ABSTRACT
The main task of this work is to improve the performance of the existing recognition system in the acoustic and phonetic phases by extracting new features joined with the baseline system feature vectors to increase the separation distance measure especially between confused syllables in Speaker Independent Large Vocabulary Mandarin Speech Recognition System (SI-LVMSRS). We demonstrate the effect of using the Non-Linear Energy Operator (NLEO) distribution based on AM-FM demodulation techniques on the error rate reduction, assuming that the individual component signals are spectrally isolated by each other and can be modeled as discrete-time mono-component AM-FM signals. Using NLEO as feature instead of the traditional energy operator (TEO) method of computing the energy, by examining many parameters in spectrum distribution in combination with the MFCC as front-end detection parameters combined with the acoustic modeling type Duration Distribution Based Hidden Markov Model (DDBHMM). The experiment shows the advantage of eliminating the pre-emphasis, while using NLEO in the feature vectors instead of TEO. The Relative Average Error Rate Reduction (RAERR) is improved when the number of candidates are increased (5.64%, 9.89%, 19.26%) when 1, 5, 25 candidates are used respectively. If we are careful to adjust the way of computing the parameters of the energy operator, these are affected by the distribution of these components in the time-frequency space.

1. INTRODUCTION
The 'energy' in a signal usually indicates the average of the sum of the squares of the magnitude of that signal. An alternate representation also commonly used is by computing the Discrete Fourier Transform (DFT) of the signal, where the square magnitude of the frequency samples of the computed transformation assumed to represent the energy in the respective frequency components. Notice for example that a 100 Hz signal has the same energy as a 2000 Hz signal. However, the energy required generating the acoustic signal for 2000 Hz is much greater than the 100 Hz signal [1,2,3,4]. It is believe that the NLEO based features are better able to reflect the nonlinear airflow structure of speech production under adverse stressful conditions [5]. In order to reflect this kind of signal difference, we will use a nonlinear method to compute the energy operator by modeling the speech signal as a kind of AM-FM signal. Due to physiological characteristics of the speech production system, voiced sections of the speech signal have an attenuation of approximately 20dB per decade. To counterbalance this negative slope, a pre-emphasis filter (PEF) is generally used before a spectral analysis. Unfortunately, PEF also raise noise spectral energy above 4 kHz and in the case of vowels and nasal energy above 4kHz these are largely considered as a noise. Many speech recognition systems have used adaptive techniques to minimize this defect in using PEF [9]. In our recent work we eliminate the PEF from the pre-processing stage.

The problem previously described can be considered as a spectrum distribution problem. The acoustic level information will be improved by using the non-linear energy operator (NLEO) in the feature vectors instead of the traditional method. The spectral energies will be considered according to frequency positions by applying an adaptive procedure of the non-linear discrete energy operator. In section 2, we present the analysis of signal processing with particular emphasis on the characteristics related to NLEO technique. In section 3, we describe the evaluation method and database. In section 4 the comparative results are presented, and in section 5 we conclude and discuss the results.

2. ALGORITHM FORMULATION

2.1 Non-Linear Energy Operator
The traditional discrete time Energy Operator (TEO) can be computed in two ways, First: In discrete time domain and given as Acoustic modeling, phonetics and speech analysis
\[ TEO(m) = \frac{1}{N} \sum_{n=0}^{N-1} x(m,i)^2 \]  

(1)

Where \( N, m \) are the frame size and frame index respectively.

**Second:** in spectral analysis as the mean value of the summation of the values of each of the frequency components of the Band Pass Filter in the Ideal DFT, and can be defined as:

\[ TEO(m) = \frac{1}{N} \sum_{n=0}^{N-1} \sqrt{[(X_r(m,n))^2 + (X_i(m,n))^2]} \]  

(2)

Where \( X_r \) and \( X_i \) are the real and imaginary components of the FFT outputs respectively. Teager’s energy operator [5] calculation is based on the fact that the speech signal can be modeled as the sum of N AM-FM signals. This model represents each component of the speech signal as a signal with a combined amplitude modulation (AM) and frequency modulation (FM) structure. AM-FM signals of the form

\[ x(t) = a(t) \left[ \int_0^t w(\tau)d\tau \right] \]  

(3)

Is very useful in analog communication systems and have been recently used in [6,7,8] to model speech resonance’s. One of the simple algorithms for signal analysis extensively used, was the following non-linear energy-tracking operator given into two forms, First: Continuous form when operating on continuous-time signals \( x(t) \) and called non-linear differential continuous-time or Teager-Kaiser energy operator which is defined as:

\[ \Psi_c[x(t)] \approx \dot{x}(t)^2 - x(t) \ddot{x}(t) \]  

