AN APPROACH TO DESCRIBE METHODS OF FRONT END PROCESSING OF SPEECH SIGNAL

Varsha Gupta, Mukul Pant

Abstract

Speech is the most natural way of information exchange. It provides an efficient means of man machine communication using speech interfacing. Today significant part of speech recognition research is focused on speaker independent speech recognition problem. Before recognition, speech processing has to be carried out to get feature vectors of the signal. So, front end analysis plays an important role. The reasons are its wide range of applications, and limitations of available techniques of speech recognition. So, in this paper we briefly discuss the different techniques of front end analysis of speech recognition including feature extraction techniques.

KeyIndex-Pre-emphasis,EPD,Framing&Windowing,LPC,MFCC,Noise,cepstrum

1.INTRODUCTION

Automatic speech recognition is currently used in many real time applications such as cellular telephones, computers and security systems. Speech recognizers consist of feature extraction stage and a classification stage which are known as front end processing and back end processing units [1]. These two units are shown in figure (1). Although these two units may appear to be independent, they are highly coupled. To be effective, the features should be capable of separating the speakers from each other in its space, whereas the classifier should be tuned to differentiate the different classes in a given feature space. This paper explains the total front end processing unit which includes the pre-processing and feature extraction techniques. Front-end processing (FEP) extracts feature vectors from raw utterances. FEP is Showing two blocks (pre-processing and feature extraction) for processing of spoken words signal [2]. In pre-processing, raw spoken words utterances converted into vector form for further signal processing. Components of pre-processing are as follows:

1 noise cancelling

2 pre-emphasis of signal

3 voice activity detection

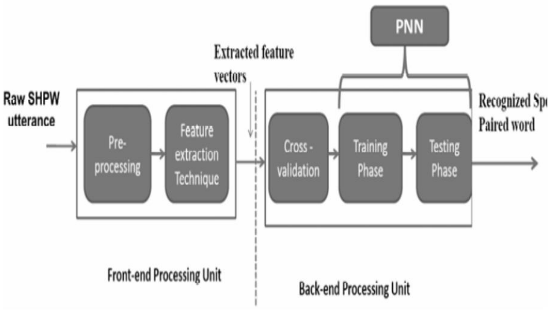


Figure 1 Block diagram of proposed word recognition system

Above components collectively converts raw spoken words utterance into desirable vector form for further processing. Here, the goal of voice activity detection method is to extract the desired spoken word utterances from the raw spoken word utterances. Extracted feature vectors can be obtained after applying feature extraction technique on utterance. It is widely accepted that feature extraction of utterance, influences the correct recognition rate.

2. PRE-PROCESSING

First step is to create feature vector. The goal in the pre-processing is to adjust the speech signal, $x(n)$, so that it will be “more appropriate” for feature extraction analysis. The pre-processing operations noise cancelling, pre-emphasis and voice activation can be seen in fig 3. The first thing to consider is if speech, $x(n)$, is corrupted by some noise,$d(n)$, for example an additive disturbance $x(n)$= $s(n)$+ $d(n)$ where $s(n)$is the clear speech signal. There are several approaches to perform noise reduction on a noisy speech signal. Two frequently used noise reduction algorithms is spectral subtraction and adaptive noise cancellation [3]. A low signal to noise ratio decrease the performance of the recognizer.

(1)Pre-emphasis: It is used to spectrally flatten the speech signal**.** This is generally done by a high pass filter. The most commonly used filter for this step is FIR filter described below:

$$H(z)= 1-0.95z^{-1}$$

The filter in the time domain will be $h(n)$ = {1,-0.95} and the filtering in the time domain will give the pre-emphasized signal $s_1(n)$ = $\sum_{k=0}^{m-1} \mathrm{h(k)\hat{s}(n-k)}$

(2)Voice Activation Detection (End Point Detection): The problem of locating the end points of an utterance in a speech signal is a major problem for the speech recognizer. Inaccurate end point detection will decrease the performance of speech recognizer. When a fair SNR is given then this task is made easier otherwise the task is difficult. Some commonly used measurements for finding speech are short time energy estimate E_{s1} or short time power estimateP_{s1} and short term zero crossing rate Z_{s1}.

$$E_{s1}(m) = \sum_{n=m-L+1}^{m} s^2\,(\mathrm{n})$$

$$P_{s1}(m) = \frac{1}{L}\sum_{n=m-L+1}^{m} s^2\,(\mathrm{n})$$

$$Z_{s1}(m) = \frac{1}{L}\sum_{n=m-L+1}^{m} \frac{|\mathrm{sgn}(\mathrm{s}_1(\mathrm{n})) - \mathrm{sgn}(\mathrm{s}_2(n-1))|}{2}$$

.

