ÇUKUROVA UNIVERSITY
INSTITUTE OF NATURAL AND APPLIED SCIENCES

MSc THESIS

Mohamad Dia ABDULKARIM

NOISE ROBUST SPEAKER RECOGNITION UNDER UNKNOWN NOISE ENVIRONMENT

DEPARTMENT OF COMPUTER ENGINEERING

ADANA, 2015

ÇUKUROVA UNIVERSITY
INSTITUTE OF NATURAL AND APPLIED SCIENCES

NOISE ROBUST SPEAKER RECOGNITION UNDER UNKNOWN NOISE ENVIRONMENT

Mohamad Dia ABDULKARIM

MSc THESIS

DEPARTMENT OF COMPUTER ENGINEERING

We certify that the thesis titled above was reviewed and approved for the award of degree of the Master of Science by the board of jury on / /2015

…….....	……	……..……..……
Assoc. Prof. Dr. Zekeriya TÜFEKÇİ	Assoc. Prof. Dr. Sami ARICA	Asst. Prof. Dr. Lütfü SARIBULUT
SUPERVISOR	MEMBER	MEMBER

This MSc Thesis is written at the Department of Institute of Natural And Applied Sciences of Çukurova University.
Registration Number:

Prof. Dr. Mustafa GÖK
Director
Institute of Natural and Applied Sciences

ABSTRACT

MSc THESIS

NOISE ROBUST SPEAKER RECOGNITION UNDER UNKNOWN NOISE ENVIRONMENT

Mohamad Dia Abdulkarim

ÇUKUROVA UNIVERSITY
INSTITUTE OF NATURAL AND APPLIED SCIENCES
DEPARTMENT OF COMPUTER ENGINEERING

Supervisor : Assoc. Prof. Dr. Zekeriya TÜFEKÇİ
Year: 2015, Pages: 59
Jury : Assoc. Prof. Dr. Zekeriya TÜFEKÇİ
: Assoc. Prof. Dr. Sami ARICA
: Asst. Prof. Dr. Lütfü SARIBULUT

Many approaches designed to increase the performance and stability of the speaker verification system under adverse conditions were studied, performance will be at its peak when no mismatch occurs between training and testing conditions. Therefore, among all these methods, the Parallel Model Combination (PMC) appears to be the most adequate and capable techniques to handle such issue, where it compensates by minimizing the mismatch occurring between the test and the training conditions. In this study the main goal is to increase the performance of the speaker verification system. In previous studies, the (PMC) method was used to estimate the noisy speech parameters by using clean speech and noise model, assuming noise statistics are known. In this study, it is assumed that noise is not known. Noise is estimated using common VAD techniques from the noisy speech.

Accordingly non-speech that is characterized by a certain VAD technique can be considered to estimate the noise model. In this study two common VAD techniques are used to to estimate the noise model, and PMC is used to estimate the noisy speech for all methods. The method that estimates the noise model directly from the noise signal is referred to the baseline method.Thereafter, the performance of the baseline is compared with that of the VAD techniques.

NIST 1988 speaker recognition databases and NOISEX-92 databases were used to evaluate the performance of the speaker verification system. Expiremental results shows that the performance of the method that used the VAD techniques to estimate the noise model is comparable with the baseline method in the case of high signal-to-noise-ratio (SNR) levels, however in the case of low SNR levels, baseline method yielded better results in terms of equall error rate (EER).

Key Words: Voice Activity Detection, Noise Estimation, Parallel Model Compensation, Equal Error Rate.

ÖZ

YÜKSEK LİSANS TEZİ

BİLİNMEYEN GÜRÜLTÜ ÇEVRE ORTAMINDA GÜVENLİ KONUŞMACI TANIMLAMA

Mohamad Dia ABDULKARIM

ÇUKUROVA ÜNİVERSİTESİ
FEN BİLİMLERİ ENSTİTÜSÜ
BİLGİSAYAR MÜHENDİSLİĞİ ANABİLİM DALI

Danışman : Doç. Dr. Zekeriya TÜFEKÇİ
Yıl: 2015, Sayfa: 59
Jüri : Doç. Dr. Zekeriya TÜFEKÇİ
: Doç. Dr. Sami ARICA
: Yrd. Doç. Dr. Lütfü SARIBULUT

Konuşmacı doğrulama sisteminin performansı ve karalalığı gürültülü ortamlar altında arttırmak için, üzerinde bir çok yaklaşım ve tasarım araştırlmıştır. Eğitim ve test koşulları arasında hiçbir uyumsuzluk olmadığı zaman sistem en yüksek performansta çalışır. Bu nendenle, tüm metodlar arasında Paralel Model Kombinasyonu (PMK), bahsedilen konuyu en uygun ele alan teknik olarak görünmektedir. Bu metod, test ve eğitim koşulları arasında olan uyumsuzluğu, minimize ederek çözmektedir. Bu çalışmanın ana amacı, konuşmacı doğrulama sisteminin performansını arrtırmaktır. Daha önceki çalışmalarda, PMK tekniği, gürültü istatistiğinin bilindiğini varsayarak, gürültüsüz konuşma ve gürültü modelleri kullanılarak, gürültülü konuşma modelinin parametrelerini tahmin etmek için kullanılmıştır. Bu çalışmada gürültünün bilinmediği varsayılmıştır. Ses Aktivite Tespiti (SAT) teknikleri kullanılarak gürültü, gürültülü konuşmadan tahmin edilmiştir.

Buna göre, karakterize edilen belirlibilir SAT tekneği konuşma olmayan kısmı, gürültü modelini tahmin etmek için kullanılabilir. Bu çalışmada, doğrudan gürültü sinyalinden gürültü modelini tahmin eden metod, temel metod olarak adlandırılmıştır. Bundan sonra, temel metod performansı SAT tekniklerinin performansı ile karşılaştırılmıştır.

Konuşmacı doğrulama sisteminin performansı değerlendirmek için NIST 1998 konuşmacı tanıma ve NOISEX- 92 gürültü veritabanları kullanılmıştır. Deneysel sonuçlar, gürültü modelini tahmin etmek için SAT teknikleri kullanılan metodunun performansı, temel metodunun performansı ile yüksek Sinyal Gürültü Oranları (SGO) için karşılaştırılabilir olduğunu göstermektedir. Fakat, düşük SGO oranlarında temel metod Eşit Hata Oranı (EHO) olarak daha iyi sonuçlar vermiştir.

Anahtar Kelimeler: Ses Aktivitesi Tespiti, Gürültü Tahmini, Parelel Model Tazmini.

ACKNOWLEDGEMENTS

The study presented in this thesis was carried out under the supervision of Assoc. Prof. Dr. Zekeriya TÜFEKÇİ. I would like to express my sincere gratitude to him for his supervision guidance, patience, motivation, useful suggestions and his valuable time that he have without any hasitation for this work.

I would like to thank members of MSc thesis jury, Assoc. Prof. Dr. Sami ARICA for his suggestions and Asst. Prof. Dr. Umut ORHAN for his suggestions and support and I would like to thank Asst. Prof. Dr. Lütfü SARIBULUT for his suggestions.

I would also like to thank my Father and Mother for thier advices and motivation. Not forgetting all of my beloved best friends, office friends, house friends and all who supported me.

CONTENTS **PAGE**

LIST OF TABLES **PAGE**

LIST OF FIGURES **PAGE**

LIST OF SYMBOLS AND ABBREVIATONS

SS	: Spectral Subtraction
g	: Gain factor
X_k	: The Vector of Coefficients Resulting from Applying the *L*-point FFT on the *k*-th Frame
L	: L is the Number of FFT Coefficients
k	: Frame Number
c_k	: The Power Cepstrum
F_0	: Fundamental Frequency
N_s	: Number of States
P_s	: Average Energy of Speech Frames
P_n	: Average Energy of Noise Frames
x	: Signal
n	: Samples number
E_s	: The energy level of each frame
S	: State
t	: Time
a_{ij}	: Transition Probability
A	: Transition Probability Matrix
o_t	: Observation Vector at Time t
$b_j(o_t)$	: Probability of observing o_t given that the system is in state S_j at time t
π	: The Initial State Distribution
π_i	: Probability of being in State i

λ : The compact notation to indicate the complete parameter set of HMM

S_e : The number of Substitution Errors

D_e : The number of Deletion Errors

N_L : The total number of Labels in the Reference Transcriptions

I_e : The number of Insertion Errors

N_{cs} : States are utulized to model clean speech

M_{ns} : States are used to model the noise

N : Frame Size

M : Number of Step Size Samples

$y(n)$: Output Signal

$x(n)$: Input Signal

$w(n)$: Window Function

$x_i(k)$: DFT of speech signal

f : Frequency

mel : Mel scale

k : Index

m_f : Number of filters

$\tilde{c}_n$: Cepstral Coefficients

n_c : Number of Cepstral Coefficients

$\tilde{p}_k$: Mel spectrum obtained from the original spectrum

S_i : the i_{th} components of the speech observation vector

N_i : the i_{th} components of the noise observation vector

E_{op} : Expectation Operator

μ_i	: the i_{th} components of the clean speech mean vectors in the mel-scaled filterbank energy domain
$\tilde{\mu}_i$	: the i_{th} components of the noise mean vectors in the mel-scaled filterbank energy domain
i	: Index
μ	: Mean
Σ	: Variance
$x(n,m)$	: Clean Signal at m frame
m	: Frame number
$y(n,m)$,	: Noisy Signal at m frame
$b(n,m)$	: Background Signal at m frame
$X(\omega,m)$	: Clean Signal Frequency Spectrum
$Y(\omega,m)$	: Noisy Signal Frequency Spectrum
$B(\omega,m)$	: Background Signal Frequency Spectrum
$\varphi_y(\omega,m)$	: The phase of the average spectrum of some non-speech parts
α_{m}	: Over-subtraction factor
β_{m}	: Spectral floor factor
c	: Constant
$\alpha_{\min}$	: Limit of the over-subtraction factor
$\alpha_{\max}$	: Limit of the over-subtraction factor
$\beta_{\max}$	: Limit of the noise floor factor
$\beta_{\min}$	: Limit of the noise floor factor
θ	: Threshold
γ	: Combination Weight
E_i	: The log-energy of each frame
ε	: Arbitrary Constant

θ_{main}	: Primary energy thresholds
$\theta_{\min}$	: Minimum energy thresholds
$E_{\max}$	: Maximum Energy
$\lvert X \rvert^2$	: The powers of noisy speech
$\lvert \hat{N} \rvert^2$	: The estimated noise
γ_s	: The subtraction domain
e	: The gain exponent
g_h	: The maximum gain for noise floor
β_s	: The maximum noise attenuation
X_k^m	: The STFT coefficients
$\lambda_N(k)$	: The noise variance
H_0	: Speech Absent
H_1	: Speech Present
S_k^m	: K-dimensional STFT vectors of speech speech
N_k^m	: K-dimensional STFT vectors of noise speech
X_k^m	: K-dimensional STFT vectors of noisy speech
$\hat{\lambda}_N^m(k)$	: The mean estimate of the noise spectrum
SNR	: In decibels of a signal, S, by computing the ratio of its summed squared magnitude to that of the noise n
GMM	: Gaussian Mixture Model
HMM	: Hidden Markov Model
MS	: Minimum Statistics
HTK	: Hidden Markov Model Toolkit
MFCC	: Mel-Frequency Cepstrum Coefficient

MMSE	: Minimum Mean-Square Error
PMC	: Parallel Model Compensation
SNR	: Signal-to-Noise Ratio
VAD	: Voice Activity Detection
LPCC	: Linear Predictive Cepstral Coefficients
STSA	: Short Time Spectral Amplitude
DCT	: Discrete Cosine Transform
DFT	: Discrete Fourier Transform
FFT	: Fast Fourier Transform
SV	: Speaker Verification
SRE	: Speaker Recognition Evaluation
PDF	: Proabibilty Density Function

1. INTRODUCTION

Many ways of communications exist in our daily lives, such as textual language, body language and speech. Moreover, speech is considered to be the richest and the most powerful communication due to its rich context and character.
In terms of signal processing, speech is a signal which carries message information or data, this signal was used in many, laboratory and real world applications. The information that the signal carries can be characterized into three main types: Speech recognition, speaker recognition and language recognition, which produces speech text, speaker identity and language respectively.

Speech recognition is the process in which a speech signal is converted to a sequence group of words by implementing algorithms in a computer program (Santosh K. Gaikwad, 2010). Recently there has been a growing interest regarding speech recognition in real world applications and conditions, where many implementations and techniques have been processed and applied. Many of these systems are in the market now, some, which are speaker-dependent and others which are discrete systems. Nowadays speech systems are being used as an alternative to keyboards due to the development of the speech recognition field. Even though speech recognition identifies what the speaker says, in other words, what is being said, but it can't identify the person saying these words.

As a part of security issues, identifying the person is necessary and useful. In order to identify the person who produces the speech, speaker recognition field with its techniques and algorithms is employed. During the last decade, there has been a lot of researches and studies regarding speaker recognition field due to its similar characteristics with speech recognition.

