Novel Statistical Voice Activity Detectors

Abhijeet Sangwan

A Thesis

in

The Department

of

Electrical and Computer Engineering

Presented in Partial Fulfillment of the Requirements

for the Degree of Master of Applied Science at

Concordia University

Montreal, Quebec, Canada

January 2006

Your file Votre référence
ISBN: 0-494-14281-2
Our file Notre référence
ISBN: 0-494-14281-2

ABSTRACT

Novel Statistical Voice Activity Detectors

Abhijeet Sangwan

In this thesis, we propose a few practical statistical voice activity detectors (VADs) which combine the voice activity information in the short-term and long-term statistics of the speech signal. Unlike most VADs, which assume that the cues to activity lie within the frame alone, the proposed VAD schemes seek information for activity in the current as well as the neighboring frames. Particularly, we develop primary and contextual detectors to process the short-term and long-term information, respectively. We use the perceptual Ephraim-Malah (PEM) model to develop three primary detectors based on the Bayesian, Neyman-Pearson (NP) and competitive NP (CNP) approaches. Moreover, upon viewing voice activity detection as a composite hypothesis where the prior signal-to-noise ratio (SNR) forms the free parameter, we reveal that a correlation between the prior SNR and the hypothesis exists, i.e., a high prior SNR is more likely to be associated with 'speech hypothesis' than the 'pause hypothesis' and vice-versa, and unlike the Bayesian and NP approaches, the CNP approach alone exploits this correlation. Further, we also develop a contextual detector which uses the statistics of the speech burst and pause periods to render decisions. Subsequently, we combine the contextual detector with the primary detectors to obtain the comprehensive VADs (CVADs), i.e., the Bayesian, NP and CNP primary detectors yield the CVAD-Bayesian, CVAD-NP and CVAD-CNP detectors. Finally, the proposed VADs are tested under various noises and different SNRs, using speech samples from the SWITCHBOARD database and are compared with the adaptive multi-rate (AMR) VADs. A number of objective and subjective evaluation parameters are used to judge

the results which show that (i) the CNP detector outperforms the NP and Bayesian detectors, and compares well to the AMR VADs, (ii) the CVAD-NP and CVAD-CNP match or outperform the AMR VADs, (iii) the contextual detection scheme gives significant improvements with minimal computational overhead, (iv) the CVAD-NP and CVAD-CNP exhibit good speech and pause detection capability, respectively, and (iv) the computational complexity of the proposed VADs is very low.

Acknowledgments

My sincerest gratitude to my thesis supervisors, Dr. W.-P. Zhu and Dr. M.O. Ahmad, who provided me with a wonderful opportunity to pursue my research goals at Concordia University. This work would not have been possible without their guidance, assistance and support.

A special thanks to my family, who are a constant source of inspiration and strength, and to all my friends and colleagues for their constant encouragement.

Abhijeet Sangwan, January 2006.

To my family: Mom, Dad and Purnima ...

Contents

List of Figures

List of Abbreviations

AMR	:	Adaptive Multi-Rate
CB	:	Critical Band
CNP	:	Competitive Neyman-Pearson
CVAD	:	Comprehensive Voice Activity Detector
SNR	:	Signal to Noise Ratio
EM	:	Ephraim-Malah
ETSI	:	European Telecommunications Standards Institute
ITU	:	International Telecommunications Union
LR	:	Likelihood Ratio
LRT	:	Likelihood Ratio Test
MFSC	:	Mel Frequency Spectral Coefficients
NP	:	Neyman-Pearson
PEM	:	Perceptual Ephraim-Malah
SS	:	Sufficient Statistics
SSP	:	Same State Period
VAD	:	Voice Activity Detector

List of Symbols

H_1	:	Speech hypothesis
H_0	:	Pause hypothesis
$E[.]$	:	Expectation
$Var[.]$	:	Variance
$\mathcal{N}$	:	Gaussian distribution
$P(.)$	:	Probability
$p(.)$	:	Probability density function
$\mathbf{\Psi}$	:	Free parameter space of the composite hypothesis
$erf(.)$	:	Standard error function
$\ln(.)$	:	Natural logarithm
$\mathbf{\Lambda}$	:	Likelihood ratio
$\bigcup$	:	Union
$\bigcap$	:	Intersection
$\sum$	:	Summation
tr	:	Trace of a matrix
ζ	:	Prior SNR

Chapter 1

Introduction

1.1 Research Motivation

CONVERSATIONS are a sequence of contiguous segments of silence and speech [1]. The ability to segregate the speech and silence components of a conversation is of interest in many speech applications, and a system which accomplishes this task is known as a voice activity detector (VAD) [2].

