Journal of Scientific & Industrial Research
Vol. 69, July 2010, pp. 515-522

Performance analysis of voice activity detection algorithm for robust speech recognition system under different noisy environment

C Ganesh Babu[1*], P T Vanathi[2], R Ramachandran[1], M Senthil Rajaa[1] and R Vengatesh[1]

[1]ECE, Bannari Amman Institute of Technology, Sathyamangalam, 638401, India

[2]ECE, PSG College of Technology, Coimbatore, 641008, India

Received 21 January 2010; revised 11 April 2010; accepted 13 April 2010

This study evaluates performance of objective measures in terms of predicting quality of noisy input speech signal using voice activity detection (VAD). Implementation process includes a speech-to-text system using isolated word recognition with a vocabulary of 10 words (digits 0-9) and statistical modeling (Hidden Markov Model - HMM) for machine speech recognition. In training period, uttered digits were recorded using 8-bit pulse code modulation (PCM) with a sampling rate of 8 KHz and save as a wave format file using sound recorder software. HMM performs speech analysis using linear predictive coding (LPC) method of degree. For a given word in vocabulary, system builds an HMM model and trains model during training phase. Training steps from VAD to HMM model building are performed using PC-based Matlab programs. Current framework uses automatic speech recognition (ASR) with HMM based classification and noise language modeling to achieve effective noise knowledge estimation.

Keywords: Hidden Markov model (HMM), Subband OSF based voice activity detection (VAD), Vector quantization

Introduction

In speech recognition system (SRS), highly affected systems are new wireless communication voice services and mobile technology. Most of the noise compensation algorithm often require voice activity detector (VAD) to estimate presence or absence of speech signal[1]. In this paper, quality of speech recognition has been enhanced by changing parameters such as order of order statistic filter (OSF) and smoothing constant. By enhancing parameters, VAD efficiency increases in SRS.

Experimental Section

Speech Characteristics

Speech signals are composed of sequence of sounds. Sounds can be classified into following three distinct classes according to mode of excitation[2,3]: i) Voiced sounds are produced by forcing air through glottis with the tension of vocal cords adjusted so that they vibrate in a relaxation oscillation, thereby producing a quasi-periodic pulse of air which vibrates the vocal tract; ii) Fricative or unvoiced sounds a regenerated by forming a constriction at some point in vocal tract and forcing air through constriction at a high enough velocity to produce turbulence; and iii) Plosive sounds result from making a complete closure and abruptly releasing it.

Voice Activity Detection (VAD)

Subband based VAD (Fig. 1) uses two order statistics filters for multi-band quintiles (MBQ) signal-to-noise ratio (SNR) estimation[4]. Implementation of both OSF is based on a sequence of $2N+1$ log-energy values $\{E(m-N,k),\ldots,E(m,k),\ldots,E(m+N,k)\}$ around the frame to be analyzed. This algorithm operates on subband log-energies. Noise reduction is performed first and VAD decision is formulated on de-noised signal. Noisy speech signal is decomposed into 25-ms frames with a 10-ms window shift. Let $X(m,l)$ be spectrum magnitude for m^{th} band at frame 1. Design of noise reduction block is based on wiener filter (WF) theory, whereby attenuation is a function of SNR of input signal. VAD decision is formulated in terms of de-noised signal, being subband log-energies processed by means of OSFs.

Drawbacks of Existing VAD

Existing algorithm assumes that noise spectrum does not significantly vary within N frame of neighbourhood of 1st frame; this is not true for highly stationary noise. Noise estimation of first frame is used to denoise 8 frames forward. Noise estimate is very low for first frame. So algorithm fails at the beginning to evaluate noise spectrum and detection afterwards could be totally erroneous.