Speaker recognition is the process of recognizing a person given a speech utterance (JP CAMPBELL JR, 1997). Speaker recognition has two approaches: to recognize a particular person, which is known as speaker identification or to verify a person who he/she claims to be, which is known as speaker verification. This study will focus on the latter approach.

Speaker verification (*SV*) can be defined as the process of recognizing or verifying an unknown speaker to be one of the speakers among *N* speakers given in a database or in a list, in other words, to verify whether the speech or voice produced from an unknown speaker is identical to the id claimed. (AARON E. ROSENBERG, 1976).

Speaker verification systems consist of two stages: a training stage and a testing stage, each one is considered as an independent module.

During the training stage, the speech data's features from an unidentified speaker are extracted, next a feature model of the extracted data is formed which correspondingly will be used in the testing stage.

In the testing stage, the features of speech data are extracted using several algorithms such as *MFCC* (Mel Frequency Cepstral Coefficients) in our case, based on that, classification is applied using classification techniques in our case *GMM* (Gaussian Mixture Modeling), and further on matching the identity. Speaker verification systems are either text-dependent, where there is some restriction on utterance that users can say or text-independent where the user can say anything without any restrictions (Frederic Bimbot et al, 2004).

When parameterization of clean speech was studied, finding out speaker-discriminative features for speaker recognition tasks was a motivation (R.Schafer et al, 1975). The importance of Cepstral features for speaker recognition, especially the Mel-cepstrum feature was established in the course of that period (B. Atal, 1976).

These features show very good performance in noise-free conditions, however, based on many studies and experiments and due to random modifications of Cepstral distributions in the presence of noisy environments, there were restrictions on using these features causing mismatches between training and testing stages accordingly the largest obstacle that speaker recognition technology challenges is achieving robustness under noisy backgrounds and environments.

Many feature compensation methods were proposed to handle the mismatch and challenges that cause this lack of robustness such as VAD (Voice Activity Detection) methods which works as a classifying method by classifying speech and

non-speech frames into different groups (J. Ramírez et al, 2004). VAD accordingly, drops out non-speech frames to increase recognition.

Another well-known method which handles the mismatch that happens in the training and testing stages is the Parallel Model Combination (PMC), PMC scheme provides a way to compensate the mismatch in the corrupted speech waveform given information about clean speech and corrupting noise, PMC (Gales and Young, 1992).

In this study, we combined the VAD method with the PMC method in order to estimate the speaker models in adverse conditions by using the noisy speech to estimate noise.

2. PREVIOUS WORK

2.1. Previous work in Noise Robust Speaker Verification and Recognition

In this part, the existing methods for noise robust speaker verification and recognition are reviewed.

Achieving robustness in a noisy interfering environment have become a challenge and a difficulty in which the performance of speaker verification and recognition systems encounter. Many methods and algorithms were proposed to overcome this difficulty and to increase the performance at the same time, most of these methods were applied in speech recognition, since speaker verification's intermediate stages are similar to those of speech recognition, the same methods can be used and applied in speaker verification systems (F. Bimbot et al, 2004).

Some methods were proposed for reducing the mismatch that happens between the corrupted waveform and the clean waveform, on of the famous methods proposed was spectral subtraction (Boll, 1979). This method assumes that speech and noise are uncorrelated signals (Nolazco and young, 1994). Spectral subtraction approach preforms well, only when SNR (Signal to Noise Ratio) is stationary and high. However, when SNR is high and non stationary noise exists the performance degrades, to overcome this conflict Minimum Statistics (MS) was introduced to the field by Rainer Martin with the subtraction rule, the noise power estimation method there was a compelling impact on the residual noise (R. Martin, 1994). Estimating the noise power spectral density still remains a challenge.

Furthermore, to overcome this issue Martin suggested a method that is based on two assumptions the background noise and the speech which are generally considered to be statistically, and that the power of a noisy speech signal frequently fade to the power level of the background noise. Accordingly, these assumptions make deriving a precise PSD possible by tracking the minimum of the noisy PSD signal (R. Martin, 2001). This tracking method requires a bias compensation due to the fact that minimum is smaller than the average value.

2.2. Previous studies in Minimum Mean Square Error

MS Knowledge of the clean speech and noise distributions make the method statistically possible. MMSE-STSA, proposed by Ephraim and Malah in 1984 is a historically important speech enhancement method (Kawamura, 2012). In particular, the amplitude and phase of the spectral parts of the noise and clean speech signal are assumed to be independent Gaussian variables, under this hypothesis that clean speech parameterization can be estimated in a statistical framework (Ephraim, 1992). The MMSE - STSA method usually relied on an explicit model of the probability distributions of clean speech and noise (Torre et. a1, 2007). The distributions of the noise and speech are given best results in the log-spectral or cepstral domain than the others. (Van Dalen, 2011)

Ephraim and Malah have proposed not only an efficient spectral gain, also proposed a reliable estimation method to get the a priori SNR (Kawamura, 2012). Ephraim's method utilizes a Gaussian model for the distribution of the spectral parts (Xiong , 2006).

MMSE methods estimate noise parameters, minimizing the distance between clean speech models and cleaned speech parameters, given the noisy speech parts (Torre et. a1, 2007).

2.3. Parallel Model Compensation

PMC(Gales and Young, 1993; Gales, 1997), noisy speech is designed with a HMM with $N_{cs} \times M_{ns}$ states, where N_{cs} states are utilized to model clean speech, and M_{ns} are used to model the noise (Torre et. a1, 2007). By combining speech and noise models in linear spectral domain, The parallel model compensation (PMC) approach compensated the acoustic model (Xiong , 2006). Thus, a standard Viterbi algorithm is done to simultaneous recognition of speech and noise. In terms of non-stationary noises, a lot of states M_{ns} would be used to model the noise. One state can be enough to symbolize the noise, in terms of stationary noises. It should be noted

that the probability distribution of the combined model at each state must be considered that one of the clean speech model and one corresponding to noise (Torre et. a1, 2007). As the noise and the speech are said to be independent and additive, by adding the parameters of the noise model in linear spectral domain, the parameters of the acoustic model are compensated. In particular, for each clean and noise state pair, the mean vectors and covariance matrices of the two models are combined. PMC has also been studied along with Mel Frequency Discrete Wavelet Coefficient features to take benefits of both noise compensation and speech features local in frequency domain (Tufekci et al., 2001).

Whilst the adaptation process, both noise model and the clean acoustic model are transform to linear spectral domain specific additive, bandwidth limited channel mismatch functions and convolutional applied in(Gales, 1998). Because of the resulting complexity of the forms, the log-normal approximation is a famous and efficient choice assumed to the sum of two log-normal distributions is approximately log-normal, although cannot be carried out with delta and delta-delta parameters (Liao and Gales , 2005).

Among the training process, the single state HMM for noise model and the multi-state HMM for clean acoustic model are trained in cepstral domain (Xiong , 2006). In the beginning, where they are combined to produce the noisy acoustic model, and then transformed back to the cepstral domain again and utilized for recognition (Xiong , 2006).

As in the use of PMC for speech enhancement, these model-based schemes depended on the quality of the noise models used (Gales , 1998) and the model accuracy of the relation to the noisy conditions with speech. Single state noise models are efficient and fast, however they can only deal with slow replacing noise statistics. The considering the independence between speech and noise production is not as good as one as speech changes regarding speech with noise level. In the noise model, more states are necessary to handle quickly changing noise, substantially rising computational complexity in the decoding to find the optimal compensation of noise and speech. In addition, the training of the noise models is significant. This method can restore performance to a level comparable to training models with noisy

speech in a 10 dB SNR environment(Gales and Young, 1996). The mixture of the speech and noise is done how the noise and speech are combined to form the noisy speech signal through a mismatch function describing (Liao and Gales, 2005). Different sources are existed for additive noise samples such as the NOISEX-92 database, however convolutional noise samples are hard to achieve. The performance of this method relies on having suitable noise models. PMC has shown to work fairly well (Liao and Gales, 2005).

2.4. Previous Works in Voice Activity Detection

Generally, voice activity detectors extract specific features for classification from the input signal. These features may have good traits in the system but cannot meet success expectations. So feature combination becomes one of the popular techniques in such systems. Since the different features bring out extra computation cost, the feature selection process should be organized very carefully in VAD algorithm. Considering the discriminative and the contribution computational complexity, five kinds of measure were selected in this study. The feature selection process is based on the features' succeed for VAD systems under different noise types and different noise levels. They are the energy measure, the spectral entropy, harmonic structure, cepstral measure, and the long term measures (Zhang, 2014).

2.5. Energy-Based Voice Activity Detection

The energy-based VAD methods are the most famous methods and are widely used in speech recognition application. A variety of energy-based algorithm approaches have been recently proposed for robust voice activity detection (Hsieh, Feng and Huang, 2009). With very little computational complexity, for clean speeches or speeches with less noise, these methods have good performance (Zhang, 2014).

Energy is a simple measure of the power of the signal. We can assume that speech is always louder than background noise, and we can assign the high-energy

frames to speech and lower ones to noise for VAD. When the SNR is low, the simple energy feature cannot separate speech and noise. Energies of sub-bands were used as features in earlier work on energy based VAD to increase the discriminative power of the VAD (Woo et. a1, 2000). Another approach to increase noise robustness is to combine energy-based features with other features, like zero-crossing rate (ZCR) (Rabiner and Sambur, 1975), or the line spectral frequency (LSF) (Benyassine et. a1, 1997).

In general, these methods work well with clean speech or high SNR conditions. But, their discriminative power falls significantly under high noise level such as when SNR falls below 10 dB. Nevertheless, through with their low computation complexity, energy based methods are still developed by some standards and different real-world applications. For example G.729 Annex B (Benyassine et. a1, 1997) uses full-band energy, ZCR and low-band energy (0 to 1 kHz) as feature vector. This criterion remains the most cited work and still being used as the baseline system for performance comparison in many studies. In (Cho and Kim , 2001), energy-based VAD method is used as a preliminary event detector for further categorization in a speech enhancement application (Khoa, 2012

3. MATERIAL AND METHOD

3.1. Material

This section includes explanations and illustrations about the material used in the proposed method. It also includes information about the databases used, the Parallel Model Compensation used for speech recognition, HMM and HTK used for modeling and training, MFCC used for feature extraction, speech scaling and noise addition.

3.1.1. Databases

3.1.1.1. NIST SRE Database

Due to the importance of textual language and the ability to identify people by their voices, the national institute of standards and technology has performed many researches and evaluations, regarding the general problem of text-independent speaker recognition. In this study the NIST 1998 database is used (G. R. Doddington et, al. , 2000). NIST SRE evaluations have researched on focused on the SV issue, which is considered to be a challenging issue due to several factors that can not be controllable such as, acoustical noise, the microphone and electrical transmission, these are considered problems that face the future of speaker recognition development and technology. In order to evaluate research ideas and speech and speaker recognition systems using this database, performance measures were established such as:

- Verification

  Since verification is considered to be a detection task, performance measurements of detection systems are supported.

- Miss/false alarm

  Generally performance of detection systems is represented in two error measures, which are E_{miss} the probability of not detecting the claimed speaker in the non presence case and E_{fa} the probability of falsely detecting the claimed speaker in the presence case. These refer to the miss and false alarm respectively. The measures are computed as

  $$E_{miss} = n_{miss}/n_{target}$$
  $$E_{fa} = n_{fa}/n_{impostor}$$

  where n_{miss} represents the number of trials where the claimed speaker was not detected, n_{target} represents the number of claimed trials, n_{fa} represents the number of trials where the claimed speaker was falsely detected and $n_{impostor}$ represents the imposter trials.

- Equal error rate

  Since miss and false alarms do not produce a single number while measuring the performance, equal error rate combines the miss and false alarm rates into a number by locating the decision threshold when the alarm rates are equal.

The NIST SRE database consists of 500 telephone conversations sampled to 8 kHz , 250 male speakers and 250 female speakers, in this study only the test and train data of the male speakers were used. There are three different training conditions:

- “One-session” training : in this session there only exists 2 minutes of speech which is only taken from one conversation.
- “Two-session” training: in this session data contains 1 minute of speech which is taken from two different conversations and the same phone number.

- “Two-session-full ” training: in this session data contains all speech data taken from two conversations which the two-session training uses.

Two-session full training data were used in this study. There are also three test conditions:

- Test segment duration: in this condition the performance is computed for the test segments with a duration of 3, 10, 30-second separately.
- Same/different number: in this condition, performance is computed for test segments separately, which use the same phone number.
- Same/different handset type: in this condition, performance is computed for segments with different numbers separately using the same handset microphone type.

In this study the test segment duration condition with 30 seconds duration was used.

3.1.1.2. NOISEX-92 Database

NOISEX-92 database is a portion of the NOISEX project, which is a pilot program carried out and developed by NATO RSG.10 (Research study Group) laboratories. The NOISEX’s aim is to motivate dialogue on future experiments and databases, and to discuss some of the problems regarding the effects of noisy environments on recent automatic speech recognition (A.P. Varga et, al. , 1992). The data included in the database are naturally artificial and the recordings are not done in noisy environments. There are eight different noises that are used in the NOISEX-92 database are: speech noise, machine gun, STITEL, Lynx, F16, car noise, factory and operations room, the sampling rate of these noises are 16 kHz according to the SAM standard. In this study, we used only three of the recorded noises which are speech noise, STITEL and F16.

