^{(i-1)}} \quad (4.21)$$

$$a_j^{(i)} = k \quad (4.23)$$

$$a_j^{(i)} = a_j^{(i-1)} - k_i \; a_{i-j}^{(i-1)}, \; 1 \le j \le i-1 \quad (4.24)$$

$$E^{(i)} = (1-k_i^2)E^{(i-1)} \quad (4.22)$$

Thus, we have

$$a_j = a_j^{(p)}, \; 1 \le j \le p \quad (4.23)$$

Instead of using directly the LPC coefficients as feature vectors, cepstral coefficients, based on LPC analysis, are usually used because of their superior recognition capabilities. The LPC-based cepstral coefficients have been derived as follows.

$$c_m = a_m + \sum_{k=1}^{m-1} \frac{k}{m} c_k a_{m-k}$$

, 1 ≤ k ≤ p (4.24)

$$c_m = \sum_{k=1}^{m-1} \frac{k}{m} c_k a_{m-k}, \; m > p \quad (4.25)$$

where c_k is an LPC-based cepstral coefficient. It has been established that LPC-based cepstral coefficients produce better recognition results when they are appropriately weighted.The weighting function is given as

$$w(m) = 1 + \frac{Q}{2}\sin\left(\frac{\pi\, m}{Q}\right), \; 1 \le m \le Q \quad (4.26)$$

where Q is the order of cepstral coefficients. In addition to the cepstral coefficients, their time-derivative approximations are used as feature vectors to account for the dynamic characteristic of speech signal. The time derivative is approximated by a linear regression coefficient over a finite window.

$$\Delta\, \hat{c}_l(m) = \left[\sum_{k=-K}^{K} k \;\; \hat{c}_{l-k}(m)\right].G, \; 1 \le m \le Q \quad (4.27)$$

5. OUTPUT EXPERIMENTAL SETUP AND RESULTS

5.4.1 Pre-emphasis

The speech signal (here, also refereed as {\it word}), s(n), is filtered with a first-order FIR filter to spectrally flatten the signal. We used one of the most widely used pre-emphasis filter of the form

$$H(z) = 1 - az^{-1},$$

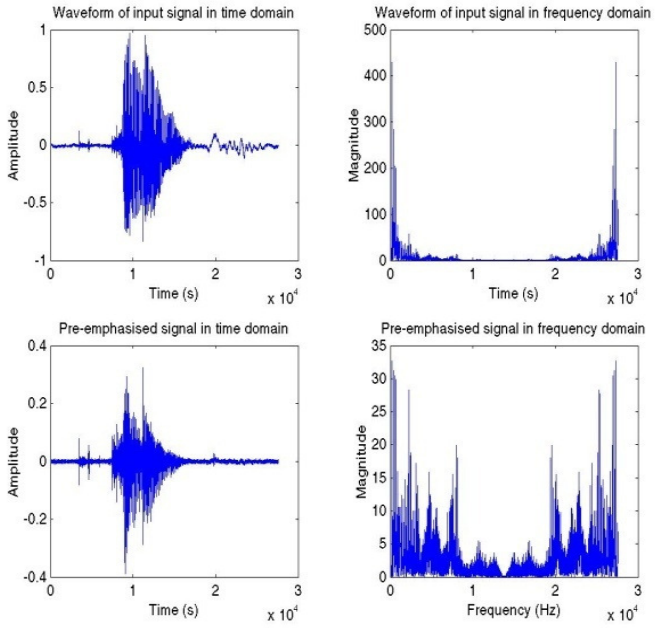

Fig. 5.1

REFERENCES

[1] Ann, S.: Current status and prospects of speech processing technology in Korea, *proceedings of 1994 International Symposium on Speech, Image Processing and Neural Networks*, Hong Kong, Vol.1, 133-136, 1994.

[2] Barra, R.; Montero, J.M.; Macías, J.; D'haro, L.F.; San Segundo, R. and De Córdoba, R.: Prosodic and segmental rubrics in emotion identification, Proc. International Conference on Acoustics, Speech and Signal Processing 206, Toulouse, pp 1085-1088, 2006.

[3] Bartkova, K. and Jouvet, D.: Selective prosodic post-processing for improving recognition of French telephone numbers, in Eurospeech'99, 6th European Conference on Speech Communication and Technology. Budapest, Hungary, 5-10 September 1999. Vol 1 pp. 267-270, 1999.

[4] Bartkova, K.: Some experiments about the use of prosodic parameters in a speech recognition system, Proceedings of the ESCA Workshop on Intonation. Athens, 18-20 September 1997. pp. 33-36, 1997

[5] Bassi, A.; Becerra Yoma, N. and Loncomilla, P.: Estimating tonal prosodic discontinuities in Spanish using HMM, Speech Communication 48, 9: 1112-1125, 2006.

[6] Baum, L.E.; Petrie, T.; Soules, G. and Weise, N.: A Maximization Technique Occurring in The Statistical Analysis of Probabilistic Function for Markov Chains, *Ann. Math. Stat.*, Vol.41, No.1, PP. 164-171, 1970.