*Author for correspondence
E-mail: bits_babu@yahoo.co.in

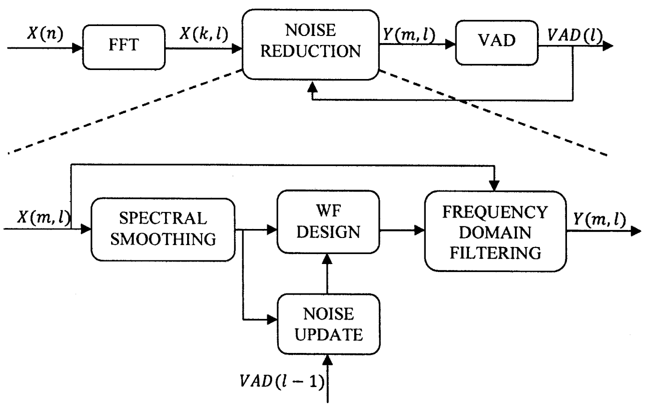

Fig. 1—Subband OSF based VAD

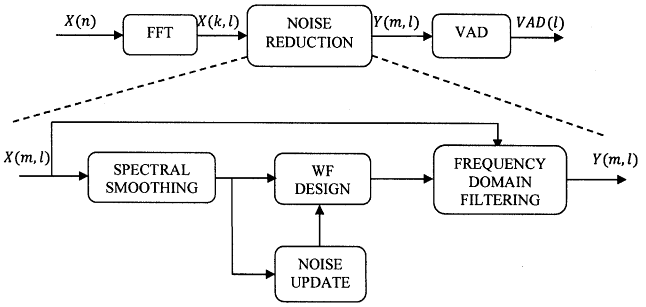

Fig. 2—Proposed VAD

Proposed VAD

Proposed algorithm does not depend on feedback loop for noise spectrum estimation. Instead, it uses a noise estimation algorithm, which updates noise for every frame. This method is best suited for highly non stationary environments, thus increasing robustness[5]. An improved VAD (Fig. 2) algorithm employs long-term signal processing and maximum spectral component tracking. It improves speech/non-speech discriminability and speech recognition performance in noisy environments[6]. In noise estimation algorithm, smoothed power spectrum of noisy speech signal is estimated using a first-order recursive formula as

$$P(\lambda,k) = \eta\, P(\lambda,k) + (1-\eta)Y(\lambda,k) \qquad \ldots(1)$$

where $Y(\lambda,k)$ is short time power spectrum of noisy speech and η is smoothing constant, wher λ is frame index and k is frequency bin index.

Speech frame is classified as speech present or speech absent. Incoming frame is classified as speech absent frame if the following condition is satisfied:

$$\zeta_L(\lambda) < \sigma \ and\ \xi_M(\lambda) < \sigma \ and\ \xi_H(\lambda) < \sigma \qquad \ldots(2)$$

where $N(\lambda,k)$ is estimate of noise power spectrum at frame λ , and LF, MF, Fs correspond to frequency bins of 1 kHz, 3 kHz and sampling frequency respectively,

where $\zeta_L(\lambda) = \dfrac{\sum_{k=1}^{LF} P(\lambda,K)}{\sum_{k=1}^{LF} N(\lambda-1,k)}$

$$\zeta_M(\lambda) = \frac{\sum_{k=LF+1}^{MF} P(\lambda,K)}{\sum_{k=LF+1}^{MF} N(\lambda-1,k)}$$

$$\zeta_H(\lambda) = \frac{\sum_{k=MF+1}^{\frac{F_s}{2}} P(\lambda,K)}{\sum_{k=MF+1}^{\frac{F_s}{2}} N(\lambda-1,k)}$$

If frame is classified as speech absent frame, noise spectrum is updated as

$$N(\lambda,k) = \varepsilon N(\lambda-1,k) + (1-\varepsilon)|Y(\lambda,k)|^2 \qquad \ldots(3)$$

In speech present frames, to update noise, frequency bins are classified as speech present or absent by tracking local minimum of noisy speech and then speech presence in each frequency bin is decided separately using ratio of noisy speech power to its local minimum. A different non-linear rule is used for tracking minimum of noisy speech by averaging past spectral values.