For each block of L samples these measures calculate some values. The short term energy estimate will increase when speech id present in$s_1(n)$. This is also the case of short term power estimate; the only thing that separates them is scaling with $\frac{1}{L}$ when calculating the short term power estimate. The short term zero crossing rate gives a measure of how many times the signal, $s_1(n)$, changes sign.

These measures will need some triggers for making decision about where the utterances begin and end. To create a trigger, one needs some information about the background noise. This is done by assuming that the starting blocks are noise. Then mean and variance will be calculated. To make a more comfortable approach the following function is used:

$$W_{s1}(m) = P_{s1}(m) \, . \left(1 - Z_{s1}(m)\right) . S_c$$

Using this function both the short-term power and the zero crossing rates will be taken into account. S_c is scale factor for avoiding small values. The trigger for this function is:

$$t_W = \mu_W + \alpha\delta_W$$

Here μ_W and δ_W are the mean and variance for $W_{s1}(m)$. After some testing, some approximation of α will give a good VAD.

3. Frame Blocking and Windowing

The next thing is to divide signal into speech frames and apply a window to each frame. We assume that each frame is K samples long, with adjacent frames being separated by P samples.

Typically values of K and P are 320 samples and 100 samples. By choosing frames of 20ms one can assume that the speech is stationary within each frame. By applying frame blocking to $x_1(n)$ one will get M vectors of length K, which corresponds to $x_1(k;m)$. The next thing is to apply a window to each frame in order to reduce signal discontinuity at either end of the block.

4. Feature Extraction

4.1 speech to feature vector

Feature extraction means creating feature vector from the speech signal. The main steps for extracting information are pre-processing, frame blocking and windowing, feature extraction and post processing [3]. See figure 2

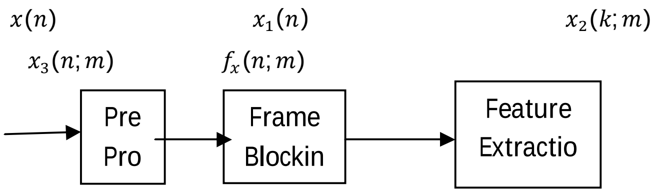


Figure 2 –Main Steps in Feature Extraction

From a correct band limited and sampled speech signal, $x(n)$,one will finaly yield the feature vectors $f_x(n;m)$ where m=0,1,2,3........,M-1 and n=0,1,2,3........,N-1 means m vector of size n. The next step is an important one, namely to extract information from the speech blocks. Some commonly used methods for speech recognition is linear prediction and mel-capstrum. These measures have been widely used and here are some reasons why:

- These measures provide a good model of the speech signal. This is particularly true in quasi steady state voiced region of speech.
- The way these measures are calculated leads to a reasonable source-vocal tract separation. This property leads to a

fairly good representation of the vocal tract characteristics(which is directly related to the speech sound being produced)

- The measures have an analytically tractable model.
- Experience has found that these measures work well in recognition application.

Other measures to add to the feature vectors are energy measures and also the calculation of delta and acceleration coefficients. The delta coefficients means that a derivative approximation of some measure (e.g.) Linear prediction coefficients) is added and the acceleration coefficients is the second derivative approximation of some measures.

(a)Linear Prediction: LPC is used to estimate basic speech parameters like pitch formants and spectra. The principle behind the use of LPC is to minimize the sum of the squared differences between the original speech and estimated speech signal over a finite duration. This could be used to give a unique set of predictor coefficients. These predictor coefficients are normally estimated in every frame, size of the frame is dependent on the word. The predictor coefficients are represented by α_k. Another important parameter is the gain (G). The transfer function of the time varying digital filter is given by (1)

$$H(z) = \frac{G}{1 - \sum_{k=1}^{p} a_k \, z^{-k}}$$

The summation is computed starting at k=1 up to p, which is taken as 12 . This means that only the first 12 coefficient are transmitted to LPC synthesizer. The two most commonly used methods to compute the coefficients are the covariance method and the auto correlation formulation. For our implementation we will be using the built in function LPC. LPC uses the Levinson-Durbin recursion to solve the numerical equations that arise from the least-squares formulation. This computation of the linear prediction coefficients is often referred to as the autocorrelation method. The reason is that this method is superior to the covariance method in the sense that the roots of the polynomial in the denominator of the above equation is always guaranteed to be inside the unit circle. Hence guaranteeing the stability of the system H(z).

