HMM-based techniques for speech segments extraction

Waleed H. Abdulla[*]
Department of Electrical and Electronic Engineering, The University of Auckland, Auckland, New Zealand

Abstract. The goal of the speech segments extraction process is to separate acoustic events of interest (the speech segment to be recognised) in a continuously recorded signal from other parts of the signal (background). The recognition rate of many voice command systems is very much dependent on speech segment extraction accuracy.
This paper discusses two novel HMM based techniques that segregate a speech segment from its concurrent background. The first method can be reliably used in clean environments while the second method, which makes use of the wavelets denoising technique, is effective in noisy environments. These methods have been implemented and shown superiority over other popular techniques, thus, indicating that they have the potential to achieve greater levels of accuracy in speech recognition rates.

1. Introduction

The increasing power of computers, combined with newly developed computational techniques, has contributed to the improvement in the performance of speech recognition systems. Current approaches make use of the advancements in neural networks, statistical models, and signal processing techniques, to develop powerful hybrid models for speech recognition applications. There are many developments that need further investigation in this field. A vitally important objective in implementing an isolated words speech recognition engine, commonly used in voice command systems, is the accurate separation of the signal of essence from its background environment. The success of this process has a crucial effect on the overall performance of isolated words automatic speech recognition (ASR) systems. It is an issue researchers have tackled since studies were first carried out in this field. In some speech recognition techniques, such as the dynamic time warping technique [27], it is necessary for the incoming spoken utterance to be as free as possible from non-speech regions to avoid such regions causing mismatching between the input and the stored templates. Also, the accurate detection of a word's start and end points means that subsequent processing of data can be kept to a minimum. The detection of the presence of speech in a background environment was classically referred to as the endpoint detection (EPD) problem [25].

The problem of detecting endpoints would seem to be relatively trivial, but, in fact, it has been found to be very difficult in practice, except in the case of very high signal to background-noise ratio (SNR). Some of the principal causes of endpoint detection failures are weak fricatives ($/f/, /h/$) or voiced fricatives that become unvoiced at the end ("has"), weak plosives at either end ($/p/, /t/, /k/$), nasals at the end ("gone"), and trailing vowels at the end ("zoo"). Thus to perform well, the algorithm must take a number of special situations into account, such as:

- Words that begin or end with low-energy phonemes (weak fricatives).
- Words that end with an unvoiced plosive.
- Words that end with a nasal.
- Words that end with a fading in intensity or a short breath.

An earlier commonly used technique used explicit features for speech non-speech discrimination such as speech signal energy and zero-crossings rate [9,21,25]. This technique is effective in the case of a low noise en-

[*] Adress for correspondence: Department of Electrical and Electronic Engineering, The University of Auckland, 20 Symonds Street, Private Bag 92019, Auckland Central, New Zealand.

ISSN 1058-9244/02/$8.00 © 2002 – IOS Press. All rights reserved

vironment, but unreliable with increasing noise and varied articulation manners such as breathing and clicks. Another approach was the pattern classifications of voiced, unvoiced and silence segments [5]. This technique implies a decision making process to improve the performance of the system but it incurs more computational load with little improvement.

Wilpon et al. benchmarked a multi-speaker digit recogniser to evaluate the effect of misalignment of word boundaries on the recognition rate [29]. The words and the reference patterns were manually extracted. The recognition rate was found to be 93%. Then a misalignment procedure was practised with the recognition error measured at each step. A similar experiment has been replicated on our system to study the recognition rate degradation due to different forced misalignments. Figure 1 shows the contour plot of the spoken digits recognition performance under different start-end constraints. The recognition rate degraded from 99% to 75% due to the signal boundary misalignment. It can be noticed that the start point misalignment allowance is less than that of the endpoint.

Recent techniques dealt with pre-silence and post-silence periods as pre and post states of hidden Markov models (HMMs). Silence her doesn't mean noise free but it could be a background signal. In this paper, background and silence will be used interchangeably. During the training phase the word modelling is carried out without including the terminal silence periods, and the silence periods are modelled as separate states. In the recognition phase, the pre and post-silence states are concatenated to the initial and final states of the words' models. Then the maximum likelihood (or any other optimisation) procedure is followed to identify the tested words. These HMM techniques, even though they are effective, still need to concatenate the silence states during the recognition phase. This consequently increases the computational cost, especially with long silence periods and increasing the number of models. In addition, the different spikes that might be issued during silence periods such as lip flaps will be embedded in the final calculation of the model likelihood which, in turn affect the ASR performance.

The interest in speech-background discrimination has intensified lately due to the increasing demand for potential use in some commercial systems. Current personal communication systems such as a cellular phone are examples of commercial systems that integrate speech recognition capabilities in their operation. These systems normally require voice commands to control them. The spoken commands need to be accurately extracted from the background to process them. The most wanted techniques are those that can work in adverse conditions such as in the car, office and other noisy places [6,20,28]. These techniques are still mostly based on measuring signal energy and the zero-crossings rate.

This paper illustrates two different novel HMM-based techniques to segregate a speech segment from its background. The first method can be reliably used in clean environments while the second method, which makes use of the wavelets denoising technique, is very effective in noisy environments. These two methods have been implemented and have shown superiority over other popular techniques, thus indicating that they have the potential to achieve greater levels of accuracy in speech recognition rates. This paper is organised as follows: Section 2 introduces the wavelet-denoising concept, which is used in one of our approaches in speech extraction. Section 3 demonstrates word extraction modelling and describes two HMM-based speech segment extraction techniques. Section 4 depicts the results of an evaluation study of the two techniques based on the computation cost and discusses the suitability of each of them. Section 5 evaluates the two techniques from a speech recognition perspective. Finally, Section 6 derives final conclusions from the research.

2. Signal denoising

Denoising is a technique for rejecting noise by damping or thresholding it in the wavelet domain [11]. It is used in wavelet literatures as a counterpart to the traditional terms of noise cancellation, noise reduction, and noise suppression used in the signal processing field. Speech background discrimination can be greatly improved by using wavelet techniques to mute noise. Wavelets have proven effective in denoising the signals from different types of noise and are even better than the traditional methods [10,12]. The idea of denoising exploits the approximation property of the wavelet coefficients. It states that only a small number of wavelet coefficients carry most of the signal power. This means that any signal can be accurately reconstructed from only a small number of coefficients. Using this notion, the low-level wavelet coefficients of the details components can be set to zero by suitable nonlinear function to eliminate the noise while keeping the signal unsusceptible. Practically, the signal is partially affected by

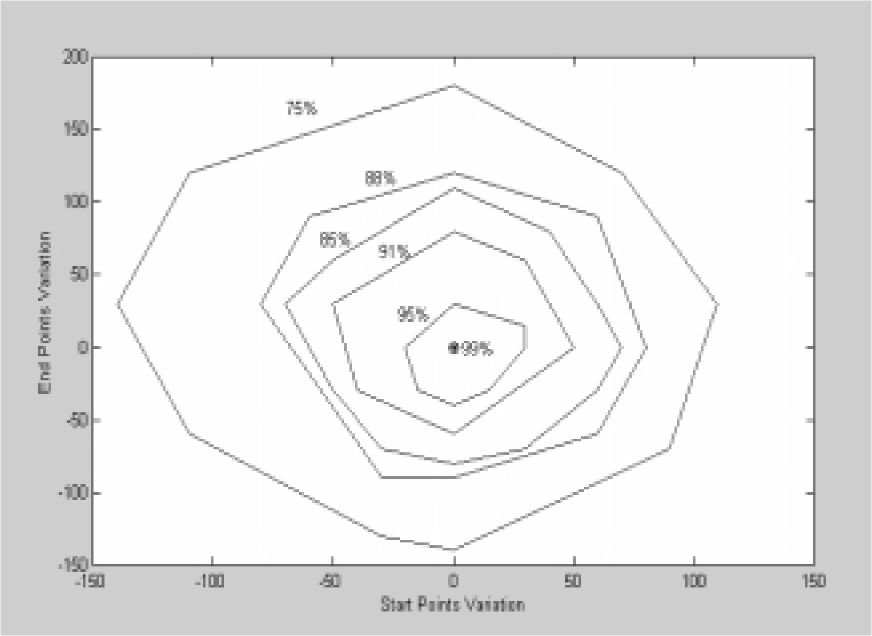


Fig. 1. Recognition performance as a function of start and end points detection.

the denoising process and this is equally applied to all the filtering techniques.

The above denoising method is called wave shrink and it is mostly effective for additive Gaussian noise governed by the following equation:

$$x(n) = u(n) + \delta.w(n) \tag{1}$$

The task of denoising is to recover the clean signal $u(n)$ from the noisy input signal $x(n)$ by suppressing the noise part $\delta.w(n)$. Where δ is the noise level, and $w(n)$ is the noise signal.

The main steps in the wave shrink algorithm are:

1 – Signal decomposition [8,16,22]

The signal is decomposed using wavelet transformation with L levels of decomposition to obtain the approximations, A, and the details, D, coefficients. Details of all of the levels of decomposition and of the last level approximations are selected for further processing, i.e. $D^1, D^2, \ldots, D^L$ and A^L.

2 – Thresholding

The absolute values of details coefficients below a carefully selected threshold level, Δ, are reset to zero, and this is called hard thresholding. This thresholding leaves unwanted discontinuities at both ends of the thresholding function, which makes the process lossy, and the original signal cannot be reconstructed exactly.