VAD is an important part of many modern speech communication systems such as hands-free telephony, mobile telephony, voice over internet protocol (VoIP), audio-conferencing, echo cancellation, speech coding, VSAT (very small aperture terminals), speech recognition and enhancement [3–7]. Voice activity detection is also an integral part of many wireless cellular and Personal Communications Systems (PCS) standards [3]. The use of VAD in second and third generation cellular systems would facilitate an efficient consumption of the available radio-frequency (RF) spectra [8]. Similarly, in VoIP systems the bandwidth consumption is reduced by selectively encoding and transmitting noisy speech frames [4, 9]. VAD also provides benefits like increasing the number of radio channels and reducing the power consumption in

portable equipments [10]. For instance, a VAD can increase the channel capacity in CDMA systems by a factor of 2 [6, 11], and increase the bandwidth utilization efficiency while improving the throughput/delay performance of the data transmission in GSM/GPRS systems with minimum impact on the service [5].

The seemingly easy task of voice activity detection becomes difficult when conversations occur in noisy backgrounds, where speech has to be detected in presence of non-stationary and unpredictable real-world noises [1]. Further, the difficulty increases if the signal-to-noise ratio (SNR) of the noisy speech is lowered [3,12]. Therefore, in practice, the mobile environment of cellular telephone systems is the most challenging scenario for voice activity detection as it is least controlled and speech is subjected to a variety of acoustical noises and SNRs. Thus, it is no surprise that the major focus of research in voice activity detection has been towards developing low complexity, efficient and robust VADs for communication systems. Herein, the International Telecommunications Union (ITU) and the European Telecommunication Standards Institute (ETSI) have adopted the G.729 Annex B and adaptive multi-rate (AMR) VAD, respectively as the de facto standards for communication systems [13,14]. Futher, various reseachers have developed different strategies to tackle the problem of voice activity detection, where the case of statistically modeled VADs is noteworthy due to their consistent performance across various noises and different SNRs [15–21]. Moreover, the statistical VADs have also presented an alternative to the heuristically designed traditional VAD schemes as the former are more tractable than the later. In general, it is far easier to tune the relevant parameters of a statistical VAD and extract an optimum performance. However, the foray into statistically modeled VAD systems has still left many unresolved issues, and the design of a low complexity robust VAD continues to be an open research problem. In this thesis, we address some of these key issues and attempt to develop a VAD scheme which

operates reliably at low SNR.

1.2 Scope and Structure of the Thesis

The objective of this thesis is to develop a VAD scheme which is capable of delivering robust performance at low SNR. The proposed VAD design is inspired by the human auditory system, where it has been observed that audition and perception are highly complex, as they utilize the multiple layers of redundancy in speech for detection and recognition. Moreover, humans perform the task of speech and sound perception with relative ease in low SNR conditions, as they increasingly exploit the diverse cues in speech available at the acoustic, linguistic and prosodic levels. On the contrary, contemporary VAD schemes are known to perform very well at high SNR but fail in noisy conditions. We believe that the reason behind the poor performance of the VADs at low SNR is the inability of the VAD to differentially utilize the various cues in the speech signal. Hence, in this thesis, we propose to combine the voice activity cues from two disparate sources in order to make a robust speech/pause decision for the given frame in question, i.e., we compute the short-term and long-term statistics of the noisy speech signal to extract the cues available at acoustical and utterance level, respectively. The short-term and long-term information is processed separately using the primary and contextual detectors, where a likelihood ratio (LR) for voice activity is developed for both and subsequently combined into a single likelihood ratio test (LRT). In this manner, we use the contextual cues to provide robustness at low SNR, when the primary cues are often corrupted and unreliable for detection.

