A ROBUST OBSERVATION MODEL FOR AUTOMATIC SPEECH RECOGNITION WITH ADAPTIVE THRESHOLDING

Sharmida A.H.[1], **Ms. Athira A.P.**[2],

[1]M.tech Student, Applied Electronics and Instrumentation, Lourdes Matha College of Science & Technology
[2]Asst.Professor, Dept. of ECE, Lourdes Matha College of Science & Technology

---***---

Abstract — *Reverberation and noise will decrease the quality of speech signal and increase the word error rate of automatic speech recognition (ASR). Mel frequency ceptral coefficient plays an important role in the ASR. This paper review an observation model, the model used to study the effect of reverberation on speech signal. From those observations a relation between clean speech signal and reverberant speech signal is obtained. The observation model's advantage is that additive noise are explicitly taken in to account and a compensation constant added based on the observation error. For detecting voice active region single pole filtering method is used. Amplitude envelope of the signal is obtained at each single frequency. Spectral variance is higher for speech signal than the noises. Mean and variance of noise compensated weighted component envelopes are taken across frequencies at each time instant. Decision logic based on the features derived from mean and variance values*

Index Terms — *Observation model, Spectral variance, weighted component envelope, word error rate.*

I. INTRODUCTION

Hands free systems allowing the user to move freely without wearing headset or holding a microphone they contribute increased convenience and safety. Reverberation is the multipath prorogation of sound from source to receiver. Both reverberation and noise can cause degradation of acoustic signals. ETSI Advanced Frontend [1] are effective against additive noise are ineffective against reverberation. An observation model is used to formulate the effect of reverberation based on the findings of observation model, feature compensation is added.

A System theoretical model for reverberation can be modeled as the convolution of speech signal and acoustic impulse response. Observation model can be broadly classified in to linear affine or additive relation between clean and reverberant features in the logarithmic Mel power spectral domain. Models that describe reverberation as an additive distortion in the power spectral domain. Models that consider reverberation as the convolution in the power spectral domain.[2][3]

Voice activity detection (VAD) is used to determine voice active region.VAD is an important step in the speaker recognition.VAD depends upon the type of algorithm and model that is used to determine speech and non-speech region. Single frequency filtering is a method that is used to discriminate between speech and non-speech. Due to correlation among samples characteristics of speech samples at each frequency is different from that of noise.SNR of speech is high for many values but for speech always SNR values will be small. For every block of data DFT is calculated and obtaining single frequency information. Temporal variation of energy taken in to consideration and a weighting factor is determined for each band. Mean and variance is used to find out a function that is used to for decision logic to discriminate speech and non-speech. The signal exploit the properties of speech signal it does not need any training data. Many studies this discrimination method compared with the adaptive multirate filtering method (AMR) [4].
Comparison of VAD with AMR is that
Adaptability: AMR is adaptable to various types of noises.
No prior information: It does not require any training data or prior information about the noise signal.
Threshold estimation: Threshold estimation does not require any non-speech beginning.

II. SIGNAL MODEL AND FEATURE EXTRACTION

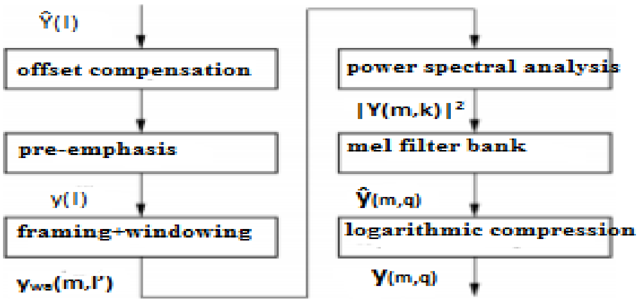


fig 1 Feature extraction

Feature extraction from the microphone signal according to a slightly modified ETSI standard Front end. It is illustrated in the figure 1.The speech signal from the microphone is passed to offset compensation it is used to eliminate dc component and a pre-emphasis state before voice activity detection. Then it is converted to a digital signal. Then, the windowed by an analysis window. Analysis window of length Lw

