ISSN (Online) 2321-2004
ISSN (Print) 2321- 5526

IJIREEICE

**International Journal of Innovative Research in
Electrical, Electronics, Instrumentation and Control Engineering**

ISO 3297:2007 Certified
Vol. 6, Issue 6, June 2018

Detection of Human Voice in Noisy Environment

Ms. Utkarsha V. Chougule[1], Dr. S. K. Shah[2]

PG Student, Department of Electronics and Telecommunication, Smt. Kashibai Navale college of Engineering, Pune, India[1]

HOD PG, Department of Electronics and Telecommunication, Smt. Kashibai Navale college of Engineering, Pune, India[2]

Abstract: Speech is a natural and easiest way of communication. Various researches have been done on voice activity detection and automatic speech recognition system. This paper presents a novel and efficient approach of human voice activity detection in noisy environment. Voice Activity Detection (VAD) is a primary method of speech processing. It refers to a class of methods which detects whether a sound signal contains speech or not i.e. it discriminates human speech signal from other sounds. In real life applications it is not feasible to focus on characteristics of noise due to their random occurrence. Hence to achieve the purpose formants are used as a main features which represents human voice characteristics very effectively. Here clean and noisy speech samples are taken and it is represented that formants of clean speech and noisy speech occurs in same range of frequencies. Thus formants are prominent feature for VAD. To check the robustness of the implemented system clean speech samples are compared with noisy speech samples at various SNR levels.

Keywords: Voice activity detection, Automatic speaker recognition, Formants

I. INTRODUCTION

Speech signal is produced when air from the lungs is passed through the vocal tract and then comes out from mouth and nose. The speech signal produced due to vibration of vocal cords is called as voiced speech. Thus sound waves are produced when speech is modulated by vocal tract system. The resonating frequency of vocal tract is called as formant. Formants can thus be considered as characteristic feature of human voice. In real life applications such as Telephone, voice conference, Automatic Speech Recognition (ASR), speech coding voice activity detection plays important role. Voice Activity Detection (VAD) is a fundamental front-end processing system used in all above applications. VAD is defined as discrimination of human speech from various other background noises. VAD systems consist of two core stages: Feature extraction and Decision making. In real life applications various types of noise can occur randomly and in an unexpected manner. That time it is not possible to derive characteristics of noise because any sound other than human voice is a single type of noise. So it is almost infeasible to have prior knowledge about them. Therefore, it is convenient to focus on human voice characteristics rather than noise.

II. LITERATURE SURVEY

Many researchers have been done till the date on voice activity detection. Many methods are present which represents characteristic features of human speech. Acoustic feature such as Zero crossing rate [2], energy [2], were used as a speaker discriminative features by Rabiner *et al*. As we know when speech signal is produces using speech organs such as lips, tongue, nose, glottis there does not exist only human voice though it is recorded in acoustic rooms. Along with human speech it contains background noises as well as silence and pauses. So in many speech related application it is more efficient to focus on human speech for VAD because there are random noises that occur and on which we cannot focus because it is time consuming. Therefore in this work formants are used more conveniently for VAD.

III. FORMANT EXTRACTION METHODS

The vocal tract is a non-uniform circular tube having proper terminating conditions. Vocal tract is open at mouth end and closed at larynx. The excitation for generating voiced and unvoiced speech is different. For voiced speech the excitation is a periodic impulse train and unvoiced signal is generated when random noise generator is present as excitation source. The excitation passed through vocal tract at some resonant frequencies called as formants. Thus formants are considered as important acoustic feature representing presence of human speech. [1].

There are various methods used for formant extraction. Following sections describes these methods in brief manner.

DOI 10.17148/IJIREEICE.2018.6611

Log spectrum: The speech signal is given as an input. This signal is analyzed over a short window of 20-30ms in order to track small variations in signal. Fast Fourier Transform is then calculated on windowed speech. Finally the log magnitude block operates on the calculated FFT spectrogram so that results can be obtained in dB. The final log spectrum plot shows that the slowly varying envelops correspond to the vocal tract parameters and formants.

Hidden Markov Model: A. Acero proposed method of tracking formant using Hidden Markov Model (HMM).Here HMM feature vector is generated using first three formant frequencies. Then the recordings are used to train the HMM model which further generates formants. This is a data-driven speech synthesizer. [3]

Linear predictive spectra: In paper [4] S. McCandless has introduced a method to track first three formants of continuous speech. Linear predictive spectra is used as input to the system which gives peaks as output. These peaks are nothing but formants.