The mathematical representation of the hard thresholding function at level i, Y^i_{hard}, can be represented by,

$$Y^i_{\text{hard}} = \begin{cases} D^i & \text{if } |D^i| > \Delta \\ 0 & \text{if } |D^i| \leqslant \Delta \end{cases} \tag{2}$$

To remove the discontinuity from the thresholding function, a denoising process called wave shrink has to be applied. During shrinking all the left details coefficients from the hard thresholding are pulled to zero level by an amount of Δ. This is also called soft thresholding and can be mathematically represented by,

$$Y^i_{\text{soft}} = \begin{cases} \text{sign } (D^i).(|D^i| - \Delta) & \text{if } |D^i| > \Delta \\ 0 & \text{if } |D^i| \leqslant \Delta \end{cases} \tag{3}$$

The denoising process needs the value of Δ to be carefully selected. Setting Δ small will result in a leakage of outliers into the signal, while setting it large will cause poor denoising performance. There are several techniques used to determine the suitable value of Δ [10,12]. The method used to determine the value of Δ depends on the nature of the signal to be denoised and it could be fixed or varied. In our problem, it is suitable to select the value of Δ fixed for all the details levels, and according to the following formula,

$$\Delta = \frac{\delta}{\sqrt{n}} \cdot \sqrt{2\log(n)} \tag{4}$$

where n is the signal length and δ is the noise level. The noise level can be estimated by using the wavelets prop-

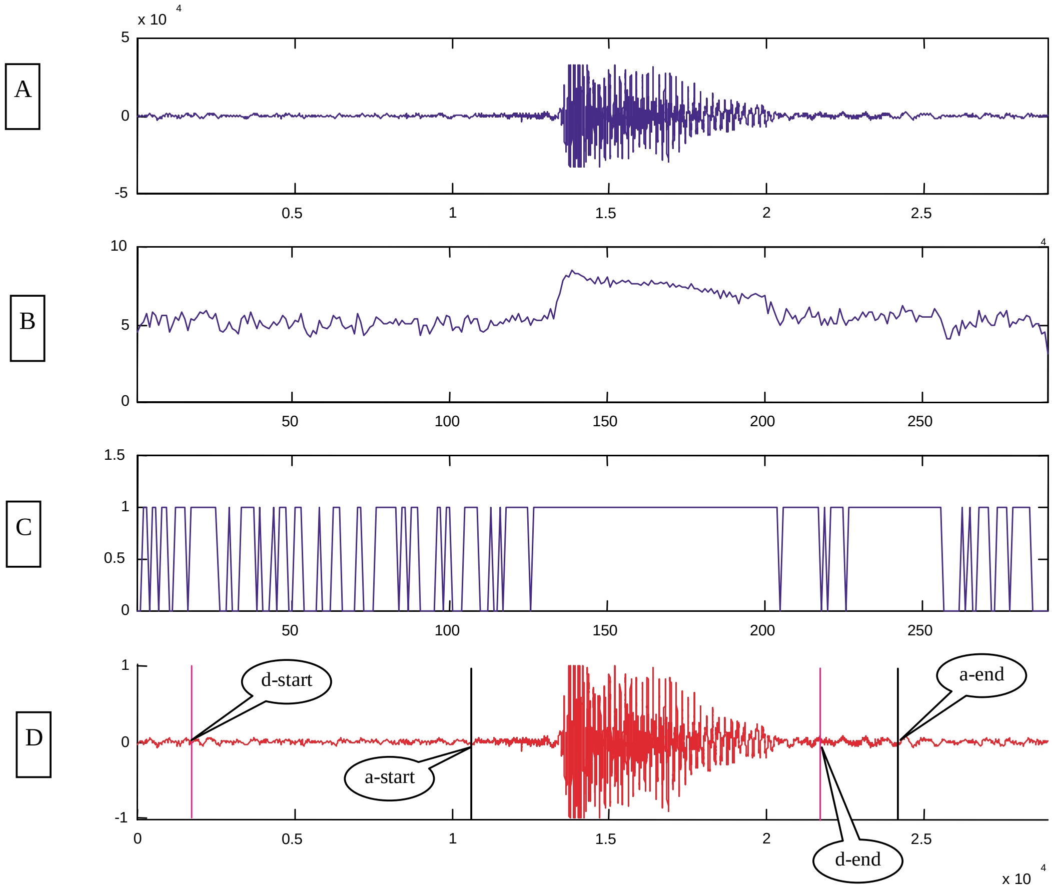


Fig. 2. Endpoints detection using the short-term energy technique. A: The input signal, B: The energy level of the signal, C: Over-Threshold energy level, D: Normalised signal with word boundary detected. d-start and d-end for the detected boundary, a-start and a-end for the actual boundary.

erties. The details coefficients D^i are mainly noise coefficients with a standard deviation δ. The engineering favourite estimate of δ is the median absolute deviation as it is more robust to the outliers.

The value of Δ can also be selected as level dependent by using the δ value of that level and applying the same above Eq. (4).

3 – Inverse wavelet transform

In this final step, the inverse wavelet transform is applied on the shrinked coefficients from step 2 to reconstruct the original signal, with suppressed noise.

3. Word extraction modelling

This section introduces two different HMM based novel techniques to build two models for extracting the spoken words, speech segments, from their background environments. We then will evaluate each technique experimentally with the aim of preparing the exact spoken signal for the word recognition models.

The datasets used for training the two models are the same. It comprises 20-50 different words spoken by different speakers in different environmental situations. The pre and post-silence periods contain different noise levels as well as some artefacts such as lip slaps, breaths, and microphone clicks. To increase the robustness of the word extraction models, we used different microphone types in recording the training dataset.

The relevant acoustic features used here are constructed from the energy and the Mel frequency cepstral coefficients (MFCC) [26]. These capture the static behaviour of the speech segments. The first order temporal derivatives, velocity coefficients of the energy and the MFCC are not used in the silence detection step since they are responsible for capturing the dy-

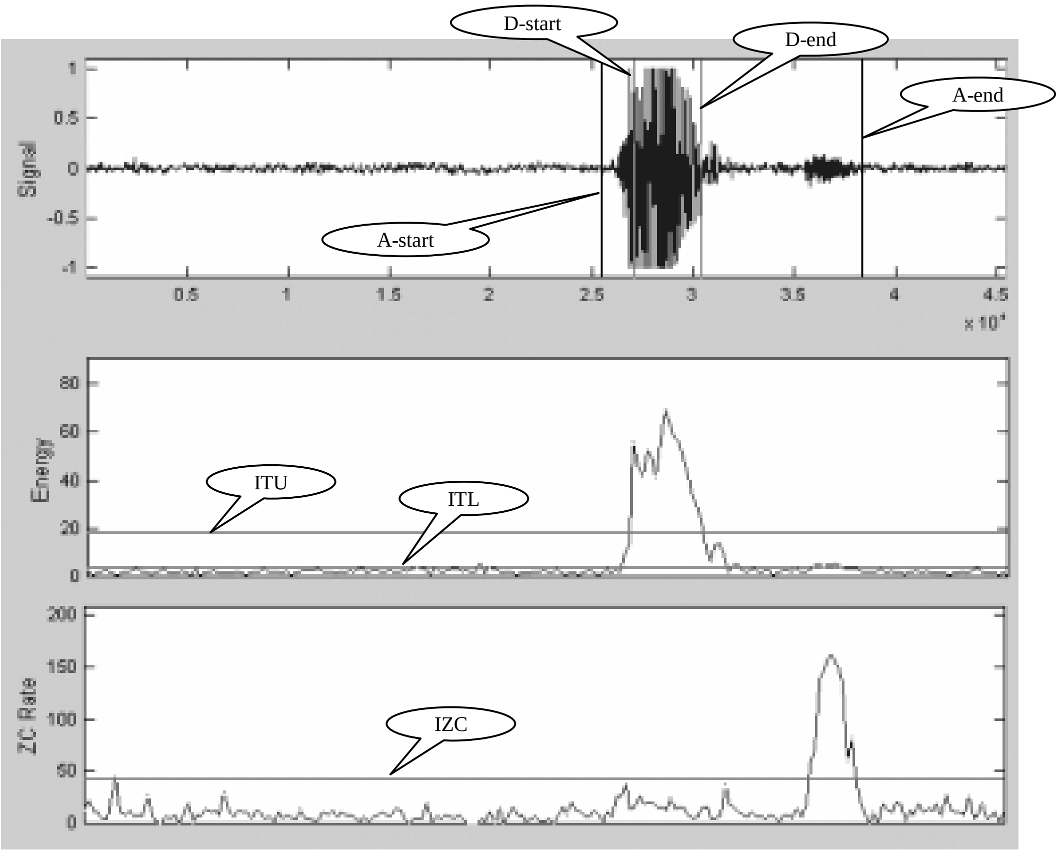


Fig. 3. Endpoints detection of a spoken digit "eight" using energy /zero-crossings technique. D-start and D-end are the detected start and end points of the signal, A-start and A-end are their corresponding actual start/end points, ITU and ITL are the upper and lower thresholds of the signal energy, IZC is the zero-crossings threshold.

namic behaviour, which has no importance here [14, 15]. The same thing can be said about the irrelevance of the second order temporal derivatives, acceleration coefficients, in our methods [18,19,23].

3.1. Commonly used speech segment extraction techniques

For the sake of comparison with our own techniques, we studied the performances of two commonly used classical techniques. One classical technique uses the energy level for the detection of endpoints of the speech segment. This technique estimates the short-term energy measure, $E(m)$, for the N-length frame of signal, $s(i)$, ending at time, m, according to the following equation,

$$E(m) = \sum_{i=m-N+1}^{m} s(i)^2 \tag{5}$$

Then, a threshold value for the energy level of the silence period is determined. Whenever a signal crosses above this level, it is considered as a speech segment. This technique is simple and performs reasonably well in a very low noise environment. It is also necessary to use a high quality, close contact microphone with a noise suppression facility. Figure 2 shows how the energy threshold technique performs when a low noise signal is presented to it. We tried to improve this technique by putting more elaborative constraints on the threshold level crossings to improve the detection of the silence and speech periods. Some improvements were achieved but these did not significantly enhance the accuracy of the speech recognition task.

In a second classical technique referred to as EZC, the performance is improved by using, in combination with the energy measure, another feature called a zero-crossings measure to segment a signal into speech and silence regions [25]. Figure 3 shows the short-term zero-crossings and energy measures plotted for the spo-

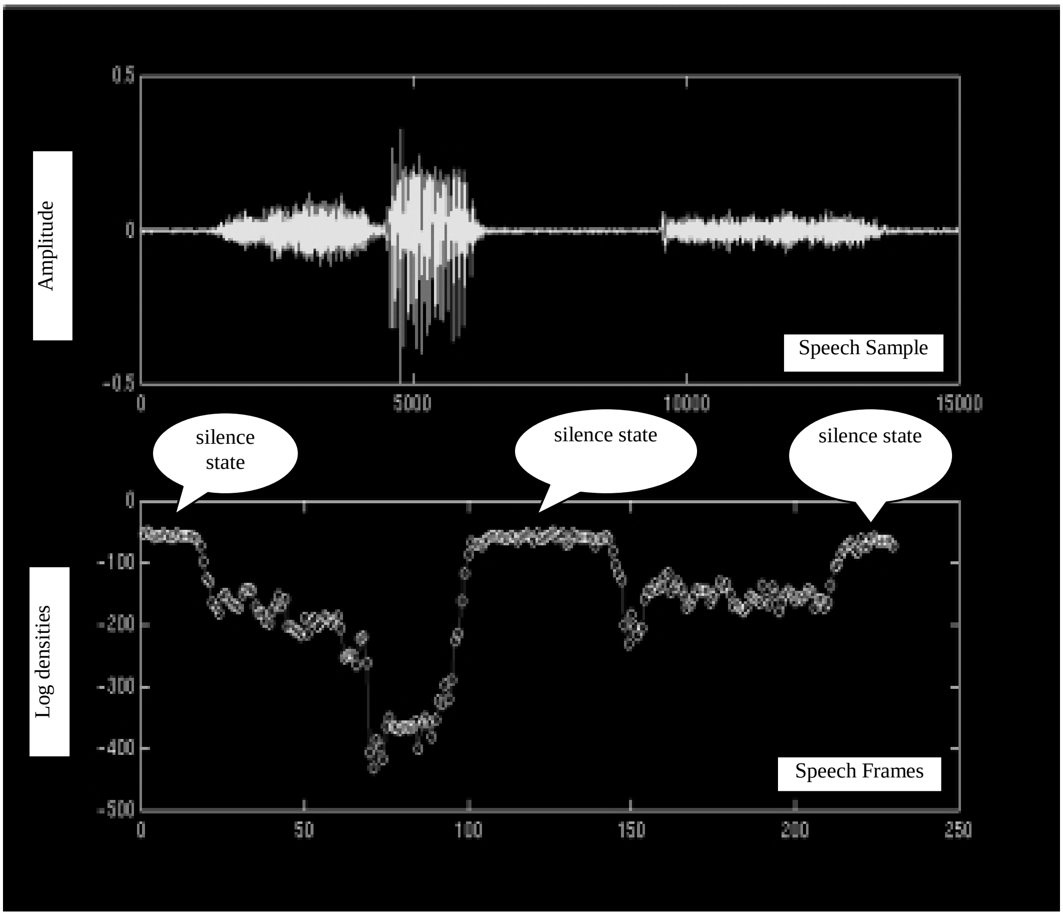


Fig. 4. Silence states in the spoken digit “six”. The silence states are identified by their high log likelihood probability densities.

ken digit “eight”. The frame rate is 100 frame/sec and the frame length is 10 ms.