If, $P_{min}\,(\lambda-1,k) < P(\lambda,K)$

Then, $P_{min}\,(\lambda,k) = \gamma P_{min}\,(\lambda-1,k)$

$$+\frac{1-\gamma}{1-\beta}\,(P(\lambda,k) - \beta\, P(\lambda-1,k)) \qquad \ldots(4)$$

Else, $P_{min}\,(\lambda,k) = P(\lambda,K)$ ...(5)

where $P_{\min}\,(\lambda,k)$ is local minimum of noisy speech power spectrum and β and γ are constants, whose values are determined experimentally.

Let $S_r(\lambda,k) = P(\lambda,k)/\,P_{\min}\,(\lambda,k)$ denote ratio between energy of noisy speech to its local minimum. This ratio is compared against a frequency-dependent threshold and if it is found to be larger than threshold, then corresponding frequency is considered to contain speech. Using $S_r(\lambda, k)$, new frequency-dependent smoothing constant can be estimated as

$$\alpha_s(\lambda,k) = \begin{cases} \alpha_1 & if\ s(\lambda,k) < \delta(k) \\ \alpha_2 & otherwise \end{cases} \quad \ldots(6)$$

where $\alpha_{1,}\alpha_2$ are smoothing constants ($\alpha_{1,} > \alpha_2$) and $\delta\ (k)$ is frequency-dependent threshold given as

$$\delta(k) = \begin{cases} 3 & 1 \le k \ \le LF \\ 3 & LF < k \ \le MF \\ 5 & MF < k \le F_s/2 \end{cases} \quad \ldots(7)$$

Finally, after computing frequency-depending smoothing factor $\alpha_s\ (\lambda, k)$, noise spectrum estimate is updated as

$$N(\ \lambda,k) = \ \alpha_s(\lambda,k)N(\lambda-1,k) + \ (1-\alpha_s(\lambda,k))\ |Y(\lambda,k|^2 \quad \ldots(8)$$

Hidden Markov Model (HMM)

Technique used to implement speech recognition is HMM[7], which represents utterance of word and calculates probability the model which created sequence of vectors[8]. There are some fundamental problems in designing of HMM for analysis of speech signal. HMM is represented as

$$\lambda\ =\ (\ \pi, A, B\ ) \quad \ldots(9)$$

where, π, initial state distribution vector; A , state transition probability matrix; and B , continuous observation probability density function matrix.

Given appropriate values of A, B and π , HMM can be used to give an observation sequence as

$$O = O_1\ O_2\ \ldots\ldots.O_T \quad \ldots(10)$$

where each observation O_t is one of the symbols from observation symbol V and T is number of observation in sequence as follows: i) Choose an initial state $q_1 = S_i$ according to initial state distribution π; ii) Set t =1; iii) Choose $O_t = v_k$ according to the symbol probability distribution in state S_i ; iv) Transit to a new state $q_{t+1} = S_j$ according to state transition probability distribution for state S_i ; and v) Set $t = t + 1$ (return to step iii) if t<T; otherwise terminate procedure.

Above procedure can be used as a generator of observations, and as a model for how a given observation sequence was generated by HMM. After re-estimate of parameters, model is given as

$$\lambda\ =\ (\ A, \mu, \Sigma).\quad \ldots(11)$$

Model is saved to represent that specific observation sequences, i.e. an isolated word. Basic theoretical strength of HMM is that it combines modeling of stationary stochastic processes (for short-time spectra) and temporal relationship among processes (via a Markov chain) together in a well-defined probability space. This combination allows to study two separate aspects of modeling a dynamic process (like speech) using one consistent framework. Another attractive feature of HMM's is relatively easy and straightforward to train a model from a given set of labeled training data (one or more sequences of observations).