In the figure m denotes the frame index and l' denote discrete time index within each frame are then transformed to the frequency domain by applying DFT.

$$\mathbf{Y_{(m,q)} = \sum^{Lw-1}{}_{l'=0}\ y_{wa}(m,l')\ e^{-j(2\Pi/K)kl'}} \ldots\ldots\ldots\ldots\ldots\ldots(1)$$

Here **k €{0,……..,K-1}**frequency bin index K is the number of frequency bins and j denote the imaginary unit. By integrating triangularly weighted power spectrum within each total of overlapping Mel frequency bands, indexed by

q €{0,………………..,Q-1}

$$\mathbf{\hat{y}_{(m,q)} = \sum^{kq(u)}{}_{kq(l)}\ {}_{|}Y_{(m,q)|}{}^{2}\Lambda q(k)} \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots(2)$$

The center frequencies for corresponding triangular weighting functions are equally spaced on the Mel scale.. The width of each Mel band is given by the difference of corresponding upper and lower bounds respectively.

Next process is the compression of Mel spectrum by natural logarithm

$$\mathbf{y_{(m,q)} = \ln \hat{y}_{(m,q)}} \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots(3)$$

Further it is notated by vector notation

$$\mathbf{{}^{\prime\prime}\Upsilon_{(m,q)} = \{y_{0},\ldots\ldots\ldots\ldots, Y_{Q-1}\}} \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots(4)$$

III. PROPOSED VAD ALGORITHM

A. Envelope of Speech Signal at Each Frequency

The discrete time speech signal s(n) is differenced so that we will get voice variation x(n)=s(n)-s(n-1) the sampling frequency is f_s.Then the signal is multiplied by a complex sinusoid of a given normalized frequency wk

$$\mathbf{x_k(n)=x(n)e^{jwkn}} \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots(5)$$

Signal is passed through single pole filter whose transfer function is given by

$$\mathbf{H(z) = 1/(1+z^{-1})} \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots(6)$$

The single pole has a pole on the real axis at a distance of r from the origin. The location of root is at z=r in the z plane, which corresponds to half the sampling frequency.

$$\mathbf{y_k(n) = -r\ y_k(n-1) + x_k(n)} \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots(7)$$

Envelope of the signal is given by

$$\mathbf{e_k(n) = \sqrt{}y_{kr}{}^{2} + y_{ki}{}^{2}} \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots(8)$$

where y_{kr} and y_{ki} are the real and imaginary components of $y_k(n)$.Since the filtering of $x_k(n)$ is done at $f_{s/2}$,the above envelope corresponds to the envelope of the signal filtered at a desired frequency .

The above method of estimating the envelope of the component at a frequency is termed as single frequency filtering (SFF) approach. The choice of the filter with a pole at z=r for estimating the envelopes of the filtered signals is likely to be more accurate, as the envelopes are computed at the highest frequency possible. Also, choosing a filter at a fixed frequency for any desired frequency avoids scaling effects due to different gains of the filters at different frequencies. If the pole is chosen on the unit circle, i.e., it may result in the filtered output becoming unstable. The stability of the filter is ensured by pushing the pole slightly inside the unit circle. Hence is chosen as 0.99.

In this study, the envelope is computed at every 20Hz in the range 300 Hz to 4000 Hz as a function of time. The frequency range 300 - 4000 Hz is chosen, as it covers the useful spectral band of speech. Thus we have envelopes for 185 frequencies as a function of time. In principle, the envelope can be computed at any desired frequency.

B. Weighted Component Envelopes of Speech Signal

Since speech signal has large dynamic range in the frequency domain, the signal may have high power at some frequencies at each instant. At those frequencies the SNR will be higher, as the noise power is likely to be less due to more uniform distribution of the power. Even for noises with non uniform distribution of power, the lower correlations of noise samples result in a lower dynamic range in the spread of noise power across frequencies, compared to speech. Note that the spectral dynamic range gives an indication of the correlation of the samples in the time domain.