It is assumed that the first 10 frames are backgrounds, and this is one limitation of the method. They are used to find the mean and the variance of each of the features. In turn, these statistics are used to set upper and lower thresholds, ITU and ITL, as shown in the figure. The method proceeds as follows. Firstly, search from the beginning until the energy crosses ITU. Then, backoff towards the signal beginning until the first point is reached at which the energy falls below ITL. This is the provisional beginning point. The provisional endpoint is detected in a similar way but starting from the end of the signal. For the actual beginning point, now examine the previous 250 ms of the signal’s zero-crossings rate. If this measure exceeds the threshold value IZCT for 3 or more times, the provisional beginning is moved to the first point, at which the IZCT threshold is exceeded. Again, perform a similar procedure for the endpoint.

The energy and zero-crossings method is commonly used in many systems, as it is a straightforward, easy to implement technique and is reasonably effective in clean environments. However, this technique still suffers from susceptibility against any slightest noise. This weakness can be seen by testing its detection ability to a recorded signal of digit “eight” spoken in an office environment using a low quality microphone (to degrade the signal) as in Fig. 3. It is apparent that the endpoints are erroneously detected, especially at the trailing edge. We notice that although the zero-crossings contour at the trailing edge of the speech segment is prominent, the algorithm neglects it since it appears a bit later than the 250 ms margin used in this technique.

3.2. *HMM based segmentation method* (*HMMseg*)

We have seen that when we use the static Mel scale coefficients as feature vectors in modelling the continuous density hidden Markov model (CDHMM) of any

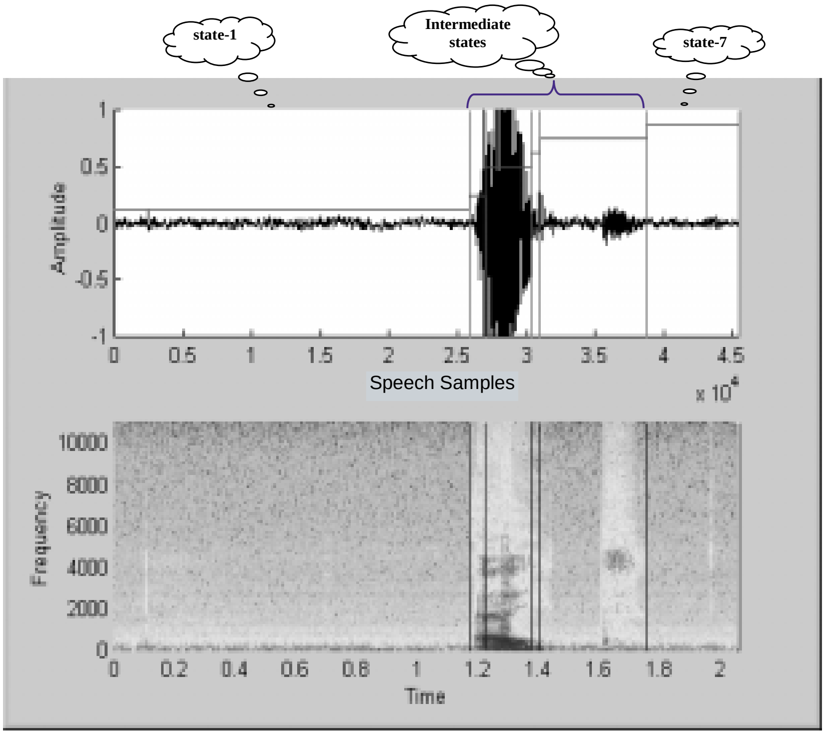


Fig. 5. Endpoints detection using 7 states CDHMM word extraction model (HMMseg).

speech signal, then the pre-silence and the post-silence periods occupy the first and last states respectively [1, 2]. The optimum number of states for successful modelling is 7 since a lesser number of states might mix the silence state with the starting or the trailing segments of the speech signal. More than 7 states will only increase the computational cost with no improvement. The intermediate states, states 2 to 6, have no relevant meanings in these models. They only perform a bridge to reach the last state of the speech signal. The topology of the CDHMM used in this method is of the left-to-right paradigm in which the state transition is constrained by a self-feedback and a left-to-right move with one state skip permitted.

The observation sequence comprises static feature vectors of one power and some of the first MFCCs, totalling 13 coefficients. For best signal tracking in state changing, the frame length was chosen to be 11.6 ms taken each 3 ms. The speech signals were preemphasised and the MFCC feature vectors were cepstrally mean normalised by subtracting their means during the training process which increased robustness toward the channel and the environment variability [17].

The extraction of the relevant signal was simply done by removing the samples, from the original input signal, belonging to the first and the seventh states while keeping the speech samples of the intermediate states for the recognition process. The states were determined by using the backtracking phase in Viterbi algorithm [24]. We will call any model used to detect the silence periods in any input signal for the above-mentioned task a word extraction model.

The tagging of the silence periods to certain states within the HMM structure can be consolidated by looking at the likelihood of detecting the silence periods in any input signal as presented to a single state model of the silence periods. Figure 4 shows the likelihood of detecting the silence periods in the spoken word “six” as presented to a single state silence model.

To evaluate how the word extraction model designed by this method performs, we presented several types of input signals and monitored the matching between the detected and actual speech boundary. As a comparison with the classical technique EZC, the HMMseg technique showed precise results when the signal of Fig. 3 was presented to its word extraction model as depicted in Fig. 5. In the HMMseg method, we don’t need any assumptions about the signal and we don’t need the preamble silence segment to do the statistical measurements. The characteristics of the speech and non-speech segments have already been implied within the model itself during the training procedure.

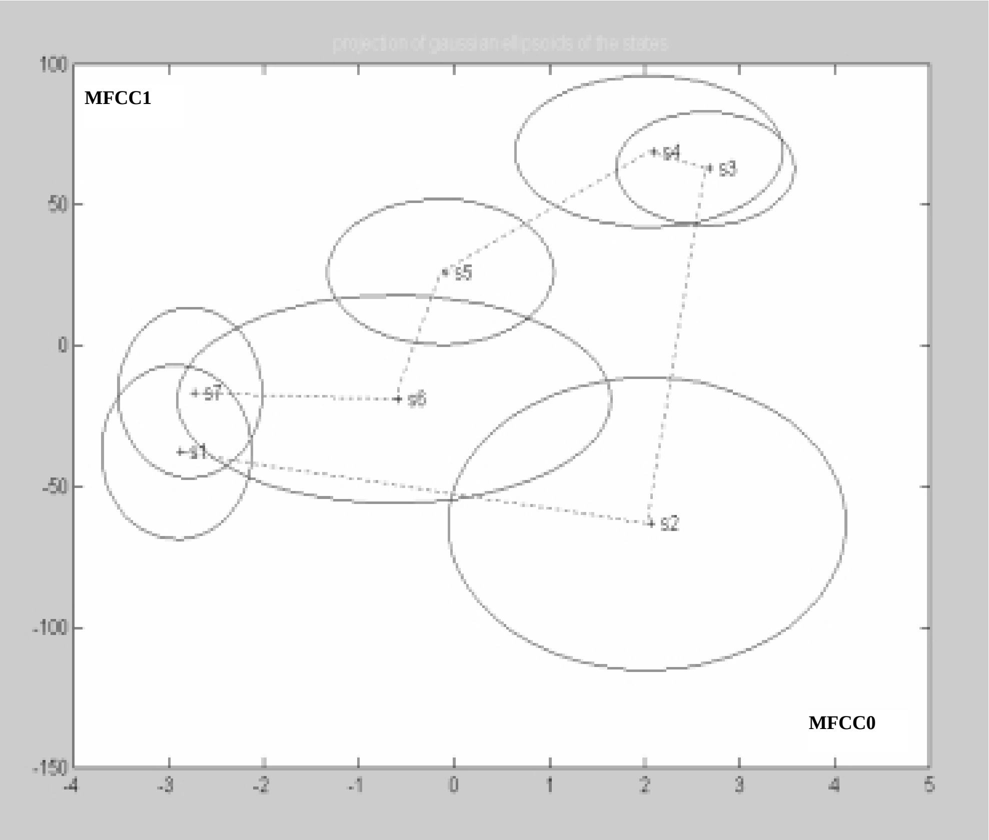


Fig. 6. Gaussian ellipsoids of the 7 states HMM word extraction model The power measurement, MFCC0, is plotted against the MFCC1.

To visualise the clustering behaviour of the states, the Gaussian ellipsoids graph is used. This graph is very useful in studying statistical models, and it represents the locus or the contour of the points that have equal probability for each cluster. In case of the Gaussian probability distribution, the locus can be found by equating its exponential term to a constant C that can be mathematically formulated by,

$$(x-\mu)^T\Sigma^{-1}(x-\mu) = C \qquad (6)$$

where x is the feature vectors set, Σ is the covariance matrix, and μ is the cluster mean vector. By taking different values for the constant C, we can plot several concentric contours.

Figure 6 shows the Gaussian ellipsoid of the word extraction model. From the clustering contours we can see the similarity between the first and the seventh states. We can also see that state 6 overlaps in places with states 1 and 7. This overlapping indicates that the trailing edge of the speech segment partially shares some spectral characteristics with the silence periods. Breathing and some non-speech signals are normally released at the end of the spoken words that are responsible for such overlapping. The other issue that we need to focus on is that the feature vectors are of dimension 13 and we can only plot the Gaussian ellipsoid of the two dimensional vectors. To compensate this problem we draw the projection of the feature vectors from different perspectives, on different planes, and share the power element, the first element in the feature vector, in all cases. This is because the power is a strong cue in determining the clusters and this is why classical techniques heavily rely on this measurement.