3.1.2. Hidden Markov Model

In this part of the thesis, a summarized study of Hidden Markov models is presented. Several ways for selecting a signal model that obtains and characterize the properties of a signal exist. These model choices can be characterized in terms of classes, into deterministic and statistical models. For the former type of models, generally provides known properties such as the signal being a sine wave, the latter type focuses on extracting only the statistical properties of a signal, such as, Gaussian models, Markov and Hidden Markov Models (HMM). HMM models are considered as probabilistic functions of Markov chains (Lawrence et al., 1989). Before getting to know HMM, Markov chains will be briefly presented.

If (Ω, Σ, P) is assumed to be the probability space, $(S, Pot(S))$ a space that is measurable, then $\{X_t, t \in T\}$ which is defined on the latter space, taking a number of values say $s \in S$ and indexed by set T(non-empty index), is referred to a stochastic process (C. Kohlschein, 2006). There are two types of processes, namely time discrete and time continuous, only the time discrete process is mentioned. A first order Markov chain or for short *Markov chain* is a stochastic process that satisfies the Markov property which is illustrated in equation (3.1), this property states that the probability of reaching a random state $t+1$ depends on no previous states except for the current state t, if the chain was an order of n, in this case the probability of reaching the next states depends on the previous n states.

$$P(X_{t+1} = s_{t+1} | X_t = s_t) = P(X_{t+1} = s_{t+1} | X_{t=s_t}, X_{t-1} = s_{t-1}, \dots, X_0 = s_0) \quad (3.1)$$

Markov chains are assumed to be time independent or constant in time as shown in the following equation:

$$p_{ij} \coloneqq P(X_{t+1} = i | X_t = j) = P(X_t = i | X_{t-1} = j) \; \forall t \in T, \forall i, j \in S \quad (3.2)$$

Meaning that the transition probabilities between the different states do not vary over time. The stochastical transition matrix M is represented as follows:

$$M = (p_{ij}), p_{i,j} \geq 0 \; \forall i, j \in S \; and \; \sum_{j \in S} p_{ij} = 1, (i \in S) \tag{3.3}$$

The vector represented in (3.4) represents the initial distribution, since the variables are stochastic, the distribution can be calculated as in (3.5).

$$\pi = (\pi_i \in S), \text{with } \pi_i = P(X_0 = i) \tag{3.4}$$

$$P(X_0 = s_0, \ldots, X_t = s_t) = \pi_{s_0} p_{s_0 s_1} p_{s_1 s_2} \ldots p_{s_{t-1} s_t} \tag{3.5}$$

Now the probability of getting from state i to state j with m steps is given as follows:

$$p_{ij}^m \coloneqq P(X_{t+m} = j | X_t = i) \tag{3.6.}$$

The m^{th} power of the transition stochastic matrix can be computed as follows:

$$p_{ij}^m = M^m(i, j) \tag{3.7}$$

Rephrasing the equations given, a time independent Markov chain containing the set of states S, the stochastic transition matrix M and the vector of the initial distribution π can be determined as:

$$\theta = (S, M, \pi) \tag{3.8}$$

An example of Markov chains is shown in figure 3.1., this figure represents the states of a Markov chain, where $s0$ is referred to the starting state, and the conditional probabilities are explained next to the edges.

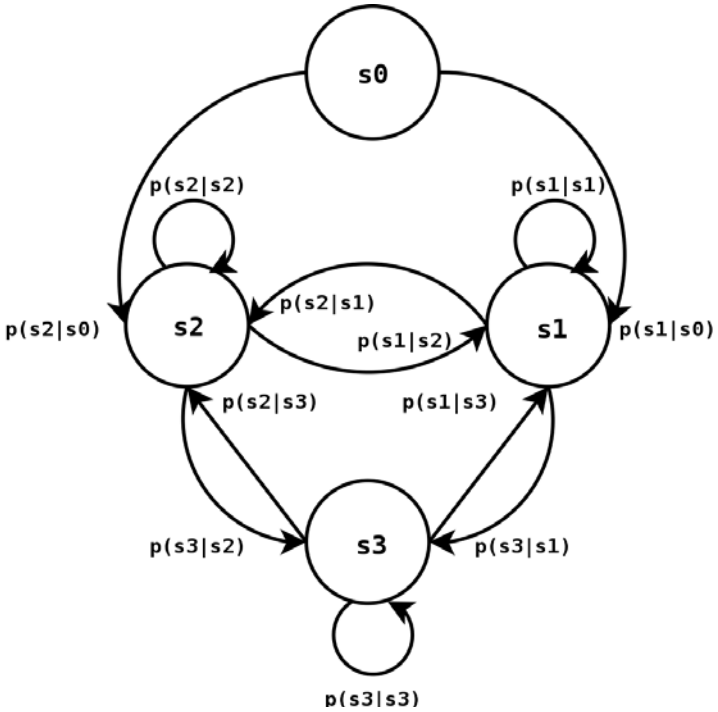


Figure 3.1. Markov chain illustrated by a directed chart.

To understand the process of Markov chains clearly, we can refer to an example presented by Christian in 2006, let S denote a set containing different weather conditions:

$$S = \{sunny, cloudy, rainy\}$$

These weather conditions can be represented in a stochastic transition matrix M illustrated in (3.9), stochastic states that the total sum of the row data is 1, where the data refer to the possibility of changes in the weather.

$$\begin{pmatrix} & sunny & cloudy & rainy \\ sunny & 0.1 & 0.2 & 0.7 \\ cloudy & 0.2 & 0.2 & 0.6 \\ rainy & 0.1 & 0.1 & 0.8 \end{pmatrix} \tag{3.9}$$

For example, according to the stated matrix the probability of the weather being sunny after it being rainy is 0.1% accordingly the probability of the weather being rainy in the following day is 0.8%. The initial distribution vector of $\pi = (P(sunny), P(cloudy), P(rainy)$ can be computed as follows:

$$\pi = (0.2,0.3,0.5)$$

In order to define HMM, it is realized that the state S from the Markov chain (3.8) is considered not directly observed at time t, as a result $s(t)$ is hidden, in other words a certain probability assigned to a symbol υ is emitted by the system, however this symbol can be observed which makes $\upsilon(t)$ visible. The probability for a kind f emission only depends on the state s at time t, therefore the conditional probability can be defined as $p(v(t)|s(t))$. According to these properties an HMM can be defined as follows:

$$\vartheta = (S, M, \Sigma, \delta, \pi) \tag{3.10}$$

Along with the definition the S, M, π were defined earlier, Σ is referred to the emission set of symbols, δ is denoted in a time independent matrix where each data entry represents the chance or the probability for a certain emission, and when the mentioned set is continuous then, probability density function (PDF) model these probabilities an example would be Gaussian distribution. And similar to Markov chains regardless of what nature they are, the total sum of all probabilities equals 1. An example of HMM is illustrated in Figure 3.2.

Let Mike Be a computer engineer who lives in a house somewhere in Adana, Turkey, that hasn't any direct contact or connections to the world outside, lets also

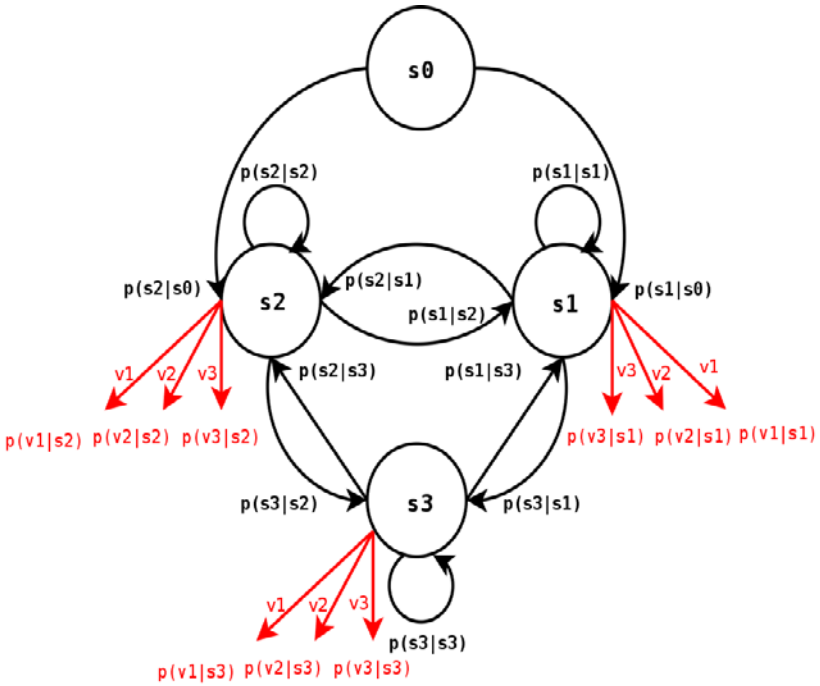


Figure 3.2. An illustration of an HMM

Assume that Mike is interested in the weather conditions of the outside world. Weather conditions can be through Markov chains over time as follows:

$$\theta_{Ad} = (S, M_{Adana}, \pi) \tag{3.11}$$

Due to the fact that his house hasn't any possibilities of observing the condition of the current weather, his only possibility to get an idea about the weather outside, he observes his cat Kitty and that is depending on the cat's fur, so when it goes outside and comes right back, its fur is one of the two states (conditions) as follows:

$$\Sigma = \{wet, dry\} \tag{3.12}$$

These two states or emissions can be observed by Mike, due to the fact the his cat Kitty is going to be in of of these states, that depends on how the weather is for sure. Therefore the emission or state probabilities can are obvious to him, and these emissions can be presented in a stochastical matrix as shown above:

$$\delta = \begin{pmatrix} & dry & wet \\ sunny & 0.7 & 0.3 \\ cloudy & 0.5 & 0.5 \\ rainy & 0.1 & 0.9 \end{pmatrix} \quad (3.13)$$

Along with the Markov chain presented earlier, the HMM model for Adana is:

$$\theta_{Ad} = (S, M_{Adana}, \Sigma, \delta, \pi) \quad (3.14)$$

Using this distinct HMM Mike can predict the weather condition in the last couple of days. Based on that, he can produce a set of sequences, depending on the weather observations, thereafter determining the maximum likelihood sequence of the weather conditions or states leading to that sequence. This study is also known as the Decoding Problem which is considered to be one of the standard or basic problems of the HMMs.

There are three basic problems presented by L.R.Raibiner in 1986, that are formulated for Hidden Markov Models, before presenting these issues, notations of an HMM are defined above:

$A =$ the distribution probability of the state transition

$B =$ the distribution probability of the type observed in some state i.e. j

$M =$ the number of observation states

$N =$ the number of states in a given model

$O =$ the observation sequence

$Q =$ the sstates, i.e.colors, weather conditions.

$T =$ the total length of the observation string (sequence)

$V =$ a discrete set of symbol possible observations

$\pi =$ the initial distribution of the state

λ= represents a specific HMM

The three problems and their solutions that should be explained in order for a model to be beneficial in real world applications are as follows:

- Problem 1: namely the evaluation problem, which implies that; given the emission (observation) sequence $O = O_1, O_2, \dots O_T$, and the model $\lambda = (A, B, \pi)$, how can we determine the probability that the observed emission was originated by the given model, so in order to define the solution, the model which best matches the emissions is to be chosen. An efficient way to deal with this problem would be the Forward algorithm, which computes all possible state sequences of a certain length T, thereafter, we can obtain he probability of O over all probable state sequences through the summation of the joint probability.
- Problem 2: namely the estimation or the decoding problem where the attemption of uncovering the hidden part of the HMM model is required in other words, choosing the $I = \mathrm{i}_1, \mathrm{i}_2, \dots, \mathrm{i}_T$ optimal state sequence, many ways to solve such problem exists, for instance, one way could be to choose the maximum likelihood of the states, which maximizes the number expected among other individual states, this solution is done by the decoding algorithm.
- Problem 3: namely the training or the learning problem, this problem focuses on maximizing the probability of the emission sequence $\Pr(O|\lambda)$, given a specific model, and that is by how to modify the mode parameters (A, B, π), this problem is considered to be the most difficult and complicated among others due to the fact the there is no direct way of solving this problem, many solutions for solving this problem exist, in this study the Baum-Welch algorithm is applied to solve this problem. This algorithm has the capability of solving the learning problem by repeatedly learning the values of parameters A, B of the model and that is done from a group of training samples.

3.1.4. Hidden Markov Model ToolKit

In this study, the main idea is to build a noise robust speaker recognizer system using HMM by implementing the HTK open source toolkit in a Linux environment. Speaker recognition was implemented using HTK version 3.4.1. The Linux operating system used is OpenSuse version 13.2, using it HTK was performed for developing the method. Along with these c ++ platform, audacity and MATLAB are used for building the program.