(4)

Where \( \dot{x} = dx(t)/dt \). Second: Discrete time energy operator presentation of the signal, can be defined as:

\[ \Psi_d[x(n)] \approx x^2(n) - x(n+1)x(n-1) \]  

(5)

Where \( n=0,1...N \). Both operators were first introduced systematically by [5,6]. There is another method to compute the ‘energy’ operator. When the energy operator is applied to an AM-FM signal, it can approximately estimate the square product of the amplitude and frequency signals in a continuous time frame:

\[ \Psi[t,x(t)] = [a(t)w_{\text{inst}}(t)]^2 \]  

(6)

And its counterpart in discrete time form is:

\[ \Psi[m,x(n)] = [A(m,n)\Omega_{\text{inst}}(m,n)]^2 \]  

(7)

Where \( A, \Omega \) are the instantaneous amplitude and frequency.

From above, the energy operator can be calculated in the time domain using (4,6), whereas in the frequency domain (5,7) is used. In this study, an algorithm is based on the non-linear Teager’s Energy Operator with Signal detection / estimation theory through spectral analysis distribution techniques. The new algorithm is evaluated by reconstructing (2,5,7) as follows:

\[ NLDTEO(m) = \frac{1}{N} \sum_{n=0}^{N-1} \left[ \sqrt{(X_r(m,n))^2 + (X_i(m,n))^2} \right] n.Fr \]  

(8)

Where \( n \) is the time base index and Fr is the frequency resolution in comparison with (6) \( a(m,n) = \sqrt{[(X_r(m,n))^2 + (X_i(m,n))^2]} \) and \( \Omega_{\text{inst}}(m,n) = n.Fr \). The modified non-linear energy calculation in the spectrum is designed to be as follows

\[ NLEO(m) = AM (m,n) \cdot FM (m,n) \cdot Wf (m,n) \]  

(9)

Where \( AM(m,n) \) is amplitude component in the demodulated spectrum space, \( FM(m,n) \) is instantaneous frequency component in the demodulated spectrum space.

\[ NLEO(m) = [(X_r(m,n))^2 + (X_i(m,n))^2] \cdot (n \cdot Fr)^2 \cdot Wf (m,n) \]  

(10)

Where, \( Wf(m,n) \) is a weighted function for the power spectrum components for smoothing and is designed as follows:

\[ Wf(m,n) = \alpha + \cos(\beta + (2\pi \cdot n / \gamma)) \]  

(11)

Let \( Cf() \) as combined function be defined as:

\[ Cf(m,n) = Wf(m,n) \cdot (n \cdot Fr)^2 \]  

(12)

Then we can re-write equation (10) as follows:

\[ NLEO(m) = (AM(m,n))^2 \cdot Cf(m,n) \]  

(13)

Figure 1 shows the effect of variation of \( \gamma = G \) in (12).

**2.2 Algorithm evaluation and spectral analysis**

Proposed a sweep test signal with a frequency varying according to (14) as a function in time domain to evaluate the NLDTEO in (8). Fig.2a depicted as the characteristic of the AM-FM spectrogram and the contour of the energy operator in fig.2b. Fig3.a and b shows the NLDTEDO with spectrograph for real mandarin speech signal recorded in our lab..

\[ x(t) = \begin{cases} 
\sin[10000(t + 0.25e^{-4t})] & t \leq 1.2 \\
\sin[-1000(t - 0.25e^{-0.244t})] & 1.2 < t < 2.4
\end{cases} \]  

(14)

We believe that the robust speech features located in the frequency domain. Frequency and time filtering of filter-bank energies are also important in feature extraction process. The signal processing can be divided into two tasks: First, the
separation of the two components AM-FM signal into components using spectral signal analysis procedure Second, information extractions from the output of this technique, such as TEO, NLDTEO, Formant tracking, spectrums steady state, etc.