Linear predictive coding: Formants are generated due to resonance frequencies of vocal tract. So one can consider vocal tract as a linear filter. Thus the formants are presented as poles of source-filter model .Linear predictive coding is a method which calculated likely samples by referring its previous samples. The peaks obtained from spectrum can be considered as formant. [5]

Cepstral method: Cepstrum can be defined as the inverse Fourier transform of logarithm of power spectrum. Cepstral coefficients give the information about vocal tract and its excitation. A Cepstral method needs further smoothing procedure in order to obtain better results. [6]

IV. PROPOSED METHODOLOGY

Formants are defined as resonance frequencies of vocal tract. In human voice first four formants are considered on primary basis. Fig. 1 shows how formant extraction is done using Linear Predictive Coding (LPC). As we know speech signal is a rapidly varying signal. Pre-emphasis is a way of compensating for the rapid decaying spectrum of speech. pre-emphasis and de-emphasis works since speech signal is bandlimited and relatively low frequency (upto 4KHz). Pre-emphasis boosts the high frequency component. After this Hilbert Transform is used so that both real and imaginary parts of the signal can be considered. Use of Hilbert transform allows us using various real and complex values as well. The output of this stage is given to adaptive filter bank. This filter bank is also called as formant tracking filters. Adaptive filter bank consists of all-zero filter (AZF) and Dynamic Tracking Filter (DTF). Both are implemented in cascading manner so that the result is calculated over bandpass filtered speech signal and then formant calculation is done on bandpassed speech signal. Poles and zeros of formant filters are updated every time so that accurate formant frequency is calculated. F1 to F4 are the four formants that are to be calculated. The first four formant frequencies of voiced speech segments are F1, F2, F3 and F4 are estimated from the four bands of the adaptive bandpass filterbank using first-order linear prediction on each band. Here hamming window is used for windowing purpose. In order to estimate a particular formant frequency from the spectrum the energy calculated in that formant band has to be above an "energy threshold level," so that speech a signal is considered as voiced. This energy threshold level is set during initial stage of the algorithm and for i^{th} formant this threshold level is updated depending on voiced part of speech. In this way F1, F2, F3 and F4 are estimated.

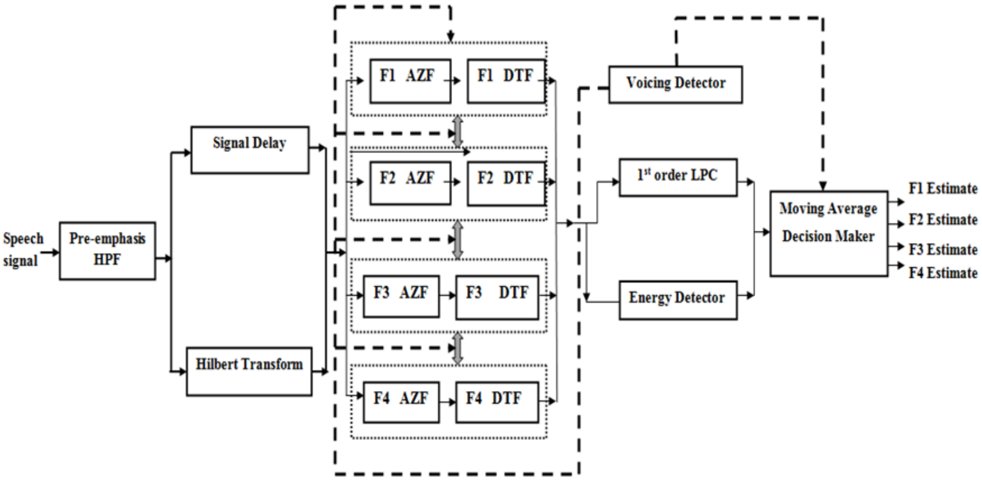

Fig. 1 Block diagram of Formant Tracker

DOI 10.17148/IJIREEICE.2018.6611

After the process of formant tracking main part is to train the speech model. In this work two models are derived one is training for formant and another is formant testing model as shown in Fig.2.

Training model- In this model clean speech signal is given as input to the system. After this pre-emphasis is done, to obtain better results. This pre-emphasised signal is given to formant tracking function. In this way formants of input signal are calculated and training formant model is formed.

Testing model- In testing model input speech signal is not clean speech. Here the input signal is noisy speech. Various noises which are stationary as well as non-stationary that are available in environment are mixed with clean speech at different SNR levels. The inputted noisy speech signal is pre-emphasised and given to formant tracker. After calculating formants of noisy speech they are compared with training model. Based on comparison decision is made on correct Voice activity detection. Finally percentage of accurately recognised VAD is calculated.