The noise power creates a floor for the envelope at each frequency, and the floor level depends on the power distribution of noise across frequency. The floor is more uniform across time if the noise is nearly stationary. Even if the noise is non-stationary, it is relatively stationary over larger intervals of time than in speech. In such cases, the floor level can be computed overlong time intervals at each frequency, if needed.

To compensate for the effect of noise, a weight value at each frequency is computed using the floor value. For each utterance, the mean (μ_k) of the lower 20% of the values of the envelope at each frequency is used to compute the normalized weight value at that frequency. The choice of 20% of the values is based on the assumption that there is at least 20% of silence in the speech utterance. The normalized weight value at each frequency is given by

$$w_{k} = (1/\mu_{k})/(\textstyle\sum^{N}_{l=0} \mu_{l}) \text{……………………………………(9)}$$

where N is the number of channels. The envelope at each frequency f_k is multiplied with the weight value w_k to compensate for the noise level at that frequency. The resulting envelope is termed as weighted component envelope. Note that by this weighting, the envelope at each frequency is divided by the estimate of the noise floor (μ_k).At each time instant, the mean ($\mu_{(n)}$) of the square of the weighted component envelopes computed across frequency corresponds approximately to the energy of the signal at the instant (Fig. 2(c)). The $\mu_{(n)}$ is expected to be higher for speech than for noise in the regions where speech signal is present, as the noise components are de-weighted. At each time instant, the standard deviation (σ(n)) of the square of the weighted component envelopes computed across frequency will also be relatively higher for speech than for noise in the regions of speech due to formant structure (Fig. 2(d)). Hence ($\mu_{(n)+}$ σ(n)) is generally higher in the speech regions, and lower in the non-speech regions. Since the spread of noise(after compensation) is expected to be lower, it is observed that the values of (σ(n)-$\mu_{(n)}$) are usually lower in the non-speech regions compared to the values in the speech regions (Fig. 2(e)). Multiplying ($\mu_{(n)}$ + σ(n)) with (σ(n)- $\mu_{(n)}$) gives (σ(n)^2 - $\mu_{(n)}{}^2$),which highlights the contrast between speech and non speech regions.Figs.2 and 3 illustrate the features of $\mu_{(n)}$, σ(n) and (σ(n))- $\mu_{(n)}$) for an utterance corrupted by pink noise at -10 dB and 5 dB ,respectively. Due to large dynamic range of the values of(σ(n)^2 - $\mu_{(n)}{}^2$),), it is difficult to observe the speech regions with small values of (σ(n)^2 - $\mu_{(n)}{}^2$).To highlight the contrast between speech and non-speech regions, the dynamic range is reduced by computing

$$\partial\ (\mathbf{n}) = \sqrt[M]{\sigma(\mathbf{n})^2 - \mu(\mathbf{n})^2} \text{……………………………..(10)}$$

where M is chosen as 64. The value of is not critical. Any value of in the range of 32 to 256 seems to provide good contrast between speech and non-speech regions in the plot of ∂(n). In computing ∂(n) , only the magnitude of (σ(n)^2 - $\mu_{(n)}{}^2$) is considered. If the sign of (σ(n)^2 - $\mu_{(n)}{}^2$) is assigned to ∂(n), the values will be fluctuating around zero in the non-speech regions for most types of noise (see Fig. 2(f) for pink noise), but the short time (20 msec) temporal average value will be small and fluctuating, making the noise floor uneven. This makes it difficult to set a threshold for deciding non-speech regions. The values of ∂(n) will have a high temporal mean value in the non-speech regions, with small temporal variance(Fig.2(g)).This helps to set a suitable threshold to isolate non-speech regions from speech regions.