Linear Predictive Coding (LPC) Analysis

To obtain observation vectors O from speech samples, s is to perform a front end spectral analysis. Type of spectral analysis that is often used (and one described here) is called LPC[2,8-11] (Fig. 3). Steps in processing are as follows:

i. Preemphasis

Digitized speech signal is processed by a first-order digital network in order to spectrally flatten signal, which is discussed as

$$\hat{s}(n) = s(n) - \alpha s(n-1) \quad \ldots(12)$$

ii. Blocking into Frames

Sections of N_A consecutive speech samples are used as a single frame. Consecutive frames are spaced M_A samples apart. Frame separation is given as

$$X_l(n) = \hat{s}(m_l + n), \qquad o \le n \le N-1;\ 0 \le l \le L-1 \quad \ldots(13)$$

iii. Frame Windowing

Each frame is multiplied by N_A sample window (Hamming Window) w(n) to minimize adverse effects of chopping N_A samples section out of running speech signal. Technique is given as

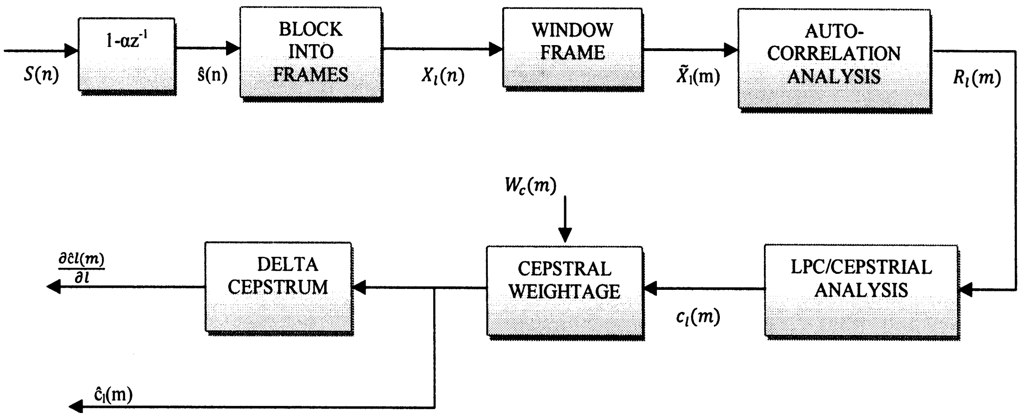

Fig. 3—Linear predictive coding

$$\tilde{X}_l(n) = x_l(n).w(n), \qquad 0 \le n \le N-1 \qquad \ldots(14)$$

iv. Auto Correlation Analysis

Each windowed set of speech sample is autocorrelated to give a set of (p + 1) coefficient, where p is order of desired LPC analysis. Autocorrelation process is given by

$$R_l(m) = \sum_{n=0}^{N-m} \tilde{X}_l(n)\tilde{X}_l(n+m), 0 \le m \le p \qquad \ldots(15)$$

v. LPC/Cepstral Analysis

A vector of LPC coefficients is computed from autocorrelation vector using a Levinson or a Durbin recursion method. An LPC derived cepstral vector is then computed up to Qth component as

$c_l(m)$=CEPSTRAL COEFFICIENT, 1 = m = Q ...(16)

vi. Cepstral Weighting

Q-coefficient cepstral vector $c_l(m)$ at time frame l is weighted by a window, $W_c(m)$[8,9]

$$W_c(m) = 1 + \left[\left(\frac{Q}{2}\right)\left(sin\left(\frac{\pi m}{Q}\right)\right)\right], \qquad 1 \le m \le Q \qquad \ldots(17)$$

$$\hat{c}_l(m) = c_l(m).w_c(m), 1 \le m \le Q \qquad \ldots(18)$$

To give

$$\hat{c}_l(m) = c_l(m).W_c(m) \qquad \ldots(19)$$

vii. Delta Cepstrum

Time derivative of sequence of weighted cepstral vectors is approximated by a first-order orthogonal polynomial over a finite length window of frames centered around current vector[10,11].

$$\Delta\hat{c}_l(m) = \left[\sum_{k=-K}^{K} k\ \hat{c}_l - k(m)\right].G \qquad \ldots(20)$$

where G is gain term to make variance of $\hat{c}_l(m)$ and $\Delta\hat{c}_l(m)$ equal, $Q_l(m) = \{\hat{c}_l(m), \Delta\hat{c}_l(m)\}$, and $\Delta\hat{c}_l(m) = \frac{\partial \hat{c}_l(m)}{\partial l}, \quad 1 \le m \le Q$.