We have experimentally observed that the first, power measurement MFCC0, and the second, MFCC1, elements of the feature vectors are sufficient to study clustering behaviour. Figure 7 shows the projection of the Gaussian ellipsoids on different planes and how they are completely consistent with each other in determining the silence states. The clustering behaviour of states 1, 6 and 7 is consistent in all cases, while this consistency has been violated for the other states. There is no overlapping region between pre-silence state 1 and the first arrivals of the speech samples state 2. This in-

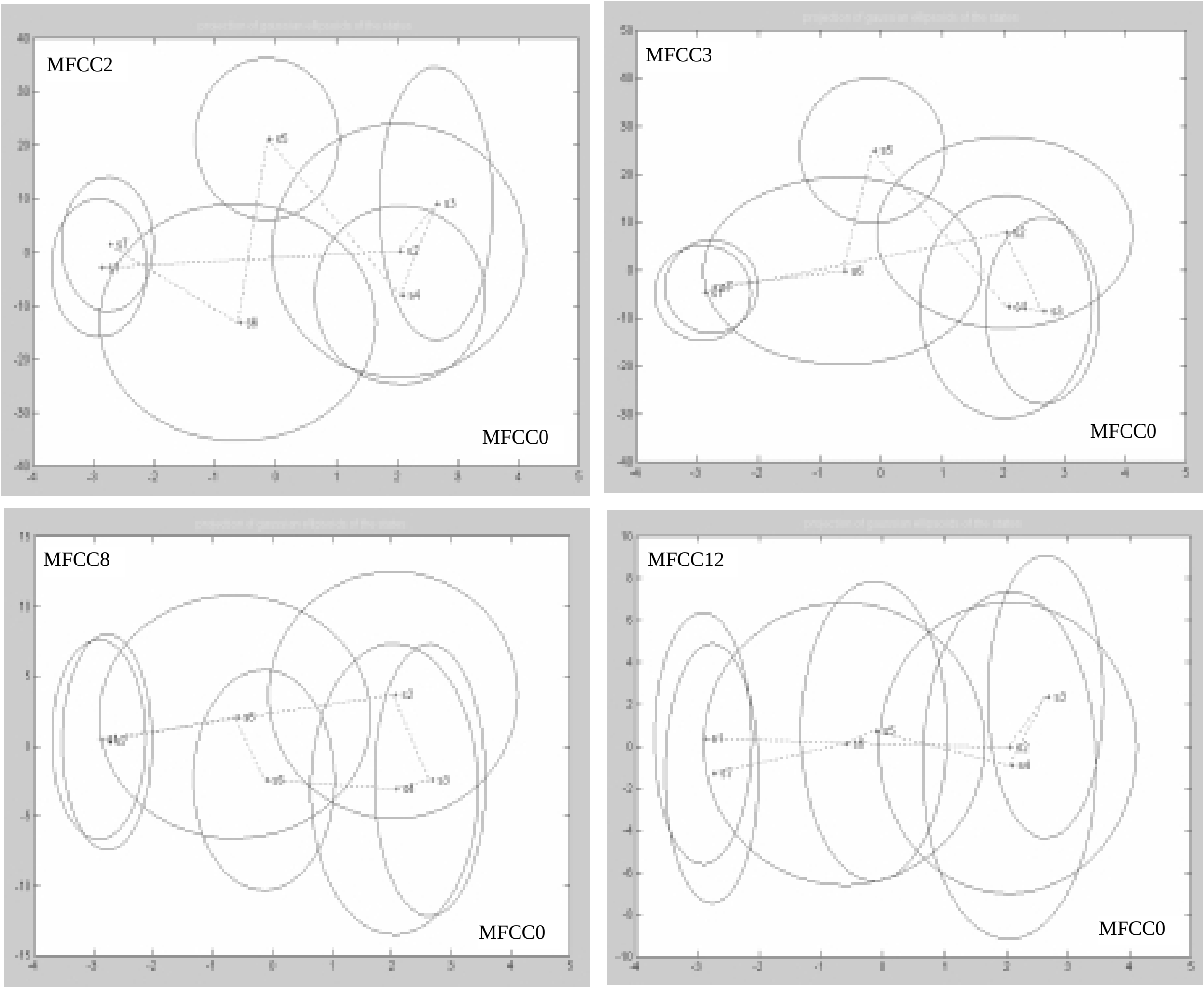


Fig. 7. The Gaussian ellipsoids of the 7 states HMM word extraction model. The power measurement, MFCC0, is plotted against different MFCCs.

dicates that the detection of the beginning of the speech segment is almost perfect. This is very important in the speech recognition process since the allowance of the erroneous detection of the speech segment leading edge is lower that that of the trailing edge, as indicated in Fig. 1. It can be noticed from this figure that the horizontal margin between the extreme points in any polygon is less in time than the vertical margin of that polygon. For example, in the outer polygon, the horizontal margin is 250 ms while the vertical margin is 320 ms. This means that for a recognition rate of 75%, the allowance in the misalignment of the start point is 250 ms around the exact start point while the corresponding allowance in the misalignment of the end point is 320 ms around the exact end point.

3.3. *HMM based segmentation with denoising method (DNHMM)*

The procedure developed in HMMseg is successful and adequate in many situations but still not immune to the increasing noise level. Our goal here is to modify HMMseg to model the incoming signal into three distinctive states representing the pre-silence, speech, and post-silence segments respectively (recalling that silence refers to the background and it might represent noisy segments of the input signal). This goal can't be modelled directly since the speech segment itself implies many different stationary spectral regions, which are translated into distinctive states as we have seen in HMMseg. If we try to use HMMseg directly and ask for a three states model to map the input signal into

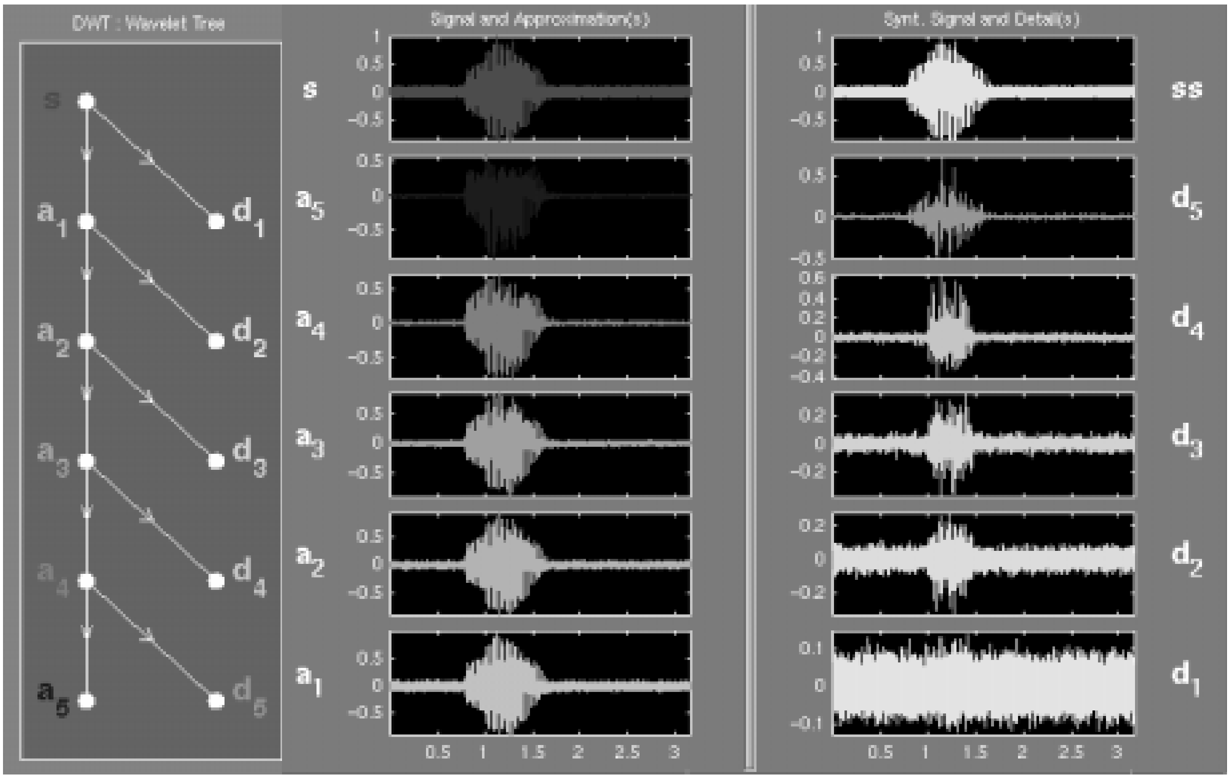


Fig. 8. A decomposition tree and reconstructed approximations and details of 5 levels of depth. s is the original signal, a_i and d_i are the approximations and details of level i respectively.

3 states directly, the model will represent each set of contiguous spectrally similar segments by a state which can be characterised as a merging property in modelling the contiguous most similar spectral regions. This will produce a chaotic situation in which the pre and post-silence segments merge with some of the beginning and trailing speech phones.

One possible procedure to achieve our goal is to suppress the changeability between different spectral regions within the speech segment and within the two silence segments so that each segment can be modelled by one stationary state according to the merging property. The aim of this section is to implement a successful procedure, which will progress towards a more robust technique in speech/silence discrimination following the above-mentioned notions. According to this procedure, a separate 3 states continuous density HMM (CDHMM) is built to efficiently discriminate the needed speech signal from the unwanted background environment. The model has the same specifications described in HMMseg regarding the topology and observation sequence type. The basic idea is to build a model that can map the input stream into three states representing the pre-silence, speech, and post-silence segments respectively. Then, the extraction of the signal is simply done by removing, from the original signal, the input samples belonging to the first and third states, while keeping the speech samples of the second state, for the recognition process.

Introducing a wavelet denoising process before modelling can dynamically suppress the different regions [3]. Denoising has a strong effect in helping the word extraction model differentiate the speech state from background states. This new model can efficiently discriminate the speech signal from the two coherent pre and post-silence segments. However, it cannot discriminate the inter-silence periods within the speech signal.