3. EVALUATION METHOD

To evaluate our algorithm we use the existing system database as followings:

3.1 Baseline System Parameters

The proposed method is implemented in THEESP [1], the speaker independent continuous mandarin speech recognition system based on the Duration Distribution Based Hidden Markov Model (DDBHMM) [7]. The acoustic model of this system is a

3.1.1 Acoustics features extractions

A high quality speech sampling frequency (16kHz) and Pre-emphasis filter H(z)=1-0.97z-1 is used in the baseline system. The acoustic features vector has a dimension of 45 components, Consisting of 14-dimensional Mel Frequency Cepstrum Components (MFCC), which are calculated on a 20-msec Hamming window advanced by 10 msec on each frame (50% overlapping). Radix-2 FFT of 512 points is performed in order to calculate a magnitude spectrum for the frame, which is averaged into 24 triangular bins arranged at equal Mel-frequency intervals. Finally, a cosine transformation is applied to such data to calculate the 14 MFCC. Moreover the normalized log NLDTEO component is also found, which is added to 14 MFCCs to form a 15-dimension static vector. The Static vector then expanded to produce a 45-dimensional (Static + Dynamic) vector upon which the DDBHMMs were built, and model the speech sub-word units were trained. Appending the first and second difference of the static coefficient as 14-dimensional $\Delta$-MFCC, 14-dimensional $\Delta\Delta$-MFCC, $\Delta$-EO, and $\Delta\Delta$-EO components extends the static vector

3.1.2 Acoustic phonetic modeling

Type: DDBHMM.'s are trained before they are used for segmentation and fixed while the post-end distance classifier is trained. Structure: Mandarin continuous speech recognition classified as VCV, CV. Syllable or/sub-syllable is chosen as the base unit of the designed recognizer. Each syllable is represented by one initial and one final phone with 2-states Left-to-Right for initial syllable phones, 4-states Left-to-Right for final syllable phones, variable length model as depicted in figure 4. Emission Densities: For each state in the DDBHMMs Single Gaussian distribution with Full Covariance Matrix $C_v$ is used. Statistical representation: DDBHMMs Left-To-Right. Training: Segmental k-means clustering, distortion threshold value is (0.0005).

3.1.3 The classifier

Number of distributions $r(k) =$ the number of reference vectors assigned for covariance matrix ($C_v$) Number of classes: 856 + One state for continuous +1 for silence. Number of Training (Ntr) tokens for all the classes: Training epochs Er: represents the procedure where all the training token is used one to train the classifier. Matrix space: [856x45] Dimensions time normalized vector space. In total there are 856 DDBMHMM's states (Str) to be trained (i.e. Str= 856).

Figure 2.(a) Sweep test signal (b) Contour of the test signal

Figure 3.(a) Spectrograph of speech signal (b) The NLDTEO
3.2 Modified THEESP Parameters

The main improvements that occurred on the base line system as described in article 3.1 are, **first**: The Traditional technique used to compute the energy in the baseline system is replaced by new Non-Linear Techniques. **Second**: Use the formula in (12) to adjust the spectrum distribution for different values of G. **Third**: Combine the NLDTEO with MFCC, and then extract the first and second derivative of the new energy operator combine with the feature vector. **Fourth**: In the front-end processing stages the pre-emphasis filter is completely eliminated in our experiments.

![Diagram](Figure 4: DDBHMM For Large vocabulary speech recognition)

3.3 Speech Data Base

The speech corpus used in these experiments comes from 54 data files recorded with the vocabulary of 7493 words in each file with a total 646 sentences in each file containing the most confused syllables. The data is divided into two groups according to the experiment conditions, first 50 of them are used for training, and other 4 data files were used for testing and evaluating the system in all experiments for the baseline and improved system. The experiments were carried out on a large-vocabulary task based on People's Daily, the training corpus is provided by National 863 High Technology Project. The testing corpus is the database established by our laboratory

Table 1. Experiment result

<table>
<thead>
<tr>
<th>Male Data Files</th>
<th>Recognition Rate: Use 50 data Files for training and 4 data files For testing</th>
</tr>
</thead>
<tbody>
<tr>
<td>Number of Candidates</td>
<td>LVCSI-AMSR Using TDEO</td>
</tr>
<tr>
<td>-------------------</td>
<td>-------------------------</td>
</tr>
<tr>
<td>1</td>
<td>5</td>
</tr>
<tr>
<td>M96</td>
<td>53.00</td>
</tr>
<tr>
<td>M97</td>
<td>66.53</td>
</tr>
<tr>
<td>M98</td>
<td>69.34</td>
</tr>
<tr>
<td>M99</td>
<td>51.27</td>
</tr>
<tr>
<td>Average</td>
<td>60.03</td>
</tr>
<tr>
<td>Relative Average Error Rate Reduction %</td>
<td>5.643</td>
</tr>
</tbody>
</table>