For creating speaker recognition systems, the HTK based on HMM is used as a software toolkit. This software or toolkit was first produced by the Cambridge University Speech Group. It has been developed and additional features were added in the early nineties.

The HTK is developed in order to be flexible enough to support both research and development of HMM systems. By using HTK tools A speaker recognition system can be examined, applied and then its results can be evaluated. Many projects can be achieved, including various kinds of speaker recognition systems, i.e. speaker verification in this study's case. HTK involve many tools that perform functions like coding data, several kinds of HMM training, including Baum-Welch re-estimation which is used in this study.

There are two main stages concerned with the HTK, the first is the training stage, in which parameters of a set of HMMs are estimated using training tools that are provided by the toolkit, this estimation is done using training utterances. The second stage is the recognition stage, in which unknown utterances are interpreted using the recognition tools that HTK provides. HTK toolkit also includes other two stages, first one is the data preparation stage, which includes converting the data files into a feature format in order for the system to implement, and the final stage, is the analysis stage, in this stage the final results (recognition results) are compared with the original results, and this means evaluating the performance of the system after it is built. (http://www.ee.columbia.edu/ln/LabROSA/doc/HTKBook21/node142.html). The processing stages of the HTK toolkit are shown in details in fig. 3.4.

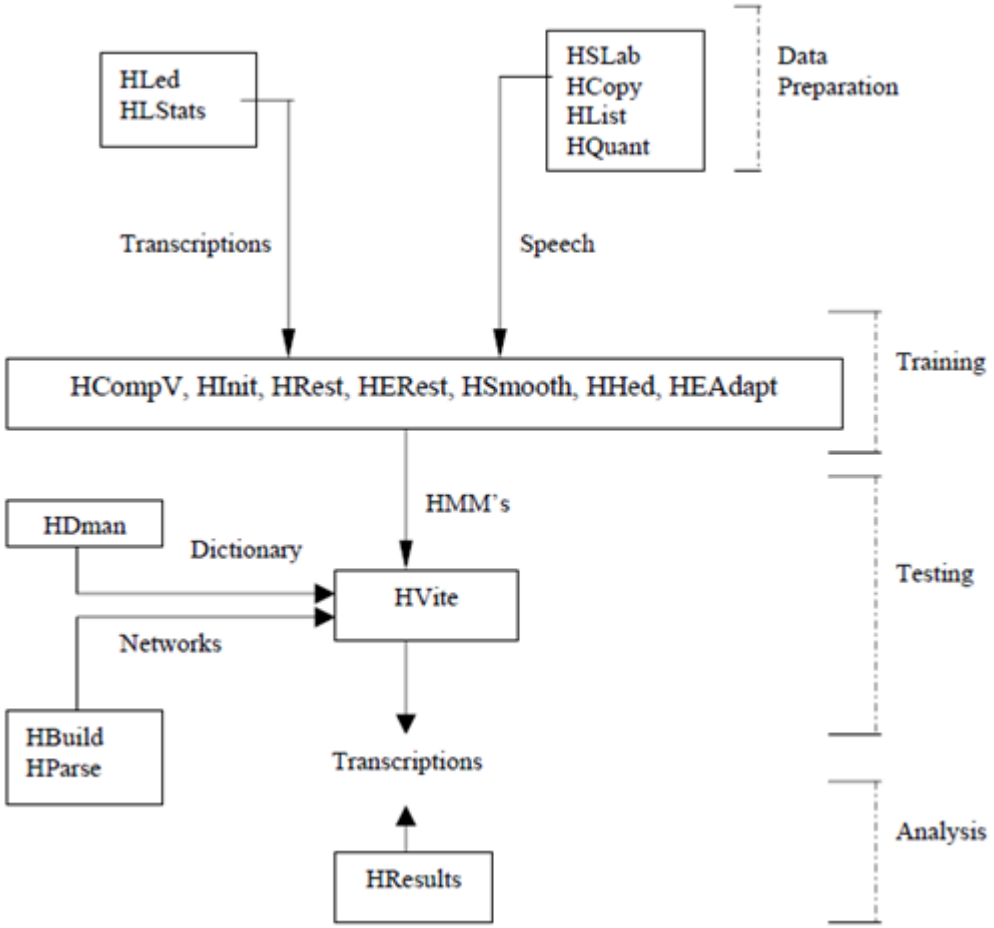


Figure 3.3. HTK Processing Stages

3.1.5. Feature Extraction Using MFFCs

Many feature extraction methods exist for obtaining the specific characteristics of the speech produced, these can be classified into physical and learned. Physical characteristics represent the sizes, i.e. the vocal tract and shapes of speech. These features should be robust under noisy conditions and they should have large variability, in addition to that the feature dimension should also be low, otherwise it cannot be handled by statistical models like GMM (Beigi, H. 2011). The features for speaker verification or recognition can be categorized into, high level features, prosodic features, voice source features, spectral-temporal features and short term spectral features, a detailed overview of these features are illustrated in fig 3.5.

Since short term spectral features are the easiest among others, it is commonly applied in speaker recognition. Using these features systems can achieve more precise recognition outcomes.

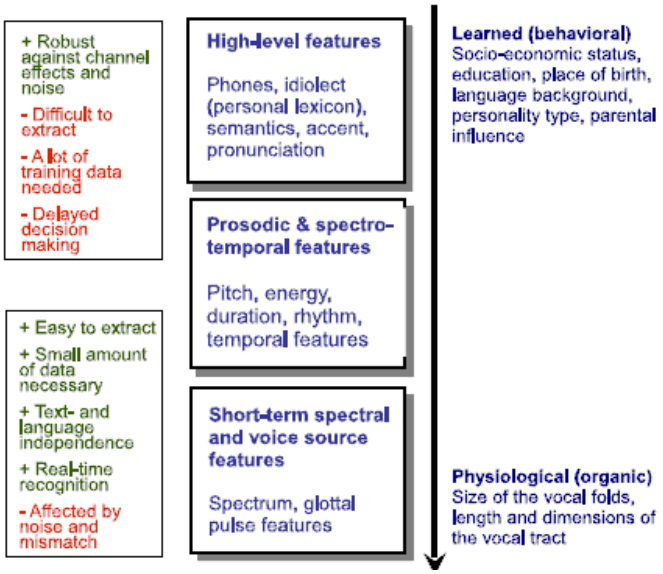


Figure 3.4. Overview of speaker recognition features

Information about the spectral envelope of the speech signal are conveyed using many features (short term specifically) such as LPCC (Linear predictive Cepstral Coefficients), MFDWC (Mel-Frequency Discrete Wavelet Coefficient), and MFCC (Mel-Frequency Cepstral Coefficients). The latter is used in this study.

MFCC features are commonly used in speaker recognition. These features combine the conceptual frequency based scale with the cepstrum analysis, thenceforth, produces a better performance in terms of recognition. Since MFCC cannot discriminate frequencies which are over 1000Hz. Overall MFCC includes two types of filter, one which is at low frequency and other above 1Khz which is referred to logarithmic spacing, a detailed process of the MFCC is illustrated above. Before MFCC is breifly discussed, it should be noticed that the importance of using feature extraction specificly MFCC, is due to the fact the waveforms cannot be processed directly using speaker recognition tools.

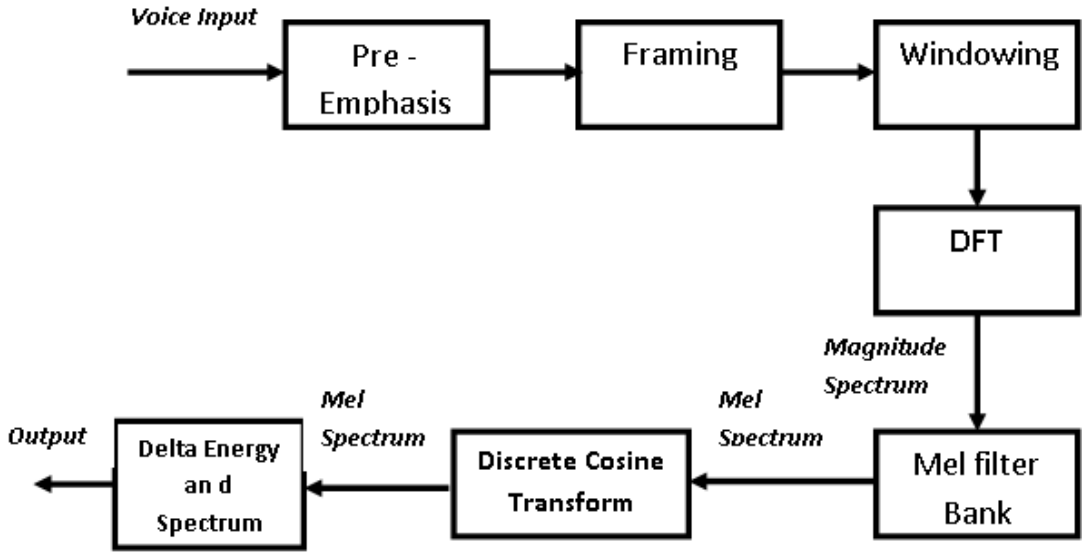


Fig 3.5. MFCC process Block Diagram

As it is obvious from the figure, there are seven steps for computing the MFCC, the following stages are as follows:

Step-1. Pre-Emphasizing the signal

Step-2. Framing the signal

Step-3. Hamming windowing of each frame

Step-4. Taking Fast Fourier Transform and Discrete Fourier Transform

Step-5. Processing Mel Filter Bank

Step-6. Applying Discrete cosine Transform

Step-7. Delta Energy and Delta Specturm

3.1.5.1. Pre-Emphasis

This is done by passing a signal through a filter such as a high pass filter,this allows higher frequencies to be emphasized and makes the signal smooth in terms of spectrum, therefore it helps reduce the noise. For instance Say b = 0.95 It means that 95% of each one sample is orginated from its previous sample.

$$\mathrm{Y}[n] = X[n] - 0.95\ X[n-1] \tag{3.15}$$

The aim of this process is to increase the quantity of energy in the high frequencies (Pandit and Bhatt, 2014).

3.1.5.2. Framing

Framing process is to segment the speech samples or signal into number of frames, generally, a small frame. The size of each frame is considered to be necessary, because the shorter the frames the fewer samples, while the longer the frames the more the signal will have large differences. The ideal frame size of a signal is 10 ms and 40 ms, its also segmented in frames overlapping with each other (M. Dua et. a1, 2012). Each frame overlaps on its next frame. So each frame is devided into N samples out of which M samples are overlapping with the next frame. Normally $M < N$ and usually the typicall values taken are N=256 M=100 but we take the value N =400 and M =160 i.e N =25ms M =10ms for framing (Pandit and Bhatt, 2014).

3.1.5.3. Windowing

After framing the signal, it it should be realized that the speech signal being processed contains unnecessary distortion. Therefore windowing will integrate the signal to all closest frequency lines. Hamming window is common to handle such issue. In this study Hamming Window is implemented for windowing (M. Dua et. a1, 2012). A windowing function is multiplied by each frame, let $x(n)$ denote the input signal, $w(n)$, $0 \leq n \leq N-1$ the window function and $y(n)$ the output signal (L.Muda, et.al., 2010).

The windowed signal as a result gives the following ;

$$Y(n) = X(n) \times W(n) \tag{3.16}$$

The hamming window as given as follows;

$$W(n) = 0.54 - 0.46\,cos\left(\frac{2\pi n}{N-1}\right) \quad , 1 \leq n \leq N \tag{3.17}$$

3.1.5.4. Fast and Discrete Fourier Transform

The FFT is used for the convertion between time and frequency domain of each frame having N samples as shown in eq. (3.17) Thereafter DFT of the windowed signal is taken by using FFT algorithm. DFT implies the spectral information of a speech signal. In fact, it gives energy level at different frequencies.

$$x_i(k) = \sum_{n=1}^{N} x_i(n) w(n)\, e^{\frac{-2j\pi kn}{N}}, \quad 1 \leq k \leq \tag{3.18}$$

where and K is the length of the DFT.

3.1.5.5. Mel Filter Bank

After computing the FFT in the previous step we realize that the spectrum is wide and the signal isn't linearly scaled. To overcome this problem Mel Scale is used. The Mel scale's main function is combining the obtained frequency of the speech signal with the current measured frequency. Any given Frequency f can be converted to Mel scale using *mel* (*f*) as shown above.

$$mel\,(f) = 2595 log\left(1 + \frac{f}{700}\right) \tag{3.19}$$

Now a filter bank consisting of 20-40 (typically 26) triangular shaped band pass filter is used on Mel Scale, after converting frequencies to Mel scale.

In this study, the filterbank consists of 26 vectors and a length of 1024. Mostly each vector is considered to be zeros, but for a certain section of the spectrum is a non-zero. The power spectrum is multiplied with each filterbank to obtain filterbank energies, then coefficents are added up . Once this is done the remaining 26 numbers indicates how much energy was left in each filterbank.