The training dataset used to prepare the word extraction model was taken from 50 words spoken in isolation by multi speakers in different environments. The signal was firstly denoised using a biorthogonal wavelet (bio2.2 which is one version of different possibilities of the biorthogonal wavelets) with a level of decomposition of 16, to mute the noise perfectly before starting the word extraction procedure [8]. Other wavelets

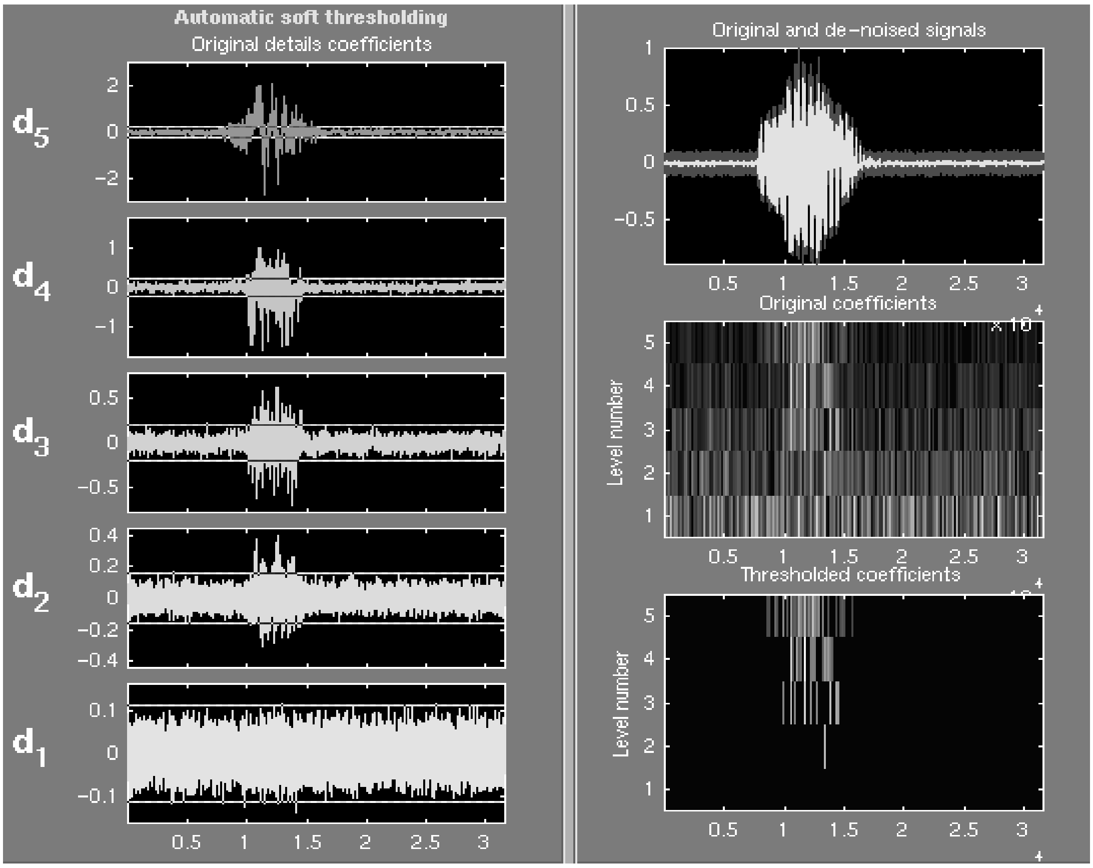


Fig. 9. Details reconstructed signal and thresholding levels (left), reconstructed denoised signal superimposed on the original noisy signal (upper right), number of details coefficient before and after thresholding (lower right).

such as symlets and Daubechies with a different level of decomposition can also be used, but our selection shows the best denoising performance in most of the environments. Figure 8 shows the decomposition tree of a noisy speech signal; only five levels are shown for illustration. This figure also shows the waveform at each level of decomposition of a spoken digit "9" in a noisy environment with a signal to noise ratio = 16.5 db.

The successive approximations become less and less noisy with the increasing depth of decomposition, because at each level of decomposition, more high frequency information is filtered out of the signal. Level 5 approximations, a5, looks clean compared to the original signal, s. However, we cannot only use the approximations' coefficients for denoising purposes, because in discarding all of the high frequency information by removing all of the details coefficients, we lose the sharpest features of the signal. This means that we will lose the fricative segments from the speech signal as we go deeper in decomposition levels.

Optimal denoising, as described in Section 2, is used when only the portions of the details that are below a certain limit (threshold) were discarded. Figure 9 depicts this technique and shows the discarding limits, the two horizontal lines, in each level of decomposition. It is clear from Fig. 9 that the details of level 1 represent a noise signal and thus they are completely discarded from the signal. The upper right side frame shows the denoised signal superimposed on the original signal. The lower right two frames feature the comparison of the number of wavelet coefficients before and after denoising.

The MFCC feature vectors are then extracted from the denoised signal. Each vector is composed of 13 coefficients (12 MFCC and one power coefficient). The window frame length was chosen to be 23 ms taken each 9 ms, which is longer in time and faster in processing than that used in HMMseg, since the signal dynamical behaviour varies more slowly with time due to the denoising effect. The feature vectors derived from several examples of denoised signals are used to train a

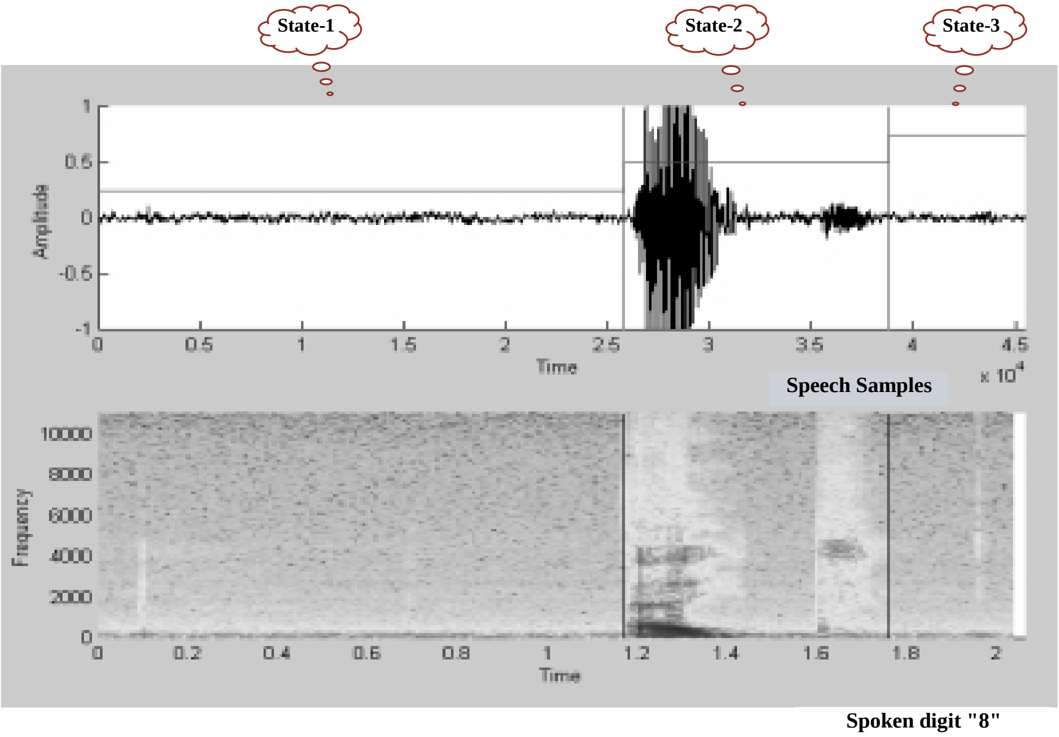


Fig. 10. Performance of the denoised word extraction model (DNHMM). (left) Spikes within pre-silence period, (right) Detection of the same signal processed in Fig. 5. States 1 and 3 correspond to the pre and post-silence periods, while state 2 is the relevant speech segment.

3-state HMM word extraction model. Figure 10 shows the performance of the trained word extraction model in segregating the same speech signal of Fig. 5 using DNHMM method.

The denoising process also makes the pre and post-silence regions gather into very similar clusters, which enables us to use the same parameters for the pre and post-silence states. This claim is consolidated by Fig. 11 that shows the projections of Gaussian ellipsoids of the word extraction model using symlets wavelets (sym4). It is obvious how similar the clusters of the pre and post-silence data are. These two clusters correspond to states 1 and 3 respectively, while the speech data cluster correspond to state 2. The projection of the Gaussian ellipsoid over different planes provides the same clustering results, as is the case in HMMseg. The post-silence state discrimination from the speech signal state is improved over that in HMMseg, which can be seen from the disappearance of the overlapping region between states 2 and 3. However, the post-silence state, state-3, is still closer to the speech signal state, state-2, than the pre-silence state, state-1.

The clustering capability of the word extraction model can be improved by using other types of wavelets. The performance of each model can be directly verified from the comparison of the detected and actual endpoints of different spoken words, which is a long and time consuming procedure. The best way to make this comparison in model evaluation is by plotting the Gaussian ellipsoids and investigating the clustering regions. Figure 12 shows the Gaussian ellipsoids of two models, one based on symlets wavelets (sym4), depicted by the dotted lines, and the other on biorthogonal wavelet (bio2.2), represented by the solid lines. It is clear that the biorthogonal wavelet model has better properties for separating the clusters of speech and silence regions. We adopted the speech/silence model based on this latter wavelet in all of our speech recognition systems after investigating many other types of wavelets similarly.

4. Evaluation of HMMseg and DNHMM from the computation cost aspect

From the speech segment detection accuracy perspective, the HMMseg and DNHMM methods perform similarly in a low noise environment. This means that

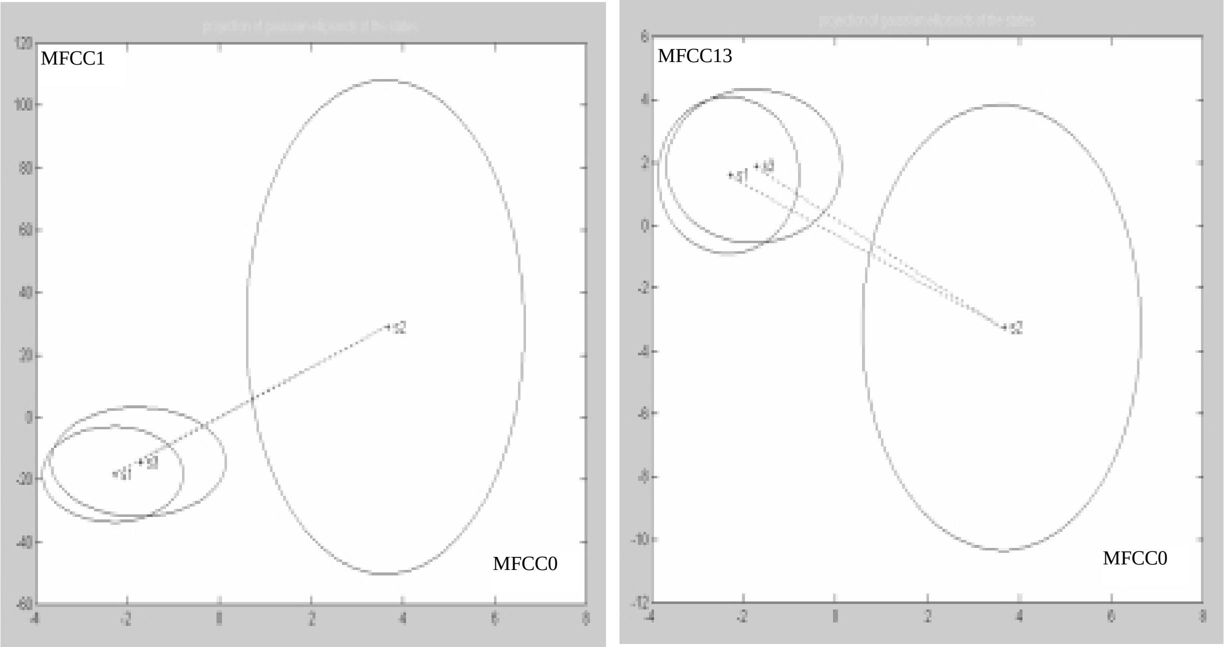


Fig. 11. The Gaussian ellipsoids of the denoised 3 states HMM word extraction model. The power measurement, MFCC0, is plotted against MFCC1 and MFCC13.

