Feature extraction for speech recognition

Mr. Ramprasad Jat

Computer Science and Engineering
Engineering college
Jhalawar

***Abstract*- automatic speech recognition (ASR) has made strides with development of digital signal processing hardware and software. But machine cannot match the performance of their human counterparts in term of accuracy and speed, speaker independent recognition. We discuss the signal modeling approach for speech recognition. The overview of basic operation involved in signal modeling. Further common used temporal and spectral analysis techniques of feather extraction.**

I. INTRODUCTION

Speech recognition system performs two basic operations: - 1. Signal modeling 2. Pattern matching
Signal modeling represents of convert speech signal into set of parameters / operation.
1. Speech shaping: - speech shaping is the process of converting the speech signal from sound pressure wave to a digital signal. Emphasizing important frequency components in signal.
2. Feature extraction: feature extraction is process obtaining such as power, pitch and vocal tract configuration from the speech signal.
3. Parameter transform:- process of converting feather into signal parameter through process of differentiation and concatenation.
4. Statistical modeling: - conversion of parameters in signal observation vectors.

II. SPECTRAL SHAPING

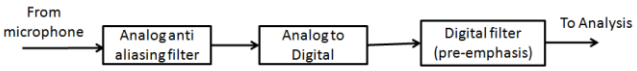

Figure 1: Basic operation of spectral shaping

Spectral shaping are two basic operation:- (a) Digitization:- conversion of analog speech signal to digital signal.
(b) Digital filtering: - emphasizing frequency components in signal.
Main purpose of digitations process is to produce a sampled data represent of speech signal with high signal to noise ratio (SNR).

$$H_{pre}(Z) = \sum_{k=0}^{Npre} a_{pre}(k) Z^{-k}$$

Where $H_{pre}(Z)$ = finite impulse response filter
One coefficient digital filter is called pre-emphasis filter.

$$H_{pre}(Z) = 1 + a_{pre} Z^{-1}$$

Range of value a_{pre} for is [-1.0,- 0.4]. The pre-emphasis filter boosts the signal spectrum approximately 20 dB per decade.
Advantage of pre-emphasis filter
1. The voice section of speech signal have negative spectral slope (attenuation of approximate 20 dB per decade). The pre-emphasis filter serves to offset this slope before spectral analysis improves the efficiency of analyses.
2. The hearing is more sensitive above 1 khz region of spectrum. The pre-emphasis filter amplifies of the spectrum.

III. FEATURE EXTRACTION

Feature extractions are two techniques for speech feather extraction:- (a) Temporal analysis – speech waveform used. (b) Speech analysis- spectral representation of speech signal used.

3.1 Spectral Analysis technicians

3.1.1 Critical band filter bank analysis

Critical band filter bank analysis is simply bank of linear phase FIR band pas filters that are arranged linearly along mel bark.
Critical band rate scale defined

$$bark = 13 \operatorname{atan}\left(0.76 * \frac{f}{1000}\right) + 3.5 \operatorname{atan}\left(\frac{f^2}{(7500)^2}\right)$$

$$melfrequency = 2595 \log_{10}\left(1 + \frac{f}{700}\right)$$

An expression for critical bandwidth is

$$BW_{critical} = 25 + 75\left[1 + 1.4\left(\frac{f}{1000}\right)^2\right]^{0.69}$$

3.1.2 **Cepstral analysis**: -

In this technique provides methodology for separating the excitation from the vocal track share.

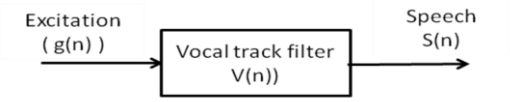

Figure 2: Linear acoustic model of speech production

Speech signal is defined

$$s(n) = g(n) * v(n)$$

Where v(n) = vocal tract impulse response

g(n): excitation signal

The frequency domain representation

$$s(f) = g(f) * v(f)$$

Taking log on both sides

$$\log(s(n)) = \log(g(n)) + \log(v(n))$$

Cepstrum is computed by taking inverse discrete Fourier transform (IDFT) of logarithm of magnitude of discrete Fourier transform.