The first filter starts at first point with value zero. It takes its maximum value one at second point. Finally it comes back to zero at third point. The second filter starts at that point at which the first filter has its maximum value. Second filter takes its maximum value at third point and comes back to zero at fourth point and so on. The triangular filters used here are defined as;

$$z_{m_f}(k) = \begin{cases} \dfrac{k - f(m_f - 1)}{f(m_f) - f(m_f - 1)} & f(m_f - 1) \le k \le f(m_f) \\ \dfrac{f(m_f + 1) - k}{f(m_f + 1) - f(m_f)} & f(m_f) \le k \le f(m_f + 1) \\ 0 & otherwise \end{cases} \quad (3.20)$$

Where, m_f is the number of filters used and $f(m_f)$ are the $m_f + 2$ Mel spaced filters. (Pandit and Bhatt, 2014).

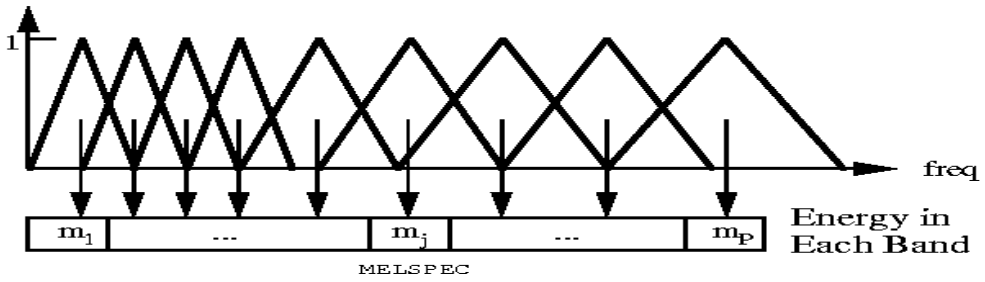


Figure 3.6. Mel-Scale Filter Bank

3.1.5.6. Discrete Cosine Tansform

In this stage the log Mel spectrum taken from the last stage is converted intot the time domain using DCT, This process is done due to the fact that filterbanks are overlapping, so their quite similar to each other. The outcome of this process is called the Mel frequency Cepstrum Coefficient, and they are used as feature vectors. They are obtained using,

$$\tilde{c}_{n_c} = \sum_{k=1}^{m_f} (\log \tilde{p}_k) \cos\left\{ n_c \left(k - \frac{1}{2} \right) \frac{\pi}{2} \right\} \quad (3.21)$$

Here n is the number of Cepstral Coefficients in each frame and m_f is the number of filters in each frame.

(http://practicalcryptography.com/miscellaneous/machine-learning/guide-mel-frequency-cepstral-coefficients-mfccs/)

Computed using the same method, however, they are not computed from the static coefficients but computed from the deltas.

3.1.5.7. Delta Energy and Delta Spectrum

This process is the final step, basically the voice signals and their frames changes, accordingly cepstral features change over time, thence adding features is required. 12 delta features and one energy feature are added, also 39 acceleration or double delta are added. The energy in any frame for some signal say x in a window from time samples ($t1$ to $t2$), is illustrated in the following equation:

$$Energy = \sum X^2 [t] \quad (3.22)$$

3.1.6. Parallel Model Compensation

Several approaches were used to achieve noise robustness, most of these approaches focused on two major methods , the first method implies that the corrupted wavform can be preprocessed preceding the pattern matching step in order to improve the SNR (Signal to Noise Ratio) , and the second method attempt to adjust the pattern matching step to record the noise effects.

This study mainly focuses on the second method mentioned, it shows good results toward noise robustness. It assumes that the observation of the noise and the speech should be abused to obtain the best outcome. This means that noise

compensation should be included in the pattern matching step in which the observed speech that will be recognized, should be fixed in the patterns being stored. The Parallel Model Combination (PMC) method is somehow similar to the HMM decomposition approach, however, it should be noticed that there are two important differences, firstly, it requires the variance states not to be straight (diagnol), and since HMM decomposition operates in the log filter bank domain instead of the cepstral domain which is preferred, it loses the advantages of the cepstral domain regarding compactness and parameter decorrelation. Secondly, due to the fact that the resulted probabilities have to be computed from both the speech distributions and the noise during run-time, it conveys a high computational cost (Gales and young, 1992).

The basic theory of the PMC approach is to estimate the adverse noise given information about both, noise and clean speech, and that is achieved regarding some function. Furthermore, PMC considers that the speech set for recognition is modelled by a given set of density HMMs which are continuous, HMMs have been trained by using clean speech data, furthermore the adverse noise, causing a mismatch between training and testing data patterns, is also modelled by the same set taking into account that it includes a single state, MFCC represents all signals. In order to compensate this occurring mismatch, a mismatch function should be defined, to retrieve such functions, many considerations (Gales and young, 1993) are made:

1. The noise and speech are independent
2. It is assumed that noise and speech are additive in the linear domain, and for noise and speech to be additive at the power spectrum level it is assumed that spectral estimate includes sufficient smoothing.
3. In order for the observation vectors to be represented in the log domain assumptions about Gaussian mixtures including sufficient information should be made.
4. The additive noise does not alter the frame state allocation

The observations provided by the mismatch function mentioned above, are as follows:

$$y_i(t) = O_i^l(t) = F(S_i^l(t), N_i^l(t)) = \log (g \exp (S_i^l(t)) + \exp (N_i^l(t))) \quad (3.23)$$

Where g is the gain matching parameter presented for explaining the level differences between the clean and noisy speech, S(t) is the clean speech and N(t) is the adverse noise. O(t) is the linear domain and $O^l\ (t)$ Is in the Cepstral domain.

Generally, if the corrupted waveform or speech is modelled by a typical HMM, then it is necessary to estimate the mean and covariance of that corrupted speech, in order to obtain the Maximum Likelihood estimation of the noise compensated speech model. However, in case the noise and speech are modelled by different HMMs that are trained on Cepstral feature vectors including Gaussian distributions with the following parameters $\{\mu^c, \Sigma^c\}$ and $\{\tilde{\mu}^c, \tilde{\Sigma}^c\}$ respectively, thus mapping these parameters to the log spectrum domain is necessary. C is a matrix denoting the discrete cosine transform. The speech parameters shown in (3.23) , (3.24) in addition to the noise parameters in (3.25) can be mapped to (3.26).

$$\mu^l = C^{-1}\mu^c \quad (3.24)$$

$$\Sigma = C^{-1}\Sigma^c(C^{-1})^T \quad (3.25)$$

$$\{\tilde{\mu}^c, \tilde{\Sigma}^c\} \quad (3.26)$$

$$\{\tilde{\mu}^l, \tilde{\Sigma}^l\} \quad (3.27)$$

Due to the fact that the linear combination of the random variables considered to be Gaussian distributed, is actually Gaussian distributed itself, no additional considerations are required. Since no close form of the mean $\mu^{\sim l}$ or the covariance $\Sigma^{\sim c}$ compensated exist, a multi- dimensional numerical integration, to obtain the

precise forms of the mean and the covariance is needed which is not recommended. However, if it is considered the sum of two distributed lognormally variables which approximately itself is distributed lognormally then calculating the mean and the variance in the linear spectrum domain, is the only step necessary to be done. Taking account to the previous consideration made, that the noise and speech are additive in the linear spectrum domain and are independent, the following equations are obtained:

$$\hat{\mu} = g\mu + \tilde{\mu} \tag{3.28}$$

$$\hat{\Sigma} = g^2\Sigma + \tilde{\Sigma} \tag{3.29}$$

The parameter μ is denoted as the mean and Σ is denoted as the covariance of the lognormal distribution correlated with the Gaussian distribution $\{\mu^l, \Sigma^l\}$ correspondly with $\{\tilde{\mu}, \tilde{\Sigma}\}$ and $\{\tilde{\mu}^l, \tilde{\Sigma}^l\}$. The parameters of the clean speech and the noise similarly that exist in the log and linear spectrum domains are associated with the equations in (3.29) and (3.30). Since the corrupted speech is considered to be lognormally distributed as metioned above, in the linear spectrum domain, by using the inverse of the previous eguations in (3.29) and (3.30), the required distribution can be acquired in the log spectrum domain, $\{\tilde{\mu}^l, \tilde{\Sigma}^l\}$, as shown in equation (3.31) and (3.32).

$$\mu_i = \exp(\mu_i^l + \Sigma_{ii}^l /2) \tag{3.30}$$

$$\Sigma_{ij} = \mu_i\mu_j \left[\exp\left(\Sigma_{ij}^l\right) - 1\right] \tag{3.31}$$

$$\hat{\mu}_i^l = \log(\hat{\mu}_i) - 1\backslash 2 \log\left(\frac{\hat{\Sigma}_{ii}}{\hat{\mu}_i^2} + 1\right) \tag{3.32}$$

$$\hat{\Sigma}_{ij}^{l} = \log\left(\frac{\hat{\Sigma}_{ij}}{\mu_i \mu_j} + 1\right) \tag{3.33}$$

In case that the Cepstral parameters are going to be used in the recognition step, then the final mapping used is as follows:

$$\mu^{c} = C\mu^{l} \tag{3.34}$$

$$\Sigma^{c} = C\Sigma^{l}C^{T} \tag{3.35}$$

This approach can be considered a similar method to estimate the Gaussian distribution of an observed corrupted waveform or speech $O^{c}(t)$ using Maximum Likelihood estimation, given the Gaussian distribution of both adverse noise and clean speech.

In order to achieve an increased recognition performance, when large databases of speakers to be recognized exist, including dynamic coefficients in the speech parameterization is considered necessary. As mentioned above PMC assumes that noise and speech are additive in the linear spectrum domain, however using dynamic coefficients in this case is not possible , in order to do such implementation a new definition of the mismatch function is required, if the equation in 3.35 as shown below, is used to parameterize the speech

$$O^{\Delta c}(t)^{T} = [O^{c}(t)^{T}, \Delta O^{c}(t)^{T}] \tag{3.36}$$

$\Delta O^{c}(t)$ represents the dynamic coefficient or the delta ccoefficient in its simplest form, furthermore

$$\begin{aligned} \Delta O^{c}(t) &= \left(O^{c}(t+1) - O^{c}(t-1)\right) \\ &= C\left(O^{l}(t+1) - O^{l}(t-1)\right) \\ &= C \log O(t+1)./O(t-1) \end{aligned} \tag{3.37}$$

Where ./ represents the elementwise division. Considering that noise and speech are additive $O(t) = S(t) + N(t)$, substituting this with the equation in 3.36 we get

$$\Delta O^c(t) = C \log S[(t+1) + N(t+1)./S(t-1) + N(t-1)] \tag{3.38}$$

The equation shown above could be represented in the linear domain, in terms the dynamic or delta coefficients of the noise $\Delta N(t)$, and speech $\Delta S(t)$ as shown below:

$$\begin{aligned} \Delta O_i(t) &= \left(\frac{S_i(t+1)}{S_i(t-1) + N_i(t-1)}\right) + \left(\frac{N_i(t+1)}{S_i(t-1) + N_i(t-1)}\right) \\ &= \Delta S_i(t)\left(\frac{\frac{S_i(t-1)}{N_i(t-1)}}{\frac{S_i(t-1)}{N_i(t-1)} + 1}\right) + \Delta N_i(t)\left(\frac{1}{\frac{S_i(t-1)}{N_i(t-1)} + 1}\right) \end{aligned} \tag{3.39}$$

It is noticed from the equation above that, it contradicts the assumption presented by the PMC method, and that is as a result of the delta coefficient at time t relying on the static coefficient at time $t-1$, consequently the speech signal waveform cannot be split with spontaneous transitions into stationary segments between them, unless the segments are long enough, then it is possible to consider the statistics of $S(t-1)$ is approximately similar to $S(t)$ and accordingly $N(t-1)$ to that of $N(t)$, taking all this into consideration, the equation 3.38 is then a suitable mismatch function.