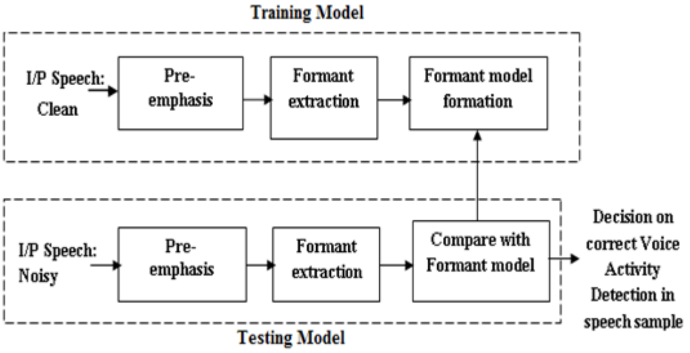

Fig. 2 VAD decision making model

V. RESULTS AND DISCUSSION

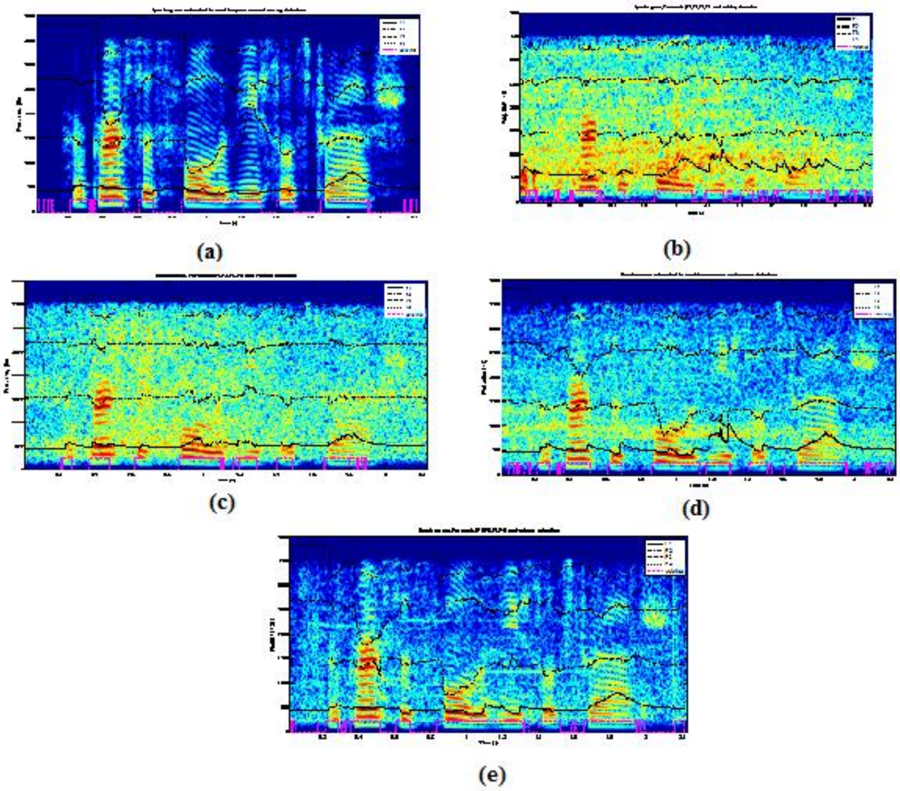

Fig. 3 Spectrogram and Formants of (a) Clean speech sample (b) Noisy speech i.e. street noise mixed speech sample at 0dB (c) Noisy speech i.e. street noise mixed speech sample at 5dB (d) Noisy speech i.e. street noise mixed speech sample at 10dB (e) Noisy speech i.e. street noise mixed speech sample at 15dB

IJIREEICE

ISSN (Online) 2321-2004
ISSN (Print) 2321- 5526

International Journal of Innovative Research in
Electrical, Electronics, Instrumentation and Control Engineering
ISO 3297:2007 Certified
Vol. 6, Issue 6, June 2018

Fig. 3 shows the Spectrograms of clean speech sample and of noisy speech samples: here street noise is used at 0dB, 5dB, 10dB and 15dB SNR levels. The main reason behind choosing street noise is it is non-stationary noise. By using this type of noise, system robustness can be achieved. First four formants are extracted and voicing detection is shown. Formants lie in both clean and noisy speech samples which in turn indicate that human voice has characteristic features which are prominent in noisy environment as well. In non-stationary noise also these formant frequencies occur, so it can be concluded that one can use these formants as a centre of all the training and testing models.