they can extract the speech signal from the background environment to the same level of accuracy in a low noise environment (low noise here refers to normal office environment using close contact, good quality microphones). However, there are some important differences from the computation cost perspective. The differences are in both training and detection steps. The computation cost of each method can be estimated by counting the number of multiplication and division operations, which are the most heavily computational operations. The difficulty here is in implying the other factors that have a direct influence on the computations. The convergence rate of each method, which is normally controlled by heuristically-defined thresholds, and the number of processed speech frames, are examples of these difficulties. We found that the best way to evaluate the two methods is to calculate the CPU time required to fulfil the same task using the same computer. The tasks are the training procedures required to construct a model and the detection procedures required to accurately detect a certain utterance. The detection task is more important than the training task since training needs to be done once, and the important thing is to train an accurate model. Detection needs to be very fast as it is required in every speech recognition procedure in which the response time is the decisive factor in selecting the technique.

Table 1
Training task. Model training time required by each method given a dataset

Method	Dataset size (word)	Training time (sec)
HMMseg	50	451.14
DNHMM	50	100.67

It is important to mention here that the two techniques can process a single word or a complete long sentence and detect the silence and speech segments. This property facilitates the identification and recognition of complete sentences by the subsequent speech recogniser. The speech-silence discrimination in a long sentence does not imply any modification of the algorithms over that used in the isolated word situation, and they are processed as if they are the same.

To evaluate the computational performance of the two techniques we presented a long utterance to them and calculated the CPU time of the discrimination process required by each. The evaluation also includes the CPU time needed to process a single spoken word to determine the relationship between the lengths of each word to the processing time of the method.

The following two tables (Tables 1 and 2) show the model training time, given a dataset, and the processing time, given a certain signal for each method.

The time measured in all cases is that needed by the CPU to execute the algorithms written in a MATLAB

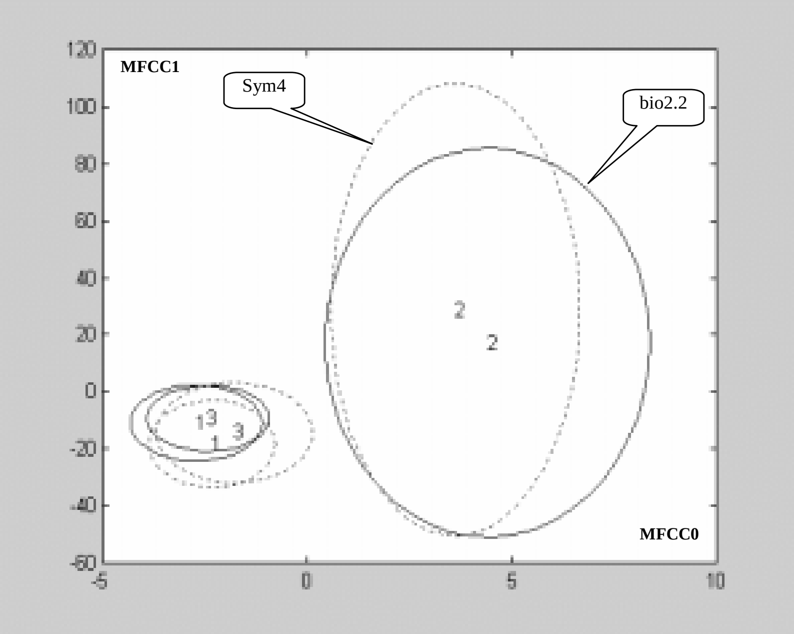


Fig. 12. Clustering properties of two word extraction models based on symlets, dotted, and biorthogonal, solid, wavelets. State's numbers are shown at the centre of the ellipsoids.

Table 2
Detection task. Processing time to detect the speech segments, given a certain signal for each method

		Detection Processing Time (sec)	
Input Signal	Signal length (sec)	HMMseg	DNHMM
Digits "5678"	4.87	5.067	4.657
Digit "6"	1.34	1.342	1.412

(MATLAB is a trademark of Math Works Inc.) environment with Version 5.3 running on a slow 400 MHz computer, 256 K RAM and NT operating system. This time can be reduced by a factor of 6–10 when the MATLAB programs are compiled into executable forms.

From the training task in Table 1, we can see that the training time required by HMMseg is longer than that required by DNHMM. However, this is not a factor to decide the adopted method, since training needs to be done once only and doesn't need to be done in real time. The detection task in Table 2 shows the time required to discriminate speech segments from the incoming signal. As we mentioned earlier, this is an important factor in determining the method to use since it has to be done in each speech recognition process. We can see from Table 2 that the two methods are nearly similar. DNHMM is faster than HMMseg in long sentences yet slower in isolated words processing. This is because of the nature of the interaction between the higher number of speech frames required in processing a signal using HMMseg compared to DNHMM which works in favour of DNHMM, and the denoising operation which increases the computational cost in DNHMM. For short signals (isolated words), the extra processing time due to the denoising operation is higher than that due to the excessive number of frames. On the other hand, in long signal (sentences), the balance is acting in vice versa.

HMMseg can be used efficiently in a low noise environment; high signal to noise ratio (SNR), while DNHMM is more robust in a noisy environment. HMMseg is not working properly in high noise environments (SNR $\ll$ 25 dB), because the high noise level could trigger the initiation of irrelevant states. Figure 13 shows the performance of DNHMM in detecting the speech segment of the word "6" spoken in a noisy environment simulated by a vacuum cleaner. The signal to noise ratio was 1.75 dB, which is a very difficult discrimination problem. This figure shows the detection of the speech segment and the spectrogram of the processed signal.

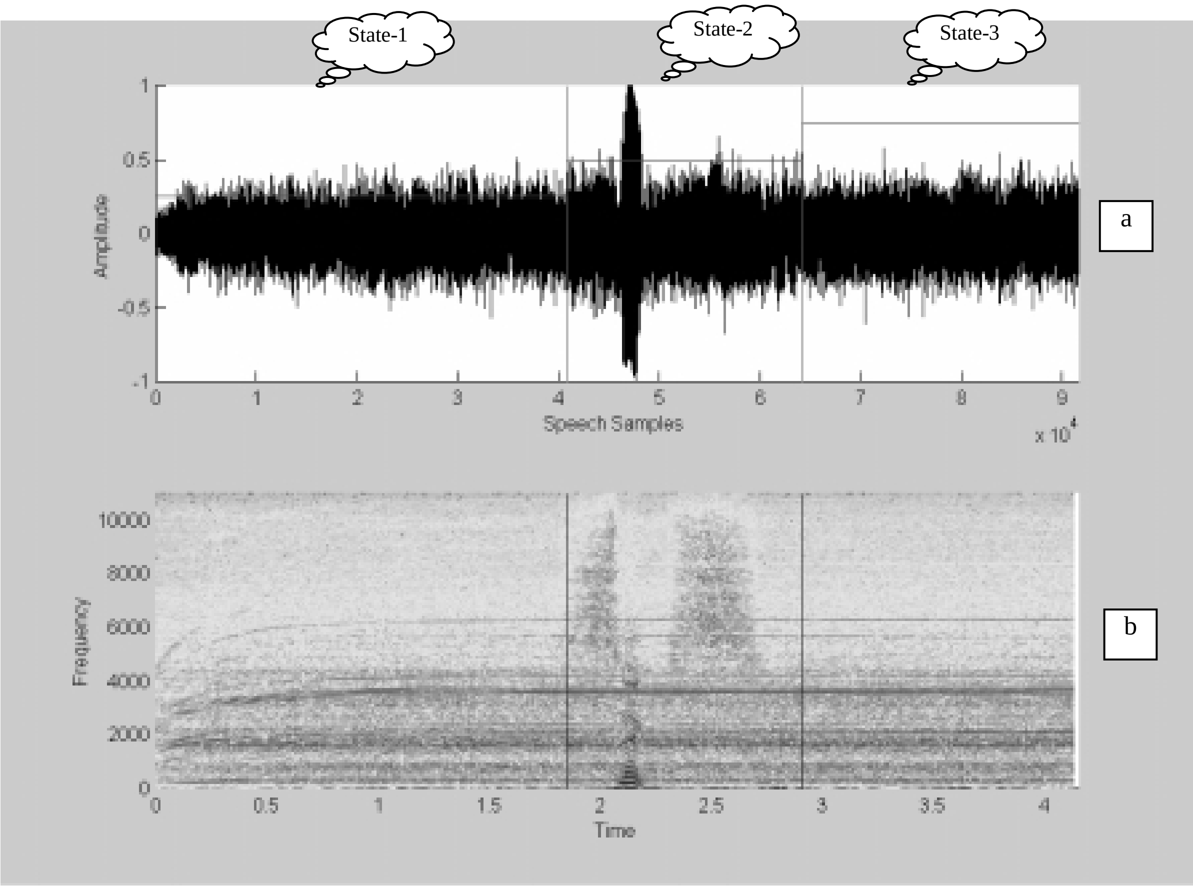


Fig. 13. Detection of the speech segment in a high noise environment (SNR = 1.75 dB). (a) Time signal with the state assignment. (b) Spectrogram with state assignment.

Figure 14 shows the enlargement of the clipped speech segment of Fig. 13 and the reconstruction of it after decomposing the signal to level 16 and then removing the low threshold coefficients, denoising, and then reconstructing again using the biorthogonal wavelets. The second lower graph clearly indicates the correctness of the detection process.

The final factor to evaluate is the discrimination of a long stream of a noise free speech signal using the two methods. The signal still includes some parasitic clicks and spikes. The example stream is composed of the connected digits “5678” as shown in Fig. 15. This figure shows a close comparison in the detection behaviour between the two methods. From this example we can see that the classification characteristics of HMMseg and DNHMM are mostly the same in low noise environments (SNR $>$ 25 dB).