Vector Quantization, Training and Recognition

To use HMM with discrete observation symbol density, a vector quantizer (VQ) is required to map each continuous observation vector into a discrete code book index. Major issue in VQ is design of an appropriate codebook for quantization. Procedure basically partitions training vector into *M* disjoin sets. Distortion steadily decreases as *M* increases. Hence, HMM (codebook size, M=32 to 256 vectors) has been used in speech recognition experiments using HMMs[12,13]. During training phase, system trains HMM for each digit in vocabulary[14]. Same weighted cepstrum matrices for various samples and digits are compared with code book

Table—1 Performance analysis of Subband OSF based VAD with and without noise estimation for 0dB for different noises

Noises	Airport			Babble			Car			Exhibition			Restaurant			Station			Street			Train		
Numbers	W	WO	%I	W	WO	%I	W	WO	%I	W	WO	%I	W	WO	%I	W	WO	%I	W	WO	%I	W	WO	%I
Zero	28	23	17.85	15	1	93.33	4	4	0.00	1	1	0.00	42	38	9.52	71	54	23.94	100	94	6.00	27	1	96.29
One	55	34	38.18	65	30	53.84	45	42	6.67	50	47	6.00	68	61	10.29	43	22	48.83	35	30	14.28	48	33	31.25
Two	81	62	23.45	31	25	19.35	68	57	16.17	70	55	21.42	72	40	44.44	86	37	56.97	23	15	34.78	71	44	38.02
Three	37	27	27.03	29	28	3.45	55	49	10.91	72	40	44.44	63	57	9.52	33	33	0.00	24	1	95.83	50	42	16.00
Four	98	27	72.45	66	28	57.58	97	53	45.36	98	31	68.37	99	38	61.62	94	33	64.89	66	40	39.39	80	35	56.25
Five	64	54	15.63	68	53	22.06	59	38	35.59	97	36	62.89	100	40	60.00	50	37	26.00	98	14	85.71	69	53	23.19
Six	1	1	0.00	1	1	0.00	81	27	66.67	66	40	39.39	82	36	56.10	1	1	0.00	1	1	0.00	24	13	45.83
Seven	69	37	46.38	30	12	60.00	81	37	54.32	58	54	6.90	60	27	55.00	62	35	43.55	56	48	14.29	92	14	84.78
Eight	58	48	17.24	39	28	28.21	46	40	13.04	65	60	7.69	53	45	15.09	75	59	21.33	62	49	20.97	58	18	68.97
Nine	25	17	32.00	11	4	63.34	19	12	36.84	8	7	12.50	19	14	26.32	22	13	40.91	20	13	35.00	21	4	80.95
Avg %	52.5	36.88	14.62	37.38	23.00	33.28	61.25	38.37	29.46	64	38	31.17	73.25	42.12	38.31	55	31.5	33.02	50.37	30.37	36.28	57.63	29.37	48.95

W, VAD with noise estimation (proposed method); WO, VAD without noise estimation (Ramirez *et al* method); %I, percentage increment from Ramirez method to proposed method; Avg%, average percentage

Table.2—Performance analysis of Subband OSF based VAD with and without noise estimation for 5dB for different noises