Table 1 shows the location of formant frequencies of clean speech and noisy speech i.e. mixture of clean speech and street noise at various SNR levels such as 0 dB, 5 dB, 10 dB, 15dB.The formants are extracted from clean speech sample S1 having length 2sec and sampled at frequency 8 kHz.

Table 1 Formant frequencies of speech sample clean_S1 and street_S1

Speech sample S1	F1	F2	F3	F4
Clean	1113.879	1435.876	2303.6	3357.65
street_0dB	708.6135	1461.9	2375.55	3478.8
street_5dB	665.8424	1466.8	2306.85	3463.95
street_10dB	761.0473	1304.193	2319.75	3434.15
street_15dB	548.3007	1458.1	2338.6	3458.4

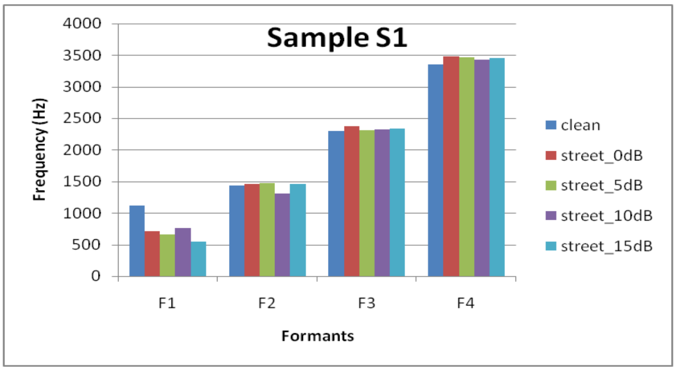

Fig. 4 Formants of speech sample clean_S1 and street_S1

For same speech sample clean_S1 and street_S1 frequencies are extracted. In order to get a better view bar graph is used which contains Formants on x-axis and Frequency (Hz) on y-axis as shown in Fig. 4. From Fig. 4 it is observed that for clean speech sample S1 and for noisy speech sample S1 at SNR levels 0dB-15dB, formants are calculated using formant tracking filters. The bar graphs shows that formants F1, F2, F3, F4 at various SNR levels such as 0dB, 5dB, 10dB, 15dB for street noise occur almost same range. The main purpose to do this is to identify whether the formants of clean speech sample and noisy speech sample occur at same range of frequencies. In this way different clean speech samples and various noisy speech samples are taken to check whether our assumption is correct. Training and testing model compares the formants calculated for both inputted speech signals i.e. clean speech and noisy speech. The decision is made based on matching evaluation metric.

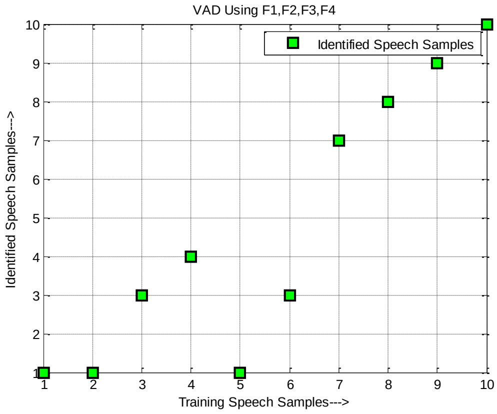

Fig. 5 Graph of Correctly detected VAD for street noise at SNR=10dB

From Fig. 5 it is observed that horizontal axis contains training speech samples and vertical axis contains identified speech samples. When sample one match with sample one then decision is correct that is VAD in sentence S1 is correctly identified as VAD in S1 otherwise decision is not correct. It shows that VAD is correct identified in speech samples S1, S3, S4, S6, S7, S8, S9, S10. Vector quantization is used to correctly identify the voice activity. Vector quantization technique forms a training codebook which contains formants of clean speech samples and formants of testing speech samples i.e. noisy speech samples are stored in testing codebook. At the time of testing if training and testing codebook of same input sample match then it is called as correct voice activity detection. Here in order to find the correct match distance between particular code vectors is measured and if it has minimum distance match is obtained.