Furthermore, an ML estimation is required to estimate the HMMs for the delta parameters, consequently calculating both the covariance and mean of the signal in the spectrum domain, hereafter the speech should be mapped to the spectrum domain as long as it is parameterized in the Cepstral domain. For speech, we have

$$(\mu^{\Delta l})^T = [(C^{-1}\mu^c)^T, (C^{-1}\Delta\mu^c)^T] \tag{3.40}$$

and

$$\Sigma^{\Delta l} = \begin{bmatrix} C^{-1}\Sigma^{c}(C^{-1})^{T} & C^{-1}\delta\Sigma^{c}(C^{-1})^{T} \\ C^{-1}(\delta\,\Sigma^{c})^{T}(C^{-1})^{T} & C^{-1}\Delta\Sigma^{c}(C^{-1})^{T} \end{bmatrix} \tag{3.41}$$

$\delta\,\Sigma^{c}$ refers to the covariance matrix that represents the correlation between the deltas and the static coefficients. Considering that a single Gaussian mixture will be estimated, the ML estimate for the delta coefficients can be given as follows:

$$\begin{aligned} &\Delta\hat{\mu}_i^l \\ &= \int_{R^n} dS^l \int_{R^n} dN^l \int_{R^n} d\Delta S^l \int_{R^n} d\Delta N^l\, p(S^l, \Delta S^l) p(N^l, \Delta N^l) \log(\gamma_i \exp(\Delta S_i^l) \\ &+ \eta_i \exp(\Delta N_i^l)) \end{aligned} \tag{3.42}$$

and

$$\begin{aligned} &\Delta\hat{\Sigma}_{ij}^l = \int_{R^n} dS^l \int_{R^n} dN^l \int_{R^n} d\Delta S^l \int_{R^n} d\Delta N^l \\ &\left\{p(S^l, \Delta S^l) p(N^l, \Delta N^l) \log\left(\gamma_i \exp\left(\Delta S_i^l\right) + \eta_i \exp\left(\Delta N_j^l\right)\right)\right\} \\ &-\Delta\hat{\mu}_i^l \Delta\hat{\mu}_j^l \end{aligned} \tag{3.43}$$

and

$$\begin{aligned} &\hat{\Sigma}_{ij}^l = \int_{R^n} dS^l \int_{R^n} dN^l \int_{R^n} d\Delta S^l \int_{R^n} d\Delta N^l \\ &\left\{p(S^l, \Delta S^l) p(N^l, \Delta N^l) \log\left(g \exp\left(S_i^l\right)\right.\right. \\ &\qquad \left.\left. + \exp\left(N_i^l\right)\right) \log\left(\gamma_j \exp\left(\Delta S_j^l\right) + \eta_j \exp\left(\Delta N_j^l\right)\right)\right\} \\ &\qquad -\Delta\hat{\mu}_i^l \Delta\hat{\mu}_j^l \end{aligned} \tag{3.44}$$

noting that

$$\gamma_i = \left(\frac{\exp(S_i^l - N_i^l)}{\exp\left(S_i^l - N_i^l\right) + 1} \right) \tag{3.45}$$

and

$$\eta_i = \left(\frac{1}{\exp\left(S_i^l - N_i^l\right) + 1} \right) \tag{2.46}$$

In order to compute the complete forms of equation 3.42 and 3.43 and by considering the variances on γ and η are trivial , thereafter the ML estimates form of the delta parameters are similar to the ones of the static coefficients, thus

$$\Delta\hat{\mu}_i = \bar{\gamma}\Delta\mu_i + \bar{\eta}_i\Delta\tilde{\mu}_i \tag{3.47}$$

$$\Delta\hat{\Sigma}_{ij} = \bar{\gamma}_i\bar{\gamma}_j\Delta\Sigma_{ij} + \bar{\eta}_i\bar{\eta}_j\Delta\tilde{\Sigma}_{ij} \tag{3.48}$$

where

$$\varepsilon[\gamma_i] = \varepsilon\left[\frac{\frac{S_i}{N_i}}{\frac{S_i}{N_i} + 1} \right] \approx \left(\frac{\frac{\mu_i}{\tilde{\mu}_i}}{\frac{\mu_i}{\tilde{\mu}_i} + 1} \right) = \bar{\gamma}_i \tag{3.49}$$

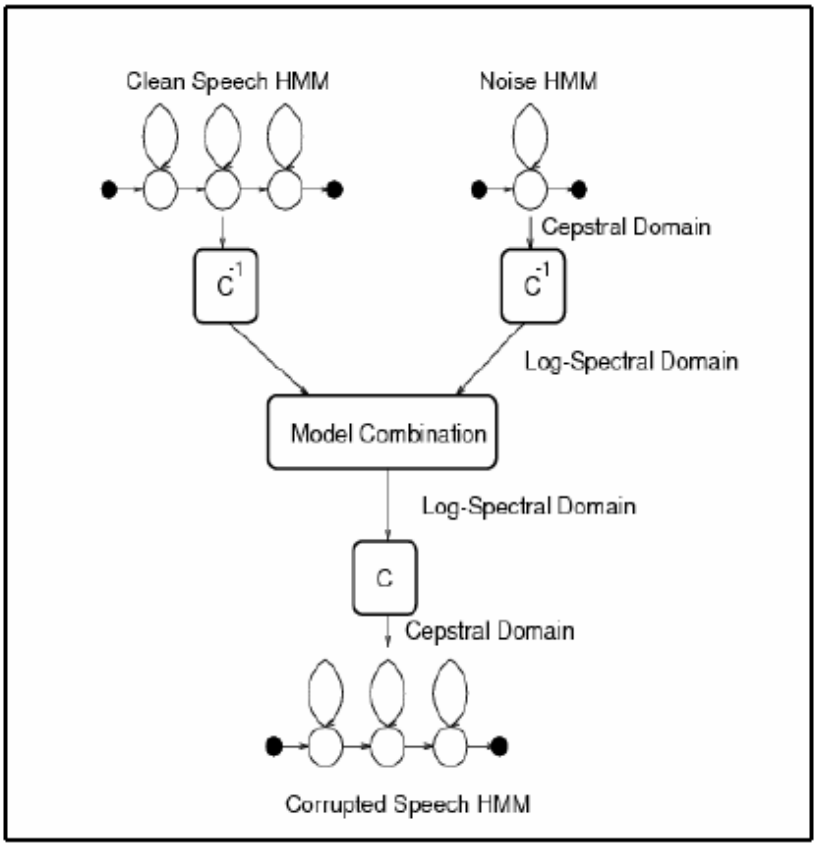
and

$$\varepsilon[\eta_i] = \varepsilon\left[\frac{1}{\frac{S_i}{N_i} + 1} \right] \approx \left(\frac{1}{\frac{\mu_i}{\tilde{\mu}_i} + 1} \right) = \bar{\eta}_i \tag{3.50}$$

Now mapping back the covariance and mean into the Cepstral domain similar to the static coefficients is possible

$$\Delta\mu^c = C\Delta\mu^l \tag{3.51}$$

$$\Delta\Sigma^c = C\Delta\Sigma^l C^T \tag{3.52}$$



3.7. The basic process of the PMC (Gales, 1995).

Overall, PMC among all other model compensation methods, is considered the most popular method in obtaining good recognition models, the basic process of PMC was presented by gales in 1995, is illustrated in figure 3.7, where the first step is transferring the input into the log or linear spectral domain, then combining both the noise and speech model using a mismatch function as mentioned earlier, thereafter the noisy speech signal is converted back to the Cepstral domain to apply the regular recognition process.

3.2. Method

The main goal of this study is to increase the performance of a speech recognition system under noisy environments. Many approaches have been proposed to improve the recognition performance of the speaker recognition system for noisy speech. In the previous researches, pre-knowledge about the noise statistic is assumed. Thus, the noise model is estimated using the noise signal. However, this is not applicable to real world applications due to the lack of the noise signal, in order to estimate the noise model. Consequently, a classifying method to label the given signal as speech and non-speech (i.e. noise) is required. One popular method is the voice activity detection (VAD). This method is well known for classifying the given signal into speech and non-speech segments. Non-speech segments which are classified using a VAD method can be used to estimate the noise model.

In this study, VAD methods are used for estimating the noise model, and PMC is proposed for estimating the noisy speech model given both the clean speech and the noise model, which is estimated using a VAD method. In this study, performances of the baseline and two well-known VAD methods are implemented and compared, for the noisy speaker recognition issue. Noise models were estimated using noise for the baseline method. For all the methods, PMC is used to estimate the noisy speech model.

Figure 3.8 explains the proposed noisy speech recognition system. First, the clean speech model is estimated, next the noise model is estimated using a VAD method under a given noise environment. Thereafter, PMC is used to estimate the noisy speech model. In this study, Noise model is estimated using noise (baseline) , VQ-VAD and MMSE.

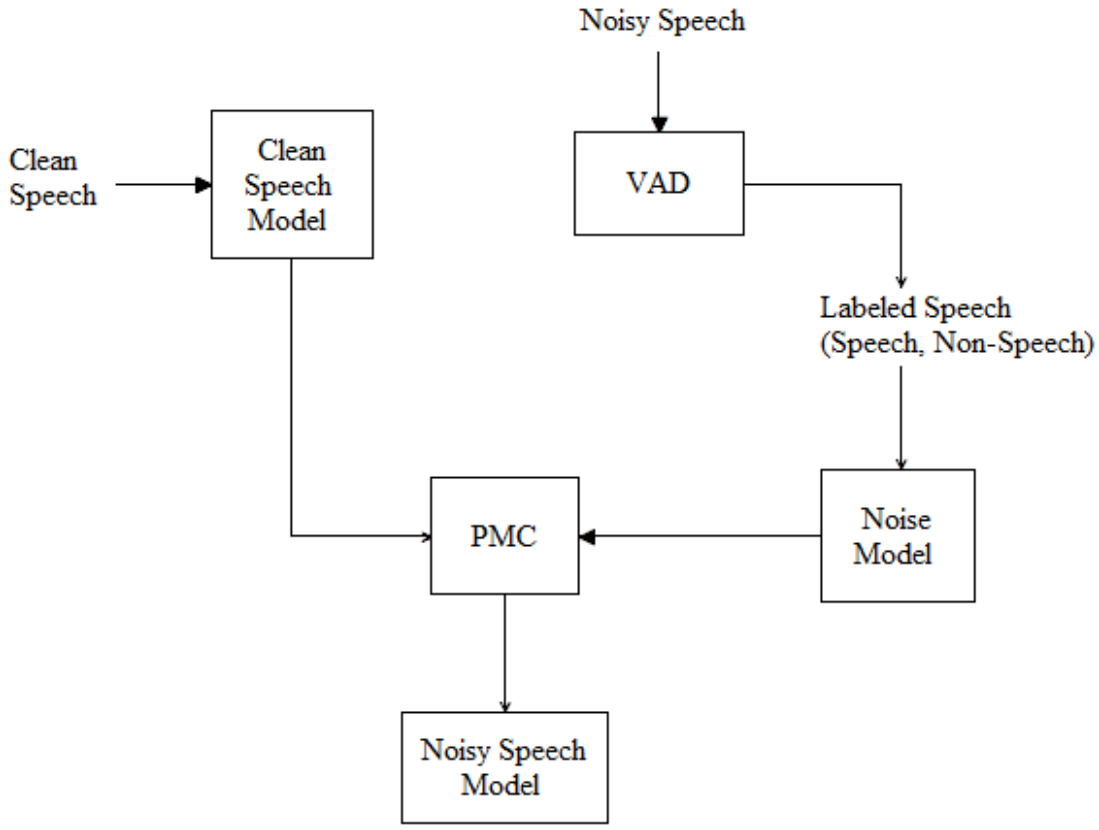


Figure 3.8. Framework of the Noisy Speech Model

3.2.1. VQ-VAD

VAD has a vital influence on robust speaker verification, specifically energy-based VAD. Generally energy-based VAD performs well in noise-free environments, however, its performance degrades in the case of the existence of adverse or unknown noisy environments. According to T.Kinunen and H.Li energy-based VAD, due to its simplicity, it is considered the most popular VAD. Energy-based VAD calculates the energy of each short-frame and correspondingly considers low and high frames, refer to non-speech and speech respectively (T.Kinnunen et al., 2010). The decision made for the energy threshold can be related to average or maximum energy utterance (M.Sahidullah et al., 2012). The threshold could also be adjusted based on GMM parameters which can be fit to the energy distribution.

As mentioned above a disadvantage of the energy-based VAD is the lack of performance under environmental conditions (H.B.Yu et al., 2011), therefore speech

enhancement is necessary especially under low levels of Signal-to-noise ratios (SNRs). According to H.B energy-based VAD with the spectral subtraction enhancement approach, can produce good results. In this research a practical VAD that does not depend on the pre-training step for the acoustic models and that is trained only from a recording, is proposed. This is attained by labeling a small number of accurate training vectors through an initial energy VAD, for the acoustic models, two mixtures are trained one for the non-speech and the other for the speech, thereafter the frames of the utterance are labeled as speech or non speech according to some likelihood ratio indicator. The suggested VAD is designed according to the some principles which are, unsupervised self-adaptive and practical (Kinnunen et al., 2013).

The adaptive energy VAD is described as follows, $x_t[n]$ denotes the n^{th} sample of the t^{th} speech frame in a given speech, first the log energy of every frame is computed by the following equation:

$$E_i = 10 \log_{10}\left(\frac{1}{N-1}\sum_{n=1}^{N}(x_i[n]-\mu_t)^2 + \varepsilon\right) \tag{3.53}$$

$\mu_i = (1/N)\sum_{n=1}^{N} x[n]$ refers to the sample mean of the frame, N is the frame size and $\varepsilon = 10^{-16}$ is a random constant to prevent log of zero. Maximum energy $E_{\max} = \max_{i=1,.....,I}\{E_i\}$ is detected for over all the T frames of the utterance. The VAD decision is a threshold comparison set based on this maximum level. Furthermore, minimum energy limit also applied to prevent utterances with low energy being falsely marked as speech. Hence, the energy VAD basis for speech existence is $(E_i > E_{\max} - \theta_{main}) \wedge (E_i > \theta_{\min})$, where θ_{main} and $\theta_{\min}$, refer to pre-set primary and minimum energy thresholds, respectively. When optimizing the spectral subtraction parameters, these fixed to θ_{main} = 30 dB and $\theta_{\min}$ = −55 dB.