Noises	Airport			Babble			Car			Exhibition			Restaurant			Station			Street			Train		
Numbers	W	WO	%I	W	WO	%I	W	WO	%I	W	WO	%I	W	WO	%I	W	WO	%I	W	WO	%I	W	WO	%I
Zero	15	11	26.66	53	30	43.39	7	6	14.28	49	10	79.59	56	50	10.71	29	27	6.89	89	86	3.37	91	0	100
One	67	52	22.38	75	57	24.00	61	50	18.03	54	54	0.00	60	42	30.00	60	58	3.33	77	36	53.24	65	64	1.54
Two	82	18	78.04	1	1	0	68	58	14.70	72	37	48.61	87	16	81.60	75	36	52.00	78	44	43.58	67	37	44.77
Three	88	36	59.09	50	42	16.00	42	34	19.04	62	38	38.70	49	43	12.24	59	40	32.2	44	43	2.27	61	46	24.59
Four	97	40	58.76	58	42	27.59	98	57	41.84	96	38	60.42	98	43	56.12	98	40	59.18	95	43	54.74	86	46	46.51
Five	5	5	0.00	95	44	53.68	55	49	10.91	65	64	1.5	59	43	27.12	82	77	6.09	75	69	8.00	96	75	21.87
Six	1	1	0.00	2	1	50.00	40	38	5.00	29	15	48.28	21	21	0.00	1	1	0.00	1	1	0.00	10	9	10
Seven	76	45	40.79	36	24	33.33	78	23	70.51	35	19	45.71	81	28	65.43	59	38	35.59	81	27	66.67	86	26	69.77
Eight	62	53	14.52	65	58	10.77	64	54	15.63	61	22	63.93	71	59	16.90	61	48	21.31	70	44	21.14	71	54	23.94
Nine	23	16	30.43	19	11	42.11	18	9	50.00	27	12	55.56	26	16	38.46	24	1	95.83	29	6	79.31	28	23	17.86
Avg %	53.86	26.00	35.72	46.25	30.12	30.99	56.12	39.37	24.28	57.75	34.36	40.35	63.87	35.75	35.40	57.87	39.62	24.41	67.5	43.63	28.98	70.25	37.88	39.88

W, VAD with noise estimation (proposed method); WO, VAD without noise estimation (Ramirez *et al* method); %I, percentage increment from Ramirez method to proposed method; Avg%, average percentage

Table.3—Performance analysis of Subband OSF based VAD with and without noise estimation for 10dB for different noises

Noises	Airport			Babble			Car			Exhibition			Restaurant			Station			Street			Train		
Numbers	W	WO	%I	W	WO	%I	W	WO	%I	W	WO	%I	W	WO	%I	W	WO	%I	W	WO	%I	W	WO	%I
Zero	42	18	57.14	50	6	88.00	33	1	96.96	91	89	2.19	25	2	92.00	38	3	92.10	79	75	5.06	13	0	100
One	70	49	30.00	77	21	72.72	76	58	23.68	56	44	21.42	67	57	14.92	65	44	32.30	68	52	23.52	58	32	39.65
Two	75	61	18.66	69	55	20.28	83	76	8.43	69	48	30.43	79	44	44.30	87	23	73.56	83	58	30.12	85	33	61.17
Three	67	49	26.86	69	53	23.19	45	40	11.11	87	84	3.44	43	36	16.28	42	38	9.52	59	49	16.94	58	47	18.96
Four	95	49	48.42	68	53	22.06	100	40	60.00	97	84	13.40	98	36	62.27	98	38	61.22	99	49	50.51	91	47	48.35
Five	81	77	4.9	61	57	6.55	81	71	12.34	89	84	5.61	65	56	13.84	94	80	14.89	67	56	16.41	91	83	8.79
Six	1	1	0	3	1	66.77	24	24	0	29	14	51.72	1	0	100	1	1	0.00	1	1	0.00	1	1	0.00
Seven	80	48	40.00	38	16	57.89	82	40	51.22	55	41	25.45	98	18	81.63	85	30	64.71	82	36	56.10	86	51	40.70
Eight	63	57	9.52	54	52	3.70	75	59	21.33	67	56	16.42	67	58	13.43	70	30	57.14	75	53	29.33	69	58	15.94
Nine	23	15	34.78	24	13	45.83	23	15	34.78	25	18	28.00	30	18	40.00	31	10	67.74	31	6	80.65	40	12	70.00
Avg %	63.87	44.13	28.24	54.37	32.75	44.68	65.50	43.75	32.96	71.62	51.00	19.20	59.50	32.88	53.15	63.75	32.13	45.54	67.25	47.00	24.83	60.38	36.75	39.70