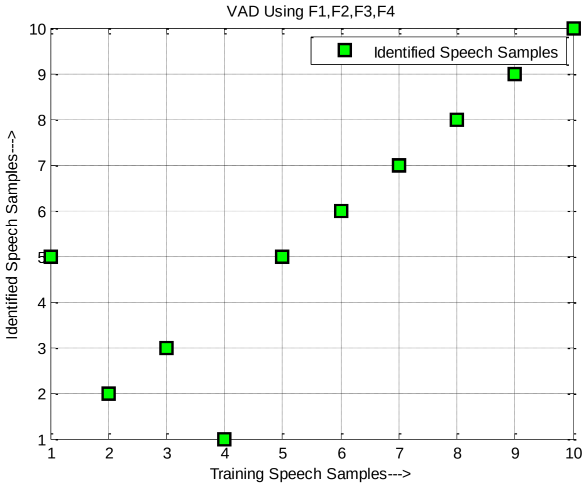

Fig. 6 Graph of Correctly detected VAD for exhibition noise at SNR=15dB

Fig. 6 shows that VAD is correct identified in speech samples S2, S3, S5, S6, S7, S8, S9, S10. The correct VAD is carried out if codebooks of training and testing models match.

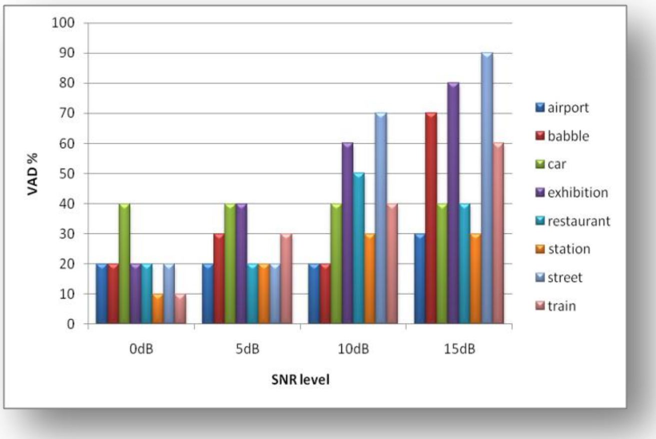

Fig. 7 VAD % for eight different types of noises

Fig. 7 shows overall results of correct VAD performed under various conditions .The observation can be made that the percentages of VAD at 0dB are very low as compared to high SNR levels. At higher SNR levels such as at 10dB and 15dB percentage of correctly detected speech samples i.e. speech samples when compared with training models identifies correct voice activities thus giving positive scores. At 10 dB and15 dB SNR level the results obtained for street, exhibition, babble, train, restaurant, car, airport and station noise are more. Thus we can say that voice activity detection is done correctly in these SNR level.

VI. CONCLUSION

Speech is a primary way of communication for human beings. Thus in many applications such as Automatic speech recognition (ASR) it is important to identify human voice in presence of background noise. Here voice activity detection system is implemented to detect the human voice when background noise is also present. For achieving the objectives formants are used in both training and testing models as a main features. This system clearly shows that formants of clean speech and noisy speech occur at same frequency range. These formants are further used for voice activity detection in different types of noise. Voice activity detection is carried out for stationary as well as non-stationary types of noise at various SNR levels. Hence, formants are prominent features representing presence of human voice and discriminates it from surrounding noise.

REFERENCES

[1] S. D. Apte, Speech and Audio Processing, Wiley-India Edition,ISBN-978-81-265-3408-1,2012

[2] L. R. Rabiner and M. R. Sambur, “An algorithm for determining the endpoints of isolated utterances,” *Bell Syst. Tech. J.,* vol. 54, no. 2, pp.297–315, 1975

[3] Acero, “Formant analysis and synthesis using hidden Markov models,” in *Proc. EUROSPEECH, 1999, pp. 1047–1050.*

[4] S. McCandless, “An algorithm for automatic formant extraction using linear prediction spectra,” *IEEE Trans. Acoust., Speech, Signal Process.*,vol. ASSP-22, pp. 135–141, 1974.

[5] Q. Yan, E. Zavarehei, S. Vaseghi, and D. Rentzos, “A formant tracking LP model for speech processing,” in Proc. *InterSpeech-ICSLP,* 2004,pp. 2409–2412.

[6] L. R. Rabiner and R. W. Schafer , Digital Processing of Speech Signals Englewood Cliffs, NJ: Prentice-Hall, 1978.

[7] K. Mustafa and I. C. Bruce,“Robust formant tracking for continuous speech with speaker variability,” *IEEE Trans. Audio, Speech, Lang. Process*. vol.14, no.2,2006

[8] C.Yoo,H. Lim, and D. Yook. “Format based robust voice activity detection”, *IEEE Trans. Audio, Speech, Lang. Process*. vol.23,no.12,2015

DOI 10.17148/IJIREEICE.2018.6611