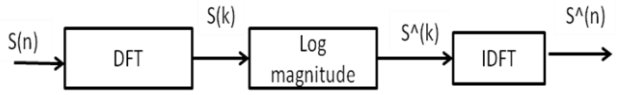

Figure 3: System for obtaining cepstrum "adapted from figure 2"

$$s(k) = \sum_{n=0}^{N-1} s(n)\exp\left(-\frac{j2\pi}{N}\right)nk$$

$$S^{\wedge}(k) = \log(S(k))$$

$$s^{\wedge}(k) = \frac{1}{N}\sum_{k=0}^{N-1} s^{\wedge}(k)\exp\left(\frac{j2\pi}{N}\right)nk$$

Where $s^{\wedge}(n)$ is cepstrum.

Cepstrum alysis is uded for tracking and pitch Vo detection.

The sample s^(n) in first 3 ms describe V(n) and can be separated from the excitation. If fs^(n) exhibits sharp periodic pulse then voice occur. The interweel between these pulses is pitch period.

If above no structure is visible in s^(n). speech is considered unvoiced.

3.1.3 Melcepstrum analysis:- Melcepstrum analysis use cepstrum with a non linear frequency axis following mel scale.

Mel cepstrum speech waveform s(n) is first windowed with analysis window (w(n)) and then DFT s(k) is computed.

The magnitude of s(k) is then weighted by a series of mel filter frequency response whose center frequencies and bandwidth match of auditory critical band filter.

$$E_{mel}(n,l) = \left(\frac{1}{A_l}\right)\sum_{k=L1}^{U1} |V_l(k) * s(k)|^2$$

E_{mel} is mel cepstrum energy.

$V_l(k)$=the frequency response at l^{th} mel scale filter.

U_l= upper frequency

L_l=lower frequency

A_l= energy of filter which normalize the filter accoding to varing bandwidth.

$$C_{mel}(n,m) = \left(\frac{1}{n}\right)\sum_{l=0}^{N-1} \log\{E_{mel}(n,l)\}\cos\left(\frac{2\pi\left(l+\frac{1}{2}\right)}{N}\right)$$

C_{mel} representation for speech spectra exploiting auditory as well as dicorrelating property of cepstrum.

3.2 Temporal Analysis: Temporal Analysis is process of waveform of speech signal like power and periodicity power estimation.

Power is computed on frame by frame

$$p(n) = \left(\frac{1}{Ns}\right)\sum_{m=0}^{Ns-1}\left(w(m) * s\left(n - \frac{Ns}{2} + m\right)\right)$$

N_s = number of sample used to computer she power

s_n = signal

w(m)= window function
n= sample index of center of the window

Hamming window is defined as

$$w(n) = \alpha_w - \left(\frac{(1-\alpha_w)\cos\left(\frac{2n\pi}{Ns-1}\right)}{\beta_w} \right)$$

W(n)=0

α_w = window constant in the rang (0, 1)

N_s = window duration in sampled

α_w = 0.54 β_w is normalization constant

IV. CONCLUSION

Different temporal and spectral analysis techniques for feather extraction have been stuied in details.

Temporal analysis techniques involve less computation but there are limited to determination simple speech parameters like power, energy and periodicity of speech. Finding vocal tract parameters we require spectral analysis techniques.

Critical band filter bank decompose the speech signal into discrete set of spectral samples containing information presented to higher levels processing in auditory system.

Cepstral analysis separates the speech signal into component representing excitation source and a component representing vocal tract impulse response. It provides the information about pitch and vocal tract configuration. But it is computation more intensive,

Mel cepstral analysis has decorrelating property of cepstral analysis and includes some aspect of audition.

REFERENCES:

1. J. W. picone, " Signal modelling technique in speech recognition" proc of IEEE, vol; 81, no 9, pp.1215-1247, sep 1993.
2. B. Goldand L. R. Rabiner, " parallel processing techniques for estimatingf pitch period of speech in the time domain," J. Acoust. Soc, America, vol 46, pt 2, no 2, pp 442-448 Aug 1969
3. H.Hermansky, B. A. Hanson, and H. Watkita, " Perceptually based processing in automatic speech recognition," Proc. IEEE Int. Conf. on Aoustic , speech and signal processing , pp. 1971-1974, apr. 1986.