W, VAD with noise estimation (proposed method); WO, VAD without noise estimation (Ramirez *et al* method); %I, percentage increment from Ramirez method to proposed method; Avg%, average percentage

Table4—Performance analysis of Subband OSF based VAD with and without noise estimation for 15dB for different noises

Noises	Airport			Babble			Car			Exhibition			Restaurant			Station			Street			Train		
Numbers	W	WO	%I	W	WO	%I	W	WO	%I	W	WO	%I	W	WO	%I	W	WO	%I	W	WO	%I	W	WO	%I
Zero	82	75	8.53	40	28	30.00	69	12	82.61	84	90	77.38	80	9	88.75	57	31	45.61	100	95	5.00	95	0	100
One	71	41	42.25	79	41	48.10	77	60	22.07	77	75	2.59	76	53	30.26	67	43	35.82	78	45	42.30	64	54	15.62
Two	92	38	58.69	90	90	0.00	86	53	36.90	87	54	37.93	88	36	59.09	79	40	49.36	87	33	62.06	89	56	37.07
Three	55	44	20.00	43	36	16.28	54	41	24.07	47	14	70.21	50	39	22.00	39	36	7.69	49	44	10.20	59	50	15.25
Four	94	44	53.19	73	36	50.68	100	41	59.00	98	14	85.71	99	39	60.61	97	36	62.89	95	44	53.68	91	50	45.05
Five	69	59	14.49	89	78	12.35	81	67	17.28	82	66	19.51	72	64	11.11	83	63	24.09	92	82	10.86	81	74	7.50
Six	1	1	0.00	5	1	80.00	8	7	12.50	7	7	0.00	9	7	22.22	1	1	0.00	4	3	25.00	39	19	51.28
Seven	82	37	54.88	46	23	50.00	85	41	51.76	43	30	30.23	82	43	47.56	85	24	71.76	88	0	100	85	37	56.47
Eight	76	44	42.11	45	44	2.22	72	51	29.17	73	19	73.97	72	59	18.06	71	51	28.17	75	45	40.00	75	61	18.67
Nine	34	18	47.06	15	15	0.00	30	19	36.67	33	10	69.70	38	16	57.89	31	11	64.52	46	15	67.39	37	21	43.24
Avg %	68.25	42.38	31.50	58.12	41.63	35.92	70.00	40.25	38.27	65.62	43.75	40.44	69.50	36.25	42.70	63.50	34.25	37.15	74.12	43.25	38.63	75.37	42.50	41.03

W, VAD with noise estimation (proposed method); WO, VAD without noise estimation (Ramirez *et al* method); %I, percentage increment from Ramirez method to proposed method; Avg%, average percentage

Table 5—Overall performance analysis of Subband OSF based VAD with and without noise estimation

Improvement	0dB	5dB	10dB	15dB
Better	Train (48.95%)	Exhibition (40.35%)	Restaurant (53.15%)	Restaurant (42.70%)
Least	Airport (14.62%)	Car (24.28%)	Exhibition (19.20%)	Airport (31.50%)

and their corresponding nearest codebook vector indices is sent to Baum-Welch algorithm to train a model for input index sequence. After training, three models for each digit correspond to three samples in vocabulary set. Then average of A, B and π matrices are found over samples to generalize models. Input speech sample is preprocessed to extract feature vector[15]. Then, nearest codebook vector index for each frame is sent to digit models. System chooses model that has maximum probability of a match.