The energy based VAD works well, for high SNRs, however, it usually marks most frames as speech, in low SNRs. Using speech enhancement methods is one strategy to overcome such problem, through increasing SNR. The spectral subtraction method can be such a method, which depends on MATLAB implementation specsub in *Voicebox*[1]. If $|X|^2$ and $|\widehat{N}|^2$, respectively, refer to the powers of noisy speech and estimated noise in a particular time-frequency FFT bin. Spectral subtraction is obtained by multiplying the noisy magnitude $|X|$ by some gain factor g whose common form is taken from (M.Berouti et al., 1979),

$$g = \max\left\{\left(1 - \left(\alpha \frac{|\widehat{N}|^2}{|X|^2}\right)^{\gamma_s/2}\right)^{e/\gamma_s}, \min\left(g_h\left(\beta_s \frac{|\widehat{N}|^2}{|X|^2}\right)^{e/2}\right)\right\} \quad (3.54)$$

where α denotes an over-subtraction factor, γ_s denotes the subtraction domain, e is the gain exponent, g_hIs the maximum gain for noise floor and β_s describes the maximum noise fading in the power domain. The phase of the noisy signal is combined with the gain and multiplied- magnitude followed by overlap-and-add signal reconstruction.

Maximum gain and maximum noise fading are fixed, g_h=1.00 and β=0.01 respectively.

Concerning the subtraction domain, magnitude domain subtraction is used by selecting (γ_s, e) = (1, 1), power domain spectral subtraction by (γ_s, e) = (2, 1) and Wiener filter by (γ_s, e) = (2, 2). In terms of the quantity of subtraction, α varies linearly from $\alpha = \alpha_{max}$ for a frame SNR of -5 dB down to α = 1 for SNR = 20 dB; the maximum over-subtraction factor, α_{max}, is used as a control parameter (Kinnunen and Rajan, 2013), MMSE (T.Gerkmann et al., 2012) and MS (R.Martin., 2001) are alternatives used as a noise estimator or noise trackers.

The Self-Adaptive VAD operates as shown in the pseudocode, presented by T. Kinnunen and P. Rajan, where the MFCCs are extracted from the given (original) signal, then by using aggressive spectral subtraction, the signal is enhanced (H.B.Yu et al., 2011), thereafter energy values are sorted and a fixed percentage is determined from the highest and the lowest frames in order to label speech and non-speech frames respectively, in reliable way, based on that non-speech and speech models are trained through using the MFCCs which corresponds to the frame indices.

Inputs: Speech signal $x[n]$, frame length (N) and hop (M)

Outputs: Binary VAD labels $VAD[k]$, $k = 1, 2, ..., K$

1. Extract MFCCs from the noisy signal

$X \leftarrow \text{ExtractMFCC}(x, N, M, MFCCParams)$;

2. Denoise the speech signal

$x_{clean} \leftarrow$ Specsub(x, SpecsubParams);

3. Compute frame energies of the enhanced signal,

$E \leftarrow \text{ComputeEnergy}(x_{clean,} N, M)$;

4. Find indices of low/high energy frames (fixed percentage)

$[i_{low,} i_{high}] \leftarrow$ FindLowestAndHighest (E , percentage);

5. Train speech and nonspeech models from the frame subsets

$\lambda_{speech} \leftarrow$Train ($\{x_k \in X \mid k \in i_{high}\}$,ModelParams);

$\lambda_{nonspeech} \leftarrow$Train ($\{x_k \in X \mid k \in i_{low}\}$,ModelParams);

6. For all frames, pick the more likely hypothesis

$VAD[t] \leftarrow \{\log p(x_t \mid \lambda_{speech}) \geq \log p(x_k \mid \lambda_{nonspeech})\} \wedge E_k \geq \theta_{\min}$;

With min-energy constraint

All MFCC processing make use of the noisy signal rather than the enhanced one containing spectral subtraction products. Both speech and non-speech models are GMMs of the form $p(x|\lambda) = \sum_{k=1}^{k} P_k N(x|\mu_k, \Sigma_k)$ with covariance matrices Σ_k, mixing the weights of P_k and mean vectors μ_k. Although various number of Gaussians can be implemented for speech models and non-speech models, the same number of Gaussians is used for speech and non-speech models (P.kenny et al., 2010).

Furthermore, it is noted that the mentioned VAD surpass the other VAD approaches and through all SNR levels, and spectral subtraction is only used to enhance the energy.

3.2.2. Minimum Mean Square Error Based VAD

ASR(Automatic Speech recognition) and speech coding which are examples of speech and speaker processing systems, are specifically designed for taking speech signals as an input, however many situations that require practical implementations, noisy backgrounds corrupt these speech signals, accordingly the resulted noisy signals are adverse to speech recognition systems, therefore in order to increase the recognition process, speech enhancement algorithms are required. Enhancement algorithms mainly rely on the presence of a robust VAD, the usage of a good VAD even without the presence of any kind of algorithm, can obtain remarkable results, especially in Automatic Speaker Recognition (ASR). VADs can be used as log-likelihood test model to evaluate likelihood of the speech absence vs. speech presence (R. J. McAulay et al., 1980) or it can be used as an accurate computation model of speech presence probability (Y. Ephraim et al, 1984), and it should be noted that most VAD algorithms depend on SNR (J. Sohn et al., 1999), hence estimating the noise power in each frame accurately is critical (Y. D. Cho et al., 2001).

Mainly, most speech recognition systems rely on the MMSE estimation of the noise power spectrum, i.e., the estimation of the expected value in every bin of the STFT (short time Fourier transform). The conventional noise power may be updated

in a recursive way (Sohn et al., 1998), or estimated during the first frames recorded assuming that these frames contain no speech.

The more noise power is higher the more its corresponding spectrum will have high variance, accordingly, although the noise is stationary, the estimation of that noise spectrum of the existing observation will not be approximate to the MMSE estimation of the noise spectrum. High levels of noise corrupt noise adaptation approaches, since they only depend on the probability of speech presence, and in the case of low SNR levels, such probabilities lack accuracy, hence speech enhancement methods and VAD techniques that show good results in high SNR levels, might possibly not perform well in low SNR levels (B. Lee et al., 2007).

This approach suggests a combination of a present noisy observation with an MMSE a posteriori noise estimation built on a priori noise estimation, engaging the uncertainty of speech presence, this approach does not depend on the assumption that noise spectrum can be predicted, hence it is not similar to any adaptation methods.

The statistical noise model of this method considers the input signal say x consists of stationary noise only say n, and it is considered that the noise is an arbitrary process having zero mean and an unknown pdf (probability density function). Hence the STFT of x may be given as follows

$$X\,{}_{k}^{m} = \sum_{n=0}^{L-1} x[n + mL] e^{-j\frac{2\pi kn}{N}} \tag{3.55}$$

The STFT coefficients asymptotically have a Gaussian pdf with zero mean, if $L \to \infty$, and assuming that the coefficients are a weighted sum of sthe samples of the arbitrary process mentioned earlier, hence the pdf of the k^{th} frequency bin for the $X\,{}_{k}^{m}$ coefficients, can be presented as follows

$$p(X_k^m) = \frac{1}{\pi \lambda_N(k)} \exp\left\{-\frac{|X_k^m|^2}{\lambda_N(k)}\right\} \tag{3.56}$$

$\lambda_N(k) = E[|X_k^m|^2]$ refers to the variance of the noise, hence the DFT noise variance $\lambda_N(k)$ is corresponding to the MMSE estimation of the noise power, if $|X_k^m|^2$ denoting the spectral component has and exponential pdf with $\lambda_N(k)$ denoting the mean.

The VAD's function is to compare the probability of the following assumption:

$$\begin{cases} H_0: X_k^m = N_k^m, & speech\ absence \\ H_1: X_k^m = S_k^m + N_k^m, & speech\ presence \end{cases} \quad (3.57)$$

N_k^m, X_k^m and S_k^m are referred to STFT vectors of noise, noisy speech and speech respectively, and are of K- dimension. (Lee and Johnson, 2007).

Error propagation is a priority problem , in high-noise environments including low SNR,. Error propagation can be over comed by using a specific quantity of prior information to the problem. For instance, if the noise process is determined as stationary, and if the first M frames of the signal are known to include non-speech, then an a priori periodogram estimate $\lambda_N(k)$ of $E\left[|N_k^m|^2\right]$ with known standard error may be calculated. If we consider no further data about $E\left[|N_k^m|^2\right]$ intervening frames are provided then

$$E\left[|N_k^m|^2 H_1\right] = \hat{\lambda}_N(k) \text{ and } \hat{\lambda}_N^m(k) = \beta_k^m \overline{\lambda}_N(k) + (1-\beta_k^m)|X_k^m|^2 \quad . (3.58)$$

If it is quite likely that the speech is present i.e., $\wedge_k^m >> 1$, so $\beta_k^m \approx 1$, then adjusts $\hat{\lambda}_N^m(k)$ to the mean estimate of the noise spectrum. Hence the described method is an example to false-positive errors, such as the autoregressive estimator, and it does not propagate error. However, same processes can be applied to false-positive frame just like any other frame about which no specific information of the

noise spectrum as provided: the noise estimate is returned to the a priori noise estimator $\hat{\lambda}_N(k)$.

A posteriori MMSE estimate of the noise power can evaluate this noise spectrum estimation method in the recent frame, when the noise process is stationary regarding high variance. According to experiments this noise estimation approach gives higher accuracy specifically for low-SNR situations (Lee and Johnson, 2007).

4. RESEARCH AND DISCUSSION

In this section, experimental and implementation setups along with their outcomes are presented. NIST 1998 and NOISEX-92 databases were used for evaluating the MFCCs using parallel model combination approach. As mentioned earlier in this study, the NIST 1998 database contains conversational data for 500 speakers, 250 male and 250 female speaker, these data or signals are sampled at 8kHz. In this study only male speakers were used for evaluation, furthermore there are three different kinds of training sessions (as mentioned earlier), this study focuses on the two- session training condition, thereafter, test data with a duration of 30 seconds using the same type of handset and gathered from the same phone number, were uses in this implementation, regarding each test file nine trails for non speaker targets versus one trail for the target speakers exist, accordingly, the total numbers exisiting for the trails is 13080.The NOISEX-92 database is normally sampled at 16 kHz, however, in this study, noises were downsampled to 8kHz in order to have the exact sampling rate of the NIST database, speech STITEL and F16 were used and added to the test speaker signals as background noise in order to attain the noisy speech signal at SNR levels of -06dB, 0dB, +06dB.

The speech signal was sampled at 8 kHz, and analyzed with 25 ms hamming windows every 10 ms. Furthermore, there were 13 MFCCs and 13 delta MFCCs for each feature vector extracted from each frame. 64 mixtures for modelling each speaker were used, additionally, 3 Gaussian mixture models were used for estimating the noise.

This implementation can be divided into three main sections, the first one is the process of training the clean speech to obtain the estimated clean speech models, before artificially adding the background noise, thereafter the second process is training the noise model,using the noise for the baseline and using the estimated noise for the VAD techniques. The final section was using PMC to estimate the noisy speech model.

In this section two VAD techniques were used, first one the VQ-VAD mentioned in the previous section, using its own proposed open source program to

generate the noise models using the noisy speech signal and the same process applies for the second VAD technique namely MMSE. The speaker recognition system was built and tested using a Hidden Markov Model tool kit HTK. The results are explained and illustrated in the following tables

Table 4.1. Comparison of equal error rates of the speech noise for all SNR levels.

SPEECH NOISE	EER (Equal Error Rate)			
	SNR (dB)	MMSE	VQ-VAD	BASELINE METHOD
	-6 dB	25.99	24.46	19.26
	-0 dB	15.21	15.59	12.30
	6 dB	9.86	9.25	8.94

Table 4.2. Comparison of equal error rates of STITEL noise for all SNR levels.

STITEL	EER (Equal Error Rate)			
	SNR (dB)	MMSE	VQ-VAD	BASE-LINE METHOD
	-6 dB	18.42	16.20	14.98
	-0 dB	11.39	10.16	10.47
	6 dB	8.18	8.40	8.10

Table 4.2. Comparison of equal error rates of F16 noise for all SNR levels.

F16	EER (Equal Error Rate)			
	SNR (dB)	MMSE	VQ-VAD	BASE-LINE METHOD
	-6 dB	31.72	30.12	23.77
	-0 dB	18.65	18.42	14.98
	6 dB	11.92	11.16	10.16

It is noted in all three tables, that, for high SNR levels, the EER results of the implemented VAD techniques were similar to the baseline method. However, for low SNR levels (-6 dB), the performance of the baseline method was significantly better than the applied VAD techniques.

By analyzing the results, it is noted that the implemented VAD techniques are good in estimating the noise for high SNR levels, but not good in the case of low SNR levels. Furthermore, VQ-VAD method slightly outperforms the MMSE method almost for all SNR cases.

5. CONCLUSION

Many researches assumed that the noise signal is known in advance, hence they estimated the noise model from the noise signal based on that pre-knowledge. However in this study a new approach is proposed.