5. Evaluation of speech extraction techniques based on speech recognition rate

In this section we study the effect of the speech segment extraction method on the overall performance of our HMM-based voice command system [1,4]. In our ASR system, two types of HMMs are needed. One is needed for speech segment extraction, which prevents the unwanted signal from being processed again and the other is needed for speech signal recognition. All the parameters of the word based speech recogniser are fixed while modifying the speech segment extraction technique. We built 26 HMM models to recognise all 26 English alphabets. Each model is of left-to-right topology with one state skip permission. Each model has 7 states and 3 multivariate Gaussian mixtures with a full covariance matrix.

The ISOLET dataset from OGI was used in all of the experiments reported in this paper [7,13]. This dataset comprises 5 sets (ISOLET-1 to ISOLET-5) collected

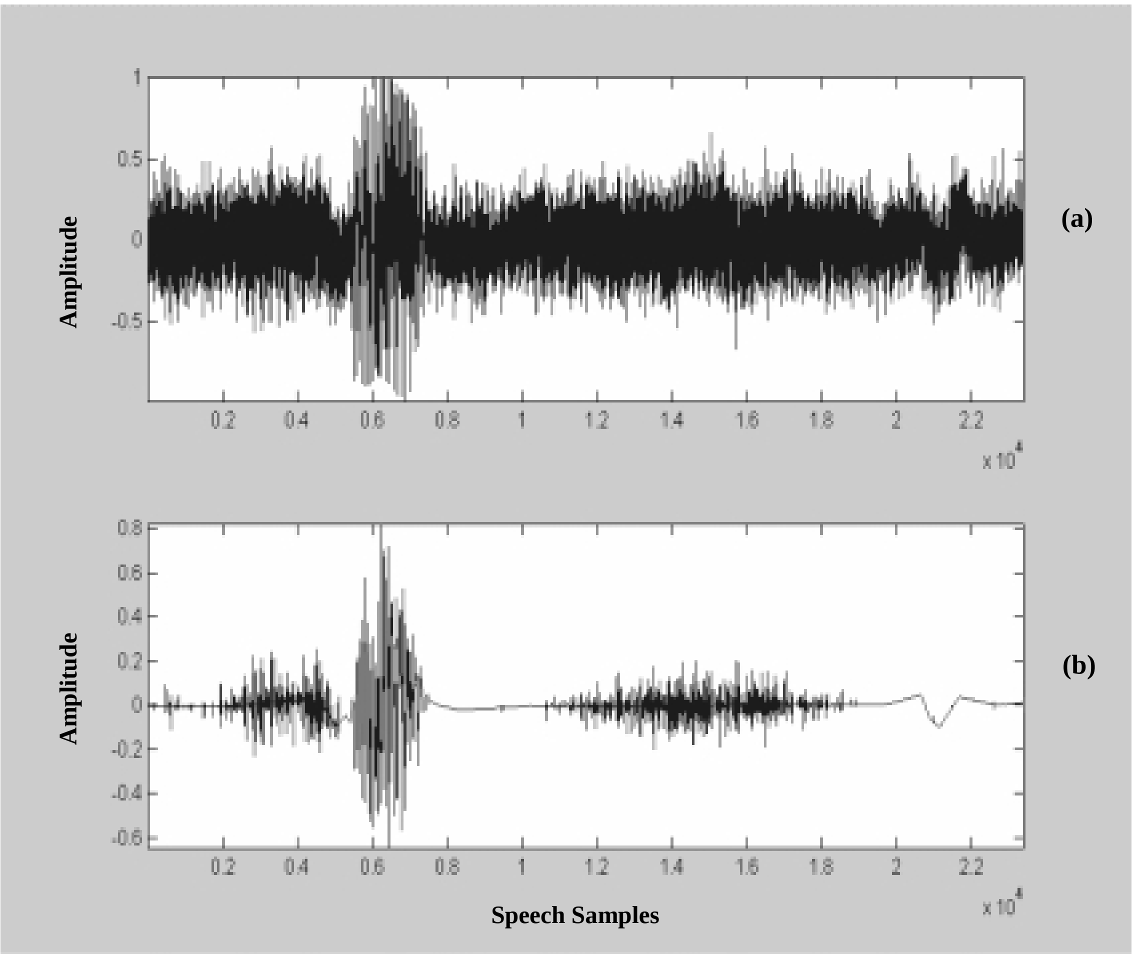


Fig. 14. The speech segment as detected by DNHMM. (a) The noisy detected speech segment, (b) The denoised detected speech segment.

from 150 speakers of English language alphabets. Each speaker spoke the name of each letter of the alphabet twice, which provides 52 training examples from each speaker. Thus, we have 6238 examples from the first four sets (used for training) and 1559 examples from the last set (used for testing) while three examples are missing from the dataset.

From the ISOLET dataset, we selected the highly confusable subset called E-set (b, c , d, e, g, p, t and v) for our experimentations. During recognition, the spoken letters were embedded in several environments of different Gaussian noise levels. Then, the contaminated signals were submitted to the different speech segments' detection techniques followed by a speech recognition stage. The recognition rate is then calculated, which in this case, a direct indication to the performance of the speech extraction technique used. The techniques used in speech segment extraction evaluation are manual extraction, which is the original ISOLET dataset, energy and zero-crossings (EZC), hidden Markov model (HMMseg), energy and zero-crossings with denoising (DNEZC), and hidden Markov model with denoising (DNHMM). Table 3 shows the recognition rates of the E-set letters in a clean environment (SNR> 30dB) using EZC and HMMseg methods as well as the recognition rate of the manually extracted alphabets. The recognition rate of the ASR system based on HMMseg outperforms that based on EZC. In the HMM based segmentation method, E and P alphabets contribute to 3.3% and 1.7% recognition error rates while the other alphabets are perfectly extracted. The last column of table 3 (labelled All) is the overall recognition rate of each technique and this information also indicates the outstanding performance of the HMM based method. More specifically, the EZC contributes to an overall recognition error rate of 10.9% (i.e. (97.5-86.9)/97.5) while the HMMseg contributes only to 0.6% of that error.

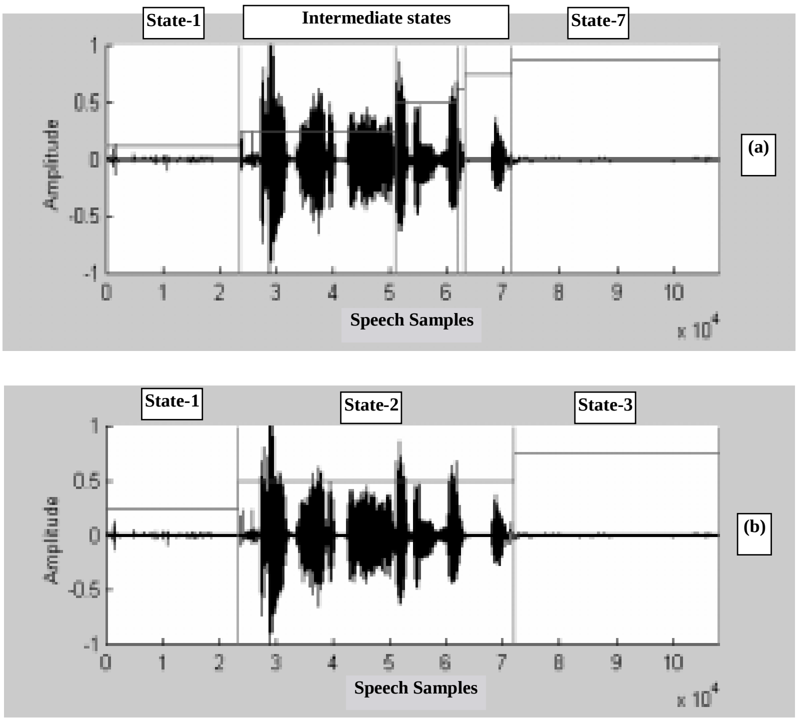


Fig. 15. The detection of connected digits "5678" using the four methods. (a) HMMseg detection function using 7 states HMM, (b) DNHMM detection function using 3 states denoised HMM model.

Table 3
Performance evaluation of EZC and HMMseg methods

	B	C	D	E	G	P	T	V	All
Manual	96.7	98.3	98.3	98.3	98.3	96.7	95	98.3	97.5
EZC	83.3	88.3	86.7	86.7	91.6	85	83.3	90	86.9
HMMseg	96.7	98.3	98.3	95	98.3	95	95	98.3	96.9

The other experiment was prepared in different noise levels environments using the DNEZC and the DNHMM techniques in speech segment extraction. In the case of DNEZC, the denoising process is used just before the EZC to enhance the performance of this technique. This is because EZC alone cannot work properly in high noise environments. Table 4 shows the recognition rate performance due to the two methods, favouring the DNHMM technique. We can figure out from Table 4 that the DNEZC technique contributes to an error rate of 10.9% in clean environment up to 24.2% in SNR = 5 dB. On the other hand, the DNHMM technique contributes to an error rate of only 0.6% up to 2.4% in the corresponding environments.

6. Conclusions

The goal of the speech segment extraction is to separate acoustic events of interest (speech segment to be recognised) in a continuously recorded signal from other parts of the signal (background). It is an essential front-end process in non-HMM based isolated words automatic speech recognition (ASR) systems. However, in HMM based paradigms this stage is excluded. A model representing the speech segment attached with two extra states, representing the pre and post non-speech periods, is used to process the complete acquired utterance. This might degrade the performance of the ASR and incur a heavy computation cost especially when we have a large number of models to be aligned with the incoming signal and the entire incoming utterance (speech plus silence) needs to be repetitively processed.

The recognition rate of many spoken command ASR systems is very much dependent on speech segment extraction accuracy. The key question here is how accurately speech must be detected so as to provide the best speech recognition performance. The definition of

Table 4
Performance evaluation of DNEZC and DNHMM methods

	Clean	15dB	10dB	5dB
Manual	97.5	93.8	88.5	80.2
DNEZC	86.9	77.1	70.8	60.8
DNHMM	96.9	92.7	87.1	78.3

'best' here is the pragmatic one – namely, the detection that provides highest recognition accuracy. To answer this question, a speaker independent digit recognition experiment was performed to determine the effect of speech detection error on recognition accuracy. Figure 1 shows the results of this experiment and we can see how the recognition rate degrades with the misalignment of speech segments. Thus our goal is to find new and sophisticated techniques to detect the speech segment as accurately as possible. The point to focus on here is to check the performance of each detection technique and how accurately it compares to the actual location of the speech segment within the entire recorded utterance (speech embedded in background).

This paper has introduced two novel methodologies (i.e. HMMseg and DNHMM) for segregating speech signal from its concurrent background. They exploit the techniques of hidden Markov modelling and wavelets denoising to boost the performance of our ASR system.

The unique feature of HMMseg is its use of the HMM technique to detect a speech segment embedded in its concurrent background. In this method we have proposed that the utterance is composed of a sequence of states (i.e. silence state – multiple speech states – silence state). This means that the relevant speech segment samples can be filtered by excluding the samples belonging to the first and the last states from the entire acquired utterance. Figure 5 shows one example of the speech segregation ability of this method.