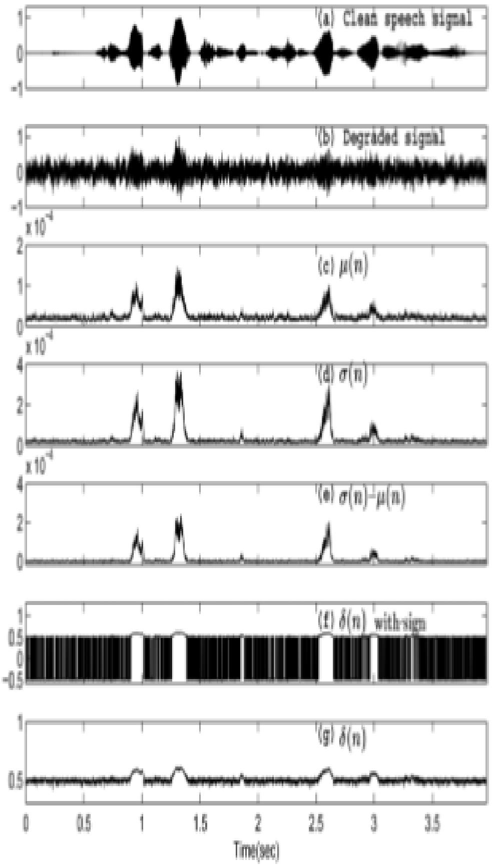


Fig. 2. (a) Clean speech signal. (b) Speech signal corrupted by pink noise at −10 dB SNR. (c) $\mu(n)$. (d) $\sigma(n)$. (e) $\sigma(n) - \mu(n)$. (f) $\delta(n)$ along with sign. (g) $\delta(n)$.

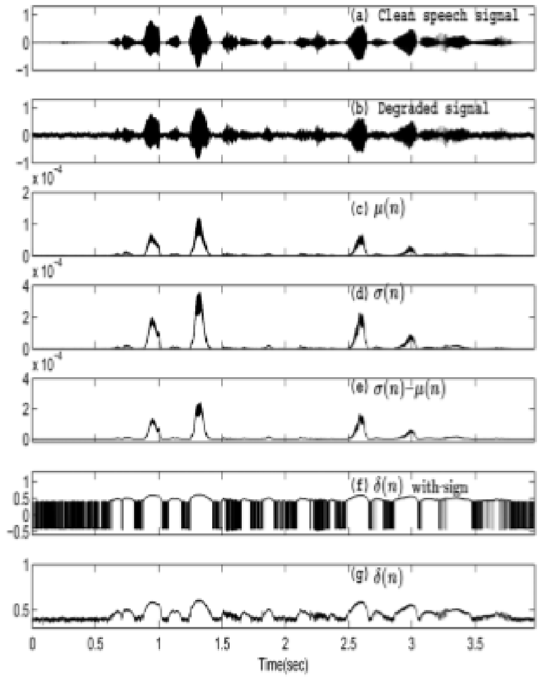


Fig. 3. (a) Clean speech signal. (b) Speech signal corrupted by pink noise at **5 dB** SNR. (c) $\mu(n)$. (d) $\sigma(n)$. (e) $\sigma(n) - \mu(n)$. (f) $\delta(n)$ along with sign. (g) $\delta(n)$.

Note that, by considering the ∂(n) values without sign, we are losing some advantage in the discrimination of non-speech regions, which has both positive and negative values, compared to speech regions which have mostly positive values. The ∂ (n) values with M=64 are used for further processing for decision making. The range of ∂ (n) with sign value (Fig. 2(f)) is different from values (Fig. 2(g)). The small temporal spread of ∂(n) values in the non-speech regions and its mean value helps to fix a suitable threshold. The ∂(n) values in the non-speech regions is dictated by the noise level. The values in non- speech regions are high for pink noise degradation at dB SNR (Fig. 2(g)) than at 5 dB SNR (Fig. 3(g)) Note the changes in the vertical scales in Figs. 2(f) and 2(g), and also in Figs.3(f) and3(g),to understand the significance of using the absolute value, i.e. ∂(n) without sign

C Decision Logic

The decision logic is based on ∂(n) for each utterance, by first deriving the threshold over the assumed(20% of the low energy) regions of noise, and then applying the threshold on temporally smoothed ∂(n) values .The window size used for smoothing ∂(n) is adapted based on an estimate of the dynamic range (ρ) of the energy of the noisy signal in each utterance, assuming that there is at least 20% silence region in the utterance. The binary decision of speech and non-speech at each time instant, denoted as 1 and 0, respectively, is further smoothed (similar to hang over scheme) using an adaptive window, to arrive at the final decision.