(b)Mel Frequency Cepstral Coefficient (MFCC): The extraction and selection of the best parametric representation of acoustic signals is an important task in the design of any speech recognition system; it significantly

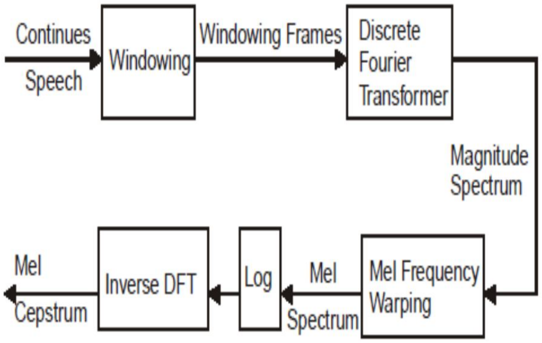
affects the recognition performance. A compact representation would be provided by a set of mel-frequency cepstrum coefficients (MFCC), which are the results of a cosine transform of the real logarithm of the short-term energy spectrum expressed on a mel-frequency scale. The MFCCs are proved more efficient. The calculation of the MFCC includes the following steps.

(i) Mel-Frequency Wrapping: Human perception of frequency contents of sounds for speech signal does not follow a linear scale. Thus for each tone with an actual frequency, f, measured in Hz, a subjective pitch is measured on a scale called the 'mel' scale [4]. The melfrequency scale is a linear frequency spacing below 1000 Hz and a logarithmic spacing above 1000Hz .As a reference point ,the pitch of a 1 KHz tone ,40dB above the perceptual hearing threshold, is defined as 1000 mels. Therefore we can use the following approximate formula to compute the mels for a given frequency *f* in Hz.

$$\mathrm{Mel}(f) = 2595*\log10(1 + f/700)$$

Ours approach to simulate the subjective spectrum is to use a filter bank, one filter for each desired mel-frequency component. That filter bank has a triangular band pass frequency response and the spacing as well as the bandwidth is determined by a constant mel-frequency interval. This corresponds to series of band pass filters with constant bandwidth and spacing on a mel frequency scale.

(ii)Cepstrum: In this final step, we convert the log mel spectrum back to time. The result is called the Mel Frequency Cepstrum Coefficients (MFCC). The cepstral representation of the speech spectrum provides a good representation of the local spectral properties of the signal for the given frame analysis. Because the mel spectrum coefficients (and so their logarithm) are real numbers, we can convert them to the time domain using the discrete cosine transform (DCT).



Complete Pipeline for MFCC

3. CONCLUSION

Overall front end process of speech recognition is described in this paper. The process with different methodology and technique wrapping in one paper. This could be helpful to see overall view on all the methods and combine the speech processing methods with each other.

4. ACKNOWLEDGMENT

The authors would like to thank the management of Dr. D.K Rajoriya for encouragement and support in this review work.

5. REFRENCE

1. Chaitra R Nanjangud, Prof Sharda C Sajjan, "LPC and Wavlet based feature extraction for speech recognition"

2. Dinesh Kumar Rajoriya, R.S. Anand and R.P. Maheshwari, “Enhanced recognition rate of spoken Hindi paired word using probabilistic neural network approach**”,** Int. J. Information and Communication Technology, Vol. 3, No. 2, 2011

3 .Mikael Nilsson, Marcus Ejnarsson, “ Speech Recognition using hidden markov model”

4. Vibha Tiwari , “MFCC and its applications in speaker recognition”, International Journal on Emerging Technologies **1**(1): 19-22(2010)

5. D.K.Freeman, G.Coiser,C.B Southcott and I.Boyd, “The Voice activity detector for the pan-European digital cellular mobile telephone service” in proc. Int. Conf. Acoust., Speech , Signal processing, Glasgrow, U.K., May 1989, pp. 369-372

6. M.J. hunt, M.Lennig and P.mermelstein,” Experiments in syllabus-based recognition of continuous speech”, proc.IEEE Intl Conf. Acoustics, Speech & Signal processing , Denver, 1980, pp.880-3

7. L.R. Ranbiner and B.H. Juang, B.Yegnanarayana, Fundamentals of speech Recognisation, 1st edition, Pearson education in south Asia, 2009

8. V.Kumar, “A statistical approach towards the recognisation of Hindi language words”, inria 00114544, ver.1, 2006, pp. 1-5.