The unique feature of DNHMM is its use of the wavelets denoising technique in addition to the HMM. This has three main advantages: first, it makes the technique robust to noise; second, it compresses the dynamics of the utterance to leave only three states in the sequence (i.e. silence state – one speech state – silence state); third, we can process fewer samples (slower frame rate and longer widow length) than for HMMseg since the states are more stable due to the effect of denoising. Figure 13 shows the performance of this method in a high noise environment.

Tables 1 and 2 depict the performance of each of the HMM based methods (i.e HMMseg and DNHMM) according to their computation costs. Additionally, Fig. 15 shows the long utterance detection ability of each one of them.

In another comparative study of classical techniques, Tables 3 and 4 specify the recognition error rates of the E set alphabets selected from the ISOLET dataset. We have secluded the error rates attributed to the speech segment extraction accuracy. From these tables, it can be concluded that the denoising process results in no improvement in the recognition rate in a clean environment for either the classical or HMM based techniques. The EZC and DNEZC techniques contribute to 10.9% of the overall error rate in a clean environment while the corresponding HMMseg and DNHMM techniques contributes to only 0.6%. In noisy environments, DNEZC contributes to an error rate of 17.8% in a SNR $=$ 15 dB environment and this increases up to 24.2 in SNR $=$ 5 dB. The corresponding DNHMM technique contributes to 1.2% to 2.4% respectively.

References

[1] W.H. Abdulla, Signal Processing and Acoustic Modelling of Speech Signal for Speech Recognition Systems, PhD Thesis, University of Otago, 2002.

[2] W.H. Abdulla and N.K. Kasabov, in: *Two pass hidden Markov model for speech recognition systems,* Singapore, 1999, pp. 175.

[3] W.H. Abdulla and N.K. Kasabov, in: *Speech recognition enhancement via robust CHMM speech background discrimination,* New Zealand, 1999, pp. 65–70.

[4] W.H. Abdulla and N.K. Kasabov, in: *Improving speech recognition performance through gender separation,* Dunedin, New Zealand, 2001, pp. 218–222.

[5] B.S. Atal and L.R. Rabiner, A pattern recognition approach to voiced-unvoiced-silence classification with application to speech recognition, *IEEE Trans.* **ASSP 24** (1976), 201–212.

[6] S.E. Bou-Ghazale and A.O. Asadi, in: *Hands-free voice activation of personal communication devices,* Istanbul, Turkey, 2000, pp. 1735–1738.

[7] R. Cole, Y. Muthusamy and M. Fanty, The ISOLET spoken letter database, Report No. Tech. Report 90-004, 1990.

[8] I. Daubechies, in CBMS-NSF Regional Conference Series in Applied Mathematics; Vol. 61 SIAM Press, Philadelphia, Pennsylvania, 1992.

[9] J.R. Deller, J.G. Proakis and J.H. Hansen, *Discrete-Time Processing of Speech Signals,* Macmillan Publishing, New York, 1993.

[10] D.L. Donoho, in: *Proceedings of the Symposia in Applied Mathematics,* I. Daubechies, ed., American Mathematical Society, 1993.

[11] D.L. Donoho, Denoising by soft-thresholding, *IEEE Trans.* **IT 41** (1995), 613–627.

[12] D.L. Donoho and I.M. Johnstone, Ideal spatial adaptation by wavelet shrinkage, *Biometrika* **81** (1994), 425–455.

[13] M. Fanty and R. Cole, *Morgan Kaufmann,* San Mateo, CA, 1991.

[14] S. Furui, Speaker independent isolated word recognition using dynamic features of speech recognition, *IEEE Trans.* **ASSP 34** (1986), 52–59.

[15] S. Furui, in: *Speaker independent isolated word recognition based on emphasized spectral dynamics,* Tokyo-Japan, 1986, pp. 1991–1994.

[16] A. Graps, An introduction to wavelets, *IEEE Computational Science and Engineering* **2** (1995).

[17] R. Haeb-Umbach, in: *Investigations on inter-speaker variability in the feature space,* Arizona-USA, 1999, pp. 1513.

[18] B.A. Hanson and T.H. Applebaum, in: *Robust speaker independent word recognition using static, dynamic, and acceleration features: experiments with lombard and noisy speech,* Albuquerque, NM, 1990, pp. 857–860.

[19] B.A. Hanson and T.H. Applebaum, in: *Features for noise-robust speaker-independent word recognition,* Kobe-Japan, 1990, pp. 1117–1120.

[20] L.-S. Huang and C.-H. Yang, in: *A novel approach to robust speech endpoint detection in car environment,* Istanbul, Turkey, 2000, pp. 1751–1754.

[21] L.F. Lamel, L.R. Rabiner, A.E. Rosenberg and J.G. Wilpon, An improved end points detector for isolated word recognition, *IEEE Trans.* **ASSP 29** (1981), 777–785.

[22] M. Misiti, Y. Misiti, G. Oppenheim and J.M. Poggi, Wavelet Toolbox, Math Works Inc., 1996.

[23] H. Ney, in: *Experiments on mixture-density phoneme modeling for the speaker of independent 1000-word speech recognition DARPA task,* Albuquerque, NM, 1990, pp. 713–716.

[24] L.R. Rabiner, A tutorial on hidden Markov models and selected applications in speech recognition, *Proc. IEEE* **77** (1989), 257–286.

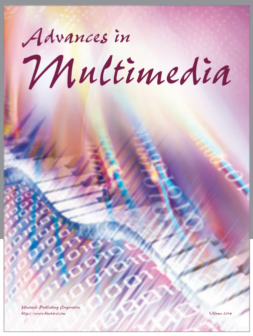
[25] L.R. Rabiner and M.R. Sambur, An algorithm for determining the endpoints of isolated utterances, *Bell Syst. Tech. J.* **54** (1975), 297–315.

[26] L.R. Rabiner, J.G. Wilpon and B.H. Juang, A model-based connected-digit recognition system using either hidden Markov models or templates, *Computer Speech & Language* **1** (1986), 167–197.

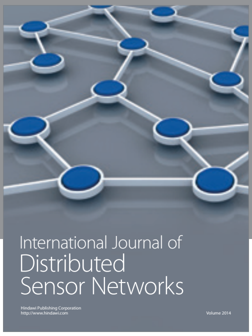
[27] H. Sakoe and S. Chiba, Dynamic programming algorithm optimization for spoken word recognition, *IEEE Trans.* **ASSP 26** (1978), 43–49.

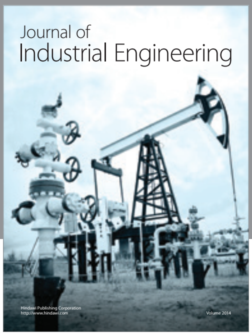
[28] W.-H. Shin, B.-S. Lee, Y.-K. Lee and J.-S. Lee, in: *Speech/non-speech classification usin multiple features for robust endpoint detection,* Istanbul, Turkey, 2000, pp. 1399–1402.

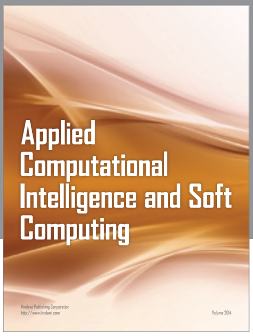
[29] J.G. Wilpon, L.R. Rabiner and T.B. Martin, An improved word -detection algorithm for telephone quality speech incorporating both syntactic and semantic constraints, *AT & T Tech. J.* **63** (1984), 479–498.

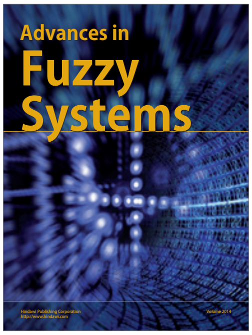

Advances in
Multimedia

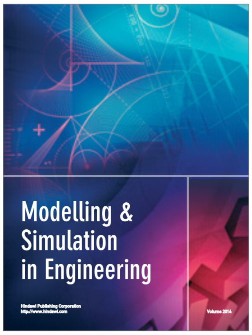
The Scientific
World Journal

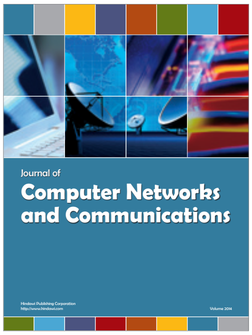

International Journal of
Distributed
Sensor Networks

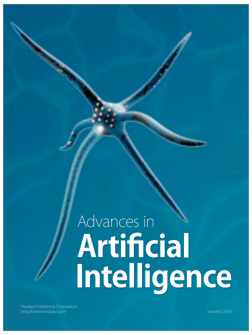

Journal of
Industrial Engineering

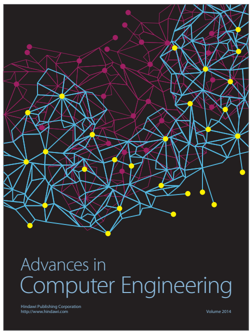

Applied
Computational
Intelligence and Soft
Computing


Advances in
Fuzzy
Systems

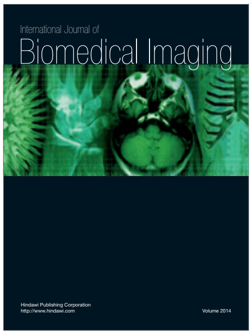

Modelling &
Simulation
in Engineering

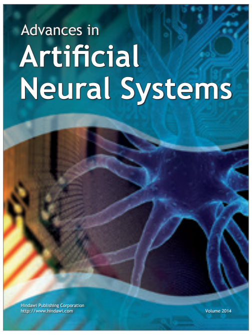

Journal of
Computer Networks
and Communications

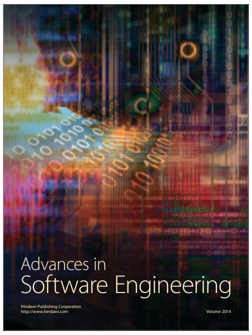

Advances in
Artificial
Intelligence


Advances in
Computer Engineering

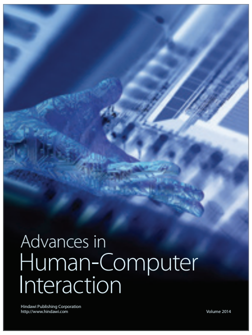
International Journal of
Computer Games
Technology

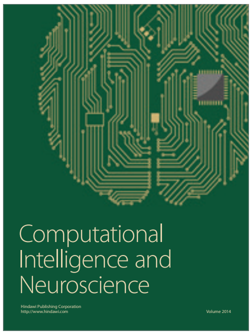

International Journal of
Biomedical Imaging


Advances in
Artificial
Neural Systems


Advances in
Software Engineering

Journal of
Robotics


Advances in
Human-Computer
Interaction


Computational
Intelligence and
Neuroscience

International Journal of
Reconfigurable
Computing

Journal of
Electrical and Computer
Engineering