1) Computation of threshold (θ): Compute the mean (μ_θ)and variance(σ_θ)of the lower20% of the values ∂ (n) of over an utterance. A threshold of $\theta = \mu_\theta + 3\,\sigma_\theta$ is used in all cases. The θ value depends on each utterance. Thus the threshold value, corresponding to the floor value of ∂(n) , is adapted to each utterance, depending on the characteristics of speech and noise in that utterance.

2) Determination of smoothing window lw:

The energy E_m of the signal x(n) is computed over a frame of 300 msec for a frame shift of 10 msec, where is the frame index. The dynamic range (ρ) of the signal is computed as

$$\boldsymbol{\rho = 10 \log (\max(E_m)/\min(E_m))} \text{.......................(11)}$$

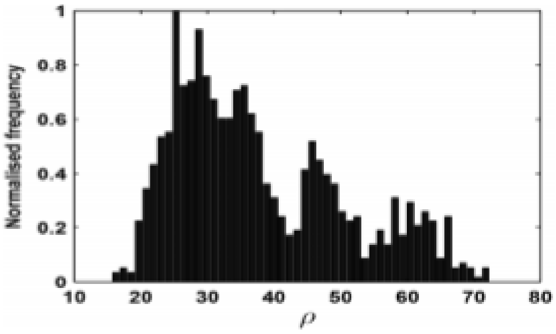


Fig. 4. Histogram of ρ values for distant speech

TABLE I
VALUES OF ρ FOR SPEECH SIGNAL DEGRADED AT SNRS OF −10 dB AND 5 dB FOR DIFFERENT TYPES OF NOISES. THE VALUE FOR CLEAN SPEECH IS 65.28

NOISE	-10 dB SNR	5 dB SNR
white	14.90	22.61
babble	19.64	36.36
volvo	41.62	56.79
leopard	27.77	43.22
buccaneer1	16.13	28.36
buccaneer2	15.67	22.75
pink	16.73	27.60
hfchannel	16.68	28.46
m109	22.44	35.87
f16	17.84	28.88
factory1	20.48	36.28
factory2	24.13	36.30
machine gun	40.52	64.84

The window length lw parameter for smoothing is obtained from the dynamic range (ρ) of the signal 5 dB for different noises. The values are high at 5dB SNR compared to the values at -10dB SNR for the same noise. The ρ values vary for different noises for the same SNR, because the degradation characteristics of noises vary. For distance speech, the histogram of ρ values for utterances in the C3 case is shown in Fig.4.The SNR for distant speech depends on the environmental conditions and on the distance of the speaker from microphone. It is observed that the ρ values for the distant speech are spread out, compared to the ρ values for different noises. This is mainly due to the effects of reverberation. The distribution of ρ values depends on the distance as well. The ρ value for each utterance is used to determine some parameter values for further processing of $\partial(n)$ and for arriving at the decision logic. In cases where the $\partial(n)$ represent the discriminating characteristics of speech and non-speech well, the corresponding ρ values are high, as observed for volvo, leopard and machine gun noises.-In such cases, small value of the smoothing window parameter lw is used. The following values of are lw chosen based on experimentation with speech degraded by different types of noises at different SNR levels:

lw =400msec **for $\rho < 30$**……………………………(12)

lw =300msec **for $30 <= \rho < =40$**………………..(13)

lw =200msec **for $\rho < 40$**……………………………(14)

3) Decision logic at each sampling instant:

The values of $\partial(n)$ are averaged over a window of size lw to obtain the averaged value $\partial(n)$ at each sample index . The decision d(n) is made as follows:

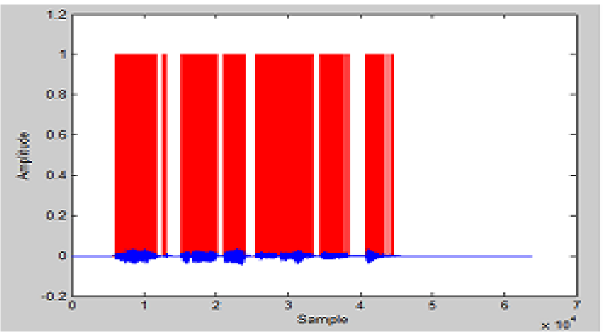
d (n)=1 **for avg($\partial(n)$)> θ**…………………………(15)

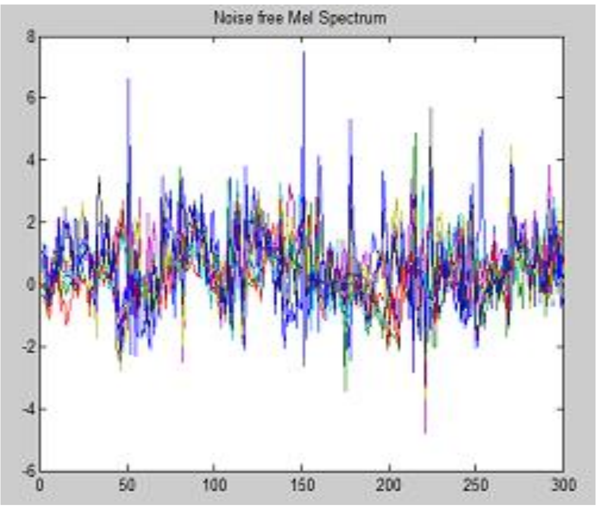
d (n)=0 **for avg($\partial(n)$)<= θ**………………………(16)

d (n)=1 represents voice and d(n)=0 represents non-speech region .After that this is given to an observation model it effectively identify speaker recognition.

IV. SIMULATION RESULTS

Voice detection can be simulated using MATLAB. By adaptive thresholding it can effectively find out the voice active region. Observation model is used to reduce the error between clean speech signal and the actual signal. Mel spectrum is obtained for noisy reverberant speech signal .From the observation model and Mel filter bank it will calculate a noise free Mel spectrum.





V. CONCLUSION

By adaptive thresholding speech and non-speech discrimination is done after reading an audio signal. A reverberant condition is applied. After that it will analyze power spectrum. Then Mel power spectral coefficients are calculated .Observation Model used to find out a relation between clean speech signal and noisy reverberant speech signal .A compensation constant is added to compensate for signal degradation. For compressing the information a Logarithmic compression is used. After this, Logarithmic Mel power ceptral coefficients (LMPSC's) are calculated then it will compare the LMPSC's of training data with this. Closest similarity is taken as the idenfied result.

REFERENCES

[1] ETSI standard document, Speech Processing, Transmission and Quality Aspects (STQ); Distributed speech recognition; Advanced frontend feature extraction algorithm; Compression algorithms, ETSI ES 202 050 V1.1.5 (2007-01), ETSI

[2] T.Takiguchi and M.Nishimura,“ Acoustic model adaptation using first order prediction for reverberant speech,”in Proc. ICASSP, May2004, pp. 869–872.

[3] ASehrand W.Kellermann,“Towards robust distant talking automatic speech recognition in reverberant environments,” in Speech and Audio Processing in Adverse Environments, ser. Signals and Communication Technology, E.Hänslerand G.Schmidt ,Eds. Berlin/Heidelberg, Germany: Springer, 2008, pp. 679–728.

[4] ETSI, Voice activity detector (VAD) for adaptive multirate (AMR) speech traffic channels, ETSIEN 301708v.7.1.1, Dec.1999