Results and Discussion

Several experiments were conducted commonly to evaluate VAD algorithm. Analysis mainly focused on error probabilities. Proposed VAD was evaluated in terms of ability to discriminate speech from non speech at different SNR's values. VADs avoid losing speech periods leading to an extremely conservative behavior in detecting speech pauses. Sub band OSF VAD identifies presence or absence of speech and extracts speech from noise corrupted speech. Proposed framework uses a speech processing module including a noise estimation algorithm with HMM based classification and noise language modeling to achieve effective noise knowledge estimation. Noise, taken from AURORA database, includes suburban train noise, babble, car, exhibition hall, restaurant, street, airport and train-station noise. In training phase, uttered words (100 samples each digits 0-9, both male and female) were recorded using 8-bit pulse code modulation (PCM) with a sampling rate of 8 KHz and saved as a wave file using sound recorder software. Automatic speech recognition systems work reasonably well under clean conditions but become fragile in practical applications involving real-world environments.

Analysis was taken at different environmental noises for digits 0-9 at different SNR values. From experimental results (Tables 1-4), proposed VAD works better than existing VAD algorithm for different noises. Accuracy increases when noise is estimated for each frame than without noise is being estimated. Looking into overall performance of proposed VAD algorithm with noise estimation (Table 5), better recognitions occur for restaurant noise at 10dB with accuracy of 53.15% and least performance for recognition accuracy about 14.62% for 0dB airport noise. With inclusion of noise estimation, proposed VAD system works better for different noises at different SNR values.

Conclusions

Proposed VAD works better than existing VAD algorithm for different noises at different SNR values. recognition for airport noise.

References

1 Ramirez, S J C, Benitez C, de la T A & Rubio A, Voice activity detection with noise reduction and long-term spectra divergence estimation, *IEEE Int Conf Acoustics, Speech Signal Processing,* **2** (2004) 1093-1096.

2 Makhoul J, Roucos S & Gish H, Vector Quantization in Speech Coding, *Proc IEEE,* **73** (1985) 1551-1558.

3 Becchetti C & Ricotti L P, *Speech Recognition Theory and C++ Implementation* (John Wiley & Sons Publication, Singapore) 2004, 121-141.

4 Ram?ez J, Segura J C, Ben?ez C, Torre ?e la & Rubio A, An effective subband OSF- based VAD with noise reduction for robust speech recognition, *IEEE Trans Speech Audio Proc,* **13** (2005) 1119-1129.

5 Alan D, Nordholm S & Togneri R, Statistical voice activity detection using low-variance spectrum estimation and an adaptive threshold, *IEEE Trans Audio, Speech Language Proc,* **14** (2006) 412-423.

6 Rangachari S & Loizou P C, A noise-estimation algorithm for highly non-stationary environments, *Speech Commun*, **48** (2005) 220-231.

7 Rabiner L R, A tutorial on Hidden Markov Model and selected applications in speech recognition, *Proc IEEE,* **77** (1989) 172-175.

8 Makhoul J, Linear prediction a tutorial view, *Proc IEEE*, **63** (1975) 215-230.

9 Markel J D & Gray Jr A H, *Linear Prediction of Speech* (Springer-Verilag, Newyork NY) 1976, 71-75.

10 Tokhura Y, A weighted cepstral distance measure for speech recognition, *IEEE Trans Acoust Speech Signal Processing*, **35** (1987) 1414-1422.

11 Juang B H, Rabiner L R & Wilpon J G, On the use of bandpass filtering in speech recognition, *IEEE Trans Acoust Speech Signal Processing*, **35** (1987) 947-954.

12 Rabiner L R, Levinson S E & Sondhi M M, On the application of vector quantization and hidden markov models to speaker-independent isolated word recognition, *Bell Syst Tech J,* **62** (1983) 1075-1105.

13 Balamuragan M T & Balaji M, SOPC- based speech to text conversion, in *NIOS-II, Embedded Processors Design Contest-Outstanding*, (Altera International Ltd., Hong Kong) 2006, 83-108.

14 Ephraim Y & Merhav N, Hidden Markov Processes, *IEEE Trans Inform. Theory,* **48** (2002) 1518-1569.

15 Rabiner L R & Schafer R W, *Digital Processing of Speech Signals* (Pearson Education Publication, New Delhi) 2004, 399-455.