Noise is estimated from the noisy speech using VAD techniques, then the estimated noise was used to estimate the noise model. Furthermore, the estimated noise model along with the speech model were combined using PMC to estimate the noisy speech model. This method is more practical in real-world applications, due to the fact that the noise model was estimated directly from the noisy speech.

Several VAD approaches were implemented for speaker verification in the presence of different noises which are Speech, Stitel and F16 noises and compared at three SNR levels -6dB ,0dB and +6dB, these techniques were also utilized to detect both speech and non speech frames.

The VAD techniques estimated the noise models from the noisy speech signal, thereafter, the results were compared with the baseline approach, the baseline approach yielded better performance than the other two VAD-based techniques as expected, namely, VQ-VAD and MMSE-VAD, due to the fact that the baseline method estimated the noise models from the noise signal rather than from the noisy speech.

Based on the results, the performance of the VAD techniques implemented was similar to one another for all SNR levels as mentioned earlier, however VQ-VAD technique shows a slightly better performance than the MMSE.

Overall, the EER results for the VAD based methods are not lower than the baseline method, so as a future research topic could be, exploring or adjusting more reliable noise estimation approaches to achieve optimal results for the speaker verification under adverse conditions case.

REFERENCES

ATAL, B. (1976). Automatic recognition of speakers from their voices. Proc. IEEE, vol. 64, no. 4, pp. 460-475.

BENYASSINE, A., SHLOMOT, E., SU, H.,-Y., and YUEN, E., 1997. A robust low complexity voice activity detection algorithm for speech communication systems. In Speech Coding For Telecommunications Proceeding. 1997 IEEE Workshop on, p. 97.

BIMBOT, F., BONASTRE, J. F., FREDOUILLE, C., GRAVIER, G., CHAGNOULLEAU, I. M., MEIGNIER, S., … GARCIA, J. O. (2004). A Tutorial on Text-Independent Speaker Verification. EURASIP J. Adv. Sig. Proc., vol. 4, pp. 430-451.

BOLL, S., F., 1979. Suppression of acoustic noise in speech using spectral subtraction. IEEE Transactions on Acoustics, Speech and Signal Processing, vol. 27, no. 2, pp. 113–120.

CAMPBELL JR, J. P. 1997. Speaker recognition: a tutorial, IEEE, vol.85, no.9, pp. 210-229.

CHEN, B., ZHU, Q., and MORGAN, N., 2004. Learning long-term temporal features in LVCSR using neural networks. In Proc. ICSLP, pp. 612–615, Citeseer.

CHO, N., and KIM, E.,-K., 2011. Enhanced voice activity detection using acoustic event detection and classification. Consumer Electronics, IEEE Transactions on, vol. 57, no. 1, p. 196.

COOKE, M., GREEN, P., JOSIFOVSKI, L., and VIZINHO, A., 2001. Robust automatic speech recognition with missing and unreliable acoustic data. Speech Communication, 34(3):267{285.

EPHRAIM, Y., and MALAH, D., 1984. Speech enhancement using a minimum mean-square error short-time spectral amplitude estimator. IEEE Transactions on Acoustics, Speech, and Signal Processing, vol. 32, no. 6, pp. 1109–1121.

EPHRAIM, Y., 1992. A bayesian estimation approach for speech enhancement using hidden Markov models. IEEE Trans, on Acoustic, Speech and Signal Processing, 40, 4, 1992, 725-735.

FLORES, N., J., A., and YOUNG, S., J., 1993. CSS-PMC: A combined enhancement/compensation scheme for continuous speech recognition in noise, Cambridge University Engineering Department. Technical Report CUED/F-INFENG/TR.128.

GAIKWAD, S., GAWALI, B. W., & YANNAWAR, P. 2010. A review on speech recognition technique. , pp. 16-24.

GALES, M., J., F., and YOUNG, S., J., 1992. An improved approach to hidden Markov model decomposition of sleep and noise. In Proc. ICASSP.

GALES, M., J., F., and YOUNG, S., J., 1996. Robust continuous speech recognition using parallel model combination. IEEE Trans. on Speech and Audio Processing.

GALES, M., J., F., 1998. Predicative model based compensation schemes for robust speech recognition. Speech Communication, 25.

GALES, M., F., J., 1997. Nice model-based compensation schemes for robust speech recognition. Proceedings of ESCA-NATO Tutorial and Research Workshop on Robust Speech Recognition for Unknown Communication Chanels.

GALES, M., F., J., and YOUNG, S., J., 1993. HMM recognition in noise using parallel modelcombination. Proceedings of EuroSpeech.

GAUVAIN, J., L., and LEE, C., H., 1994. Maximum a posteriori estimation for multivariate Gaussian mixture observations of Markov chains. IEEE Trans. on Speech and Audio Processing, pp. 291-298, Vol. 2.

GHOSH, P., TSIARTAS, A., and NARAYANAN, S., 2011. Robust voice activity detection using long-term signal variability. Audio, Speech, and Language Processing, IEEE Trans- actions on, vol. 19, no. 3, pp. 600–613.

GONG, Y., 1995. Speech recognition in noisy environments: A survey", Speech communication. pp. 261-291, Vol. 16.

HSIEH, C., -H., FENG, T., -Y., and HUANG, P., -C., 2009. Energy-based VAD with grey magnitude spectral subtraction. Speech Communication, vol. 51, no. 9, pp. 810–819.

HTK, http://www.ee.columbia.edu/～/node142.html

VAN HAMME, H., 2003. Robust speech recognition using missing feature theory in the cepstral or LDA domain. Proceedings of EuroSpeech, Geneva, Switzerland, pp. 3089-3092.

HUO, Q., CHAN, C., and LEE, C., H., 1995. Bayesian adaptive learning of the parameters of hidden Markov model for speech recgonition", IEEE Trans. on Speech and Audio Processing, pp. 334-345, Vol. 3.

IVAN, J., T., 2009. Sound Capture and Processing: Practical Approaches. Wiley.

KHOA, P., C., 2012. Noise Robust Voice Activity Detection.Ms. Thesis, 2012.

KINNUNEN, T., and RAJAN, P., 2013. A Practical, Self-Adaptive Voice Activity Detector for Speaker Verification with Noisy Telephone and Microphone Data. ICASSP .

KINNUNEN, T., CHERNENKO, E., TUONONEN, M., FRNTI, P., and LI, H., 2007. Voice activity detection using MFCC features and support vector machine. Int. Conf. on Speech and Computer, vol. 2, pp. 556–561.

LEE, B., and JOHNSON, M., H., 2007. Minimum Mean-Squared Error a Posteriori Estimation of High Variance Vehicular Noise. In-Vehicle Corpus and Signal Processing for Driver Behavior. Springer, pp.221-232

LEGGETTER, C., J., and WOODLAND, P., C., 1995. Maximum likelihood linear regression for speaker adaptation of continuous density hidden Markov models. Computer Speech and Language, pp. 171-185, No. 9.

LIAO, H., and GALES, M., J., F., 2005. Uncertainty Decoding for Noise Robust Automatic Speech Recognition. In Proceeding in Interspeech. pp. 3129-3132.

LIBERMAN, A., 1996. Speech: A special code. The MIT Press.

LIM, J., S., and OPPENHAIM, A.,V., 1979. Enhancement and bandwidth compression of noisy speech. Proceeding of the IEEE, vol.67, no.12, pp.1586-1604.

LIU, L., ALAM, M., and FU, X., 2005. Voice Activity Detection and Noise Reduction.

MAK, M., -W., and YU, H., -B., 2012. Comparison of Voice Activity Detectors for Interview Speech in NIST Speaker Recognition Evaluation. INTERSPEECH 2011, 12th Annual Conference of the International Speech Communication Association, Florence, Italy, August 27-31, 2011

MARTIN, A., CHARLET, D., and MAUUARY, L., 2001. Robust speech/non-speech detection using LDA applied to MFCC. Acoustics, Speech, and Signal Processing, 2001. Proceedings. (ICASSP'01). 2001 IEEE International Conference on, vol. 1, pp. 237–240.

MARTIN, R. Spectral Subtraction Based on Minimum Statistics, in: Proceedings of European Signal Processing Conference (EUSIPCO), (Edinburgh, Scotland, Great Britain), Sept. 1994, pp. 1182–1185.

MOLAU S., 2003a. Normalization in the acoustical feature space for improved s peech recognition. PhD thesis, Aachen University.

MOLAU, S., HILGER, F., and NEY, H., 2003b. Feature space normalization in adverse acoustic conditions. Proceedings of ICASSP 2003, Hong Kong.

MORENO, P., J., 1996. Speech Recognition in Noisy Environments, PhD thesis, Carnegie Mellon University, Pittsburgh, Pennsylvania.

MORENO, P., J., EBERMAN, B., 1997. A new algorithm for robust speech recognition: the delta vector taylor series approach. Proceedings of EuroSpeech-97.

MORENO, P., J., RAJ, B., and STERN, R., 1998. Data-driven environmental compensation for speech recognition: a unified approach. Speech Communication, 24, 4,1998,267-288.

NOLAZCO, J. A., & YOUNG, S. J. (1994). Continuous speech recognition in noise using spectral subtraction and HMM adaptation. Proceedings of ICASSP: IEEE, International Conference on Acoustics, Speech and Signal Processing, Adelaide, South Australia, Australia, pp. 409-412.

PANDIT, P., and BHATT, S., 2014. Automatic Speech Recognition of Gujarati digits using Dynamic Time Warping. International Journal of Engineering and Innovative Technology Volume 3, Issue 12.

PAPOULIS, A., and PILLAI, S., U., 2002. Probability, Random Variables and Stochastic Processes. McGraw-Hill, fourth ed.

PETTERSEN, S., G., S., 2008. Robust Speech Recognition in_the Presence of Additive Noise. Phd. Thesis, 2008.

PRACTICALCRYPTOGRAPHY,http://practicalcryptography.com/miscellaneous/machine-learning/guide-mel-frequency-cepstral-coefficients-mfccs/

RABINER, L., and SAMBUR, M., 1975. An algorithm for determining the endpoints of isolated utterances. Bell System Tech. Jour., vol. 54, no. 2, pp. 297–315.

RABINER, L., and SCHAFER, R., 1978. Digital processing of speech signals. Prentice Hall.

RABINER, L., 1977. On the use of autocorrelation analysis for pitch detection. Acoustics, Speech and Signal Processing, IEEE Transactions on, vol. 25, no. 1, pp. 24–33.

RAJ, B., SELTZER, M., L., and STERN, R., M., 2004. Reconstruction of missing features for robust speech recognition.Speech Communication, vol. 43, no. 4, pp. 275-296.

RAMIREZ, J., GORRIZ, J., M., and SEGURA, J., C., 2007.Voice Activity Detection.Fundamentals and Speech Recognition System Robustness. ISBN 987-3-90213-08-0, pp.460, I-Tech, Vienna, Austria.

RAMIREZ J., SEGURA, J., C., BENITEZ, C., DE, LA, TORRE, A., and RUBIO, A., 2004a. Efficient voice activity detection algorithms using long-term speech information. Speech communication, vol. 42, no. 3-4, pp. 271–287.

RAMIREZ J., SEGURA, J., C., BENITEZ, C., DE, LA, TORRE, A., and RUBIO, A., 2004b. Voice activity detection with noise reduction and long-term spectral divergence estimation. In Acoustics, Speech, and Signal Processing, 2004. Proceedings. (ICASSP ’04). IEEE International Conference on, vol. 2, p. ii.

ROSENBERG, A. E. 1975. New techniques for automatic speaker verification. IEEE Trans. Acoust., Speech, Signal Processing, vol. ASSP-23, pp.169-176

SCHAFER, R., & RABINER, L. (1975). Digital representations of speech signals. Proc. IEEE, vol. 63, no.4, pp, 662-667.

TUFEKCI, Z., GOWDY, J., N., GURBUZ, S., and PATTERSON, E., 2001. Applying Parallel Model Compensation with Mel-Frequency Discrete Wavelet Coefficients for Noise Robust Speech Recognition. Eurospeech.

TUFEKCI, Z., GOWDY, J., N., GURBUZ, S., and PATTERSON, E., 2006. Applied mel-frequency discrete wavelet coefficients and parallel model compensation for noise-robust speech recognition. Speech Communication 48 (2006) 1294–1307. Science Direct.

WOO, K., YANG, T., PARK, K., and LEE, C., 2000. Robust voice activity detection algorithm for estimating noise spectrum. Electronics Letters, vol. 36, no. 2, pp. 180–181.

XIONG, X., 2006. Speech Enhancement with Applications in Speech Recognition. Phd. Thesis, 2006.

ZHANG, Y., TANG, Z., -M., LI, Y., -P., and LUO, Y., 2014. A Hierarchical Framework Approach for Voice Activity Detection and Speech Enhancement. The Scientific World Journal Volume 2014.

CURRICULUM VITAE

Mohamad Dia Abdulkarim was born in Michigan, USA in 1987. He completed his university education in the department of Computer Engineering of Al-Mamoun private University, in cooperation with Sunderland University, UK in 2010.