Most statistical VADs use the Ephraim-Malah (EM) model, which assumes that the Fourier transform coefficients of speech and noise are statistically independent zero-mean Gaussian random variables [22]. However, this model fails to accomodate

the perceptual properties of speech. The benefit of incorporating perception has been observed in automatic speech recognition (ASR), where the perceptual feature 'mel frequency cepstral coefficients' (MFCCs) has become the industry standard. Hence, in this thesis, we propose a perceptual EM (PEM) model for the design of the primary detector, where a mel based feature is used instead of the Fourier transform.

The problem of voice activity detection using the PEM (or EM) model can be viewed as a composite hypothesis [23], with the prior SNR acting as the free parameter. Particularly in this composite hypothesis, there exists an intuitive relationship between the free parameter and the hypotheses, i.e., a high value of prior SNR is more likely to indicate 'speech hypothesis' than 'pause hypothesis' and vice-versa. It is worth mentioning that so far this crutial prior information about the free parameter (prior SNR) has been ignored by the EM based VADs, which use the prior SNR estimates solely for computing the test statistics. The use of prior information has shown improved performance in other detection problems like 'dipole detection by using magnetoencephalography (MEG) and electroencephalography (EEG)' [24], and 'radiodense versus radiolucent tissue detection in digitized mammograms' [25], which motivates the use of prior information in voice activity detection as well. Hence, in order to incorporate the partial prior information about the free parameter into the detector, we analyze the Bayesian, Neyman-Pearson (NP) and competitive NP (CNP) design approaches [26], and show that the CNP approach alone is capable of modeling the prior information about the free parameter into the detector design. It is useful to note that the CNP approach was recently proposed by Levitan and Merhav [26], and it is yet to see an application in VAD. Hence, via the CNP approach, this thesis pioneers the use of prior information in statistically modeled VADs which have so far depended entirely upon posterior information in the noisy speech signal for detection.

Lastly, a contextual detector which uses the information in the long-term speech

and pause durations to render decisions is also developed. From a functional perspective, the contextual detector is similar to a hang-over scheme as it attempts to correct the errors made by the primary detector. However, unlike contemporary hang-over schemes which work on the individual decisions of the primary detector, the contextual detector aggregates the individual speech and pause decisions into speech bursts and pause periods, respectively. Thereafter, it uses statistical models of the speech and pause durations to build a contextual LR. Lastly, it may be noted that the contextual and primary detectors are developed independently, which increases their efficacy as it maintains the possibility of using the primary detector alone, and combining the contextual scheme with other existing statistical VAD schemes.

The layout of the thesis is as follows: In Chap. 2, we discuss the background material on voice activity detection, where we start with a primer on speech processing, followed by a review of the contemporary VAD systems and approaches. In Chap. 3, we develop the proposed primary VADs using the Bayesian, NP and CNP approaches, and the PEM model. We also show the superiority of the CNP over the NP as a more generalized design approach. Finally, we develop the contextual detector which is subsequently combined with the primary detectors to yeild the comprehensive VADs. In Chap. 4, we test the proposed VADs using computer simulation, and compare the performances with AMR VADs. Further, we evaluate the results obtained based on a number of objective and subjective parameters.

Chapter 2

Background

IN this chapter, we present the background material for voice activity detection. The first part of the chapter deals with the fundamentals of speech signal such as voicing, articulation, acoustics, masking, critical-band phenomenon and context, which are relevant towards understanding the design and operation of a VAD. In the second part of the chapter, we discuss the problem of voice activity detection in detail and review several contemporary VADs. We also highlight the advantages and drawbacks of the different approaches towards voice activity detection.

2.1 Fundamentals of the Speech Signal

Speech signals are time varying pressure waves that are transmitted by a speaker and serve to communicate information [27]. The source of speech is an egressive airstream from the lungs. The airstream passes through the oral and nasal cavity (together known as the vocal tract) which consists of a number of organs such as the tongue, teeth, velum etc. The organs of the oral and nasal cavity are collectively known as the articulators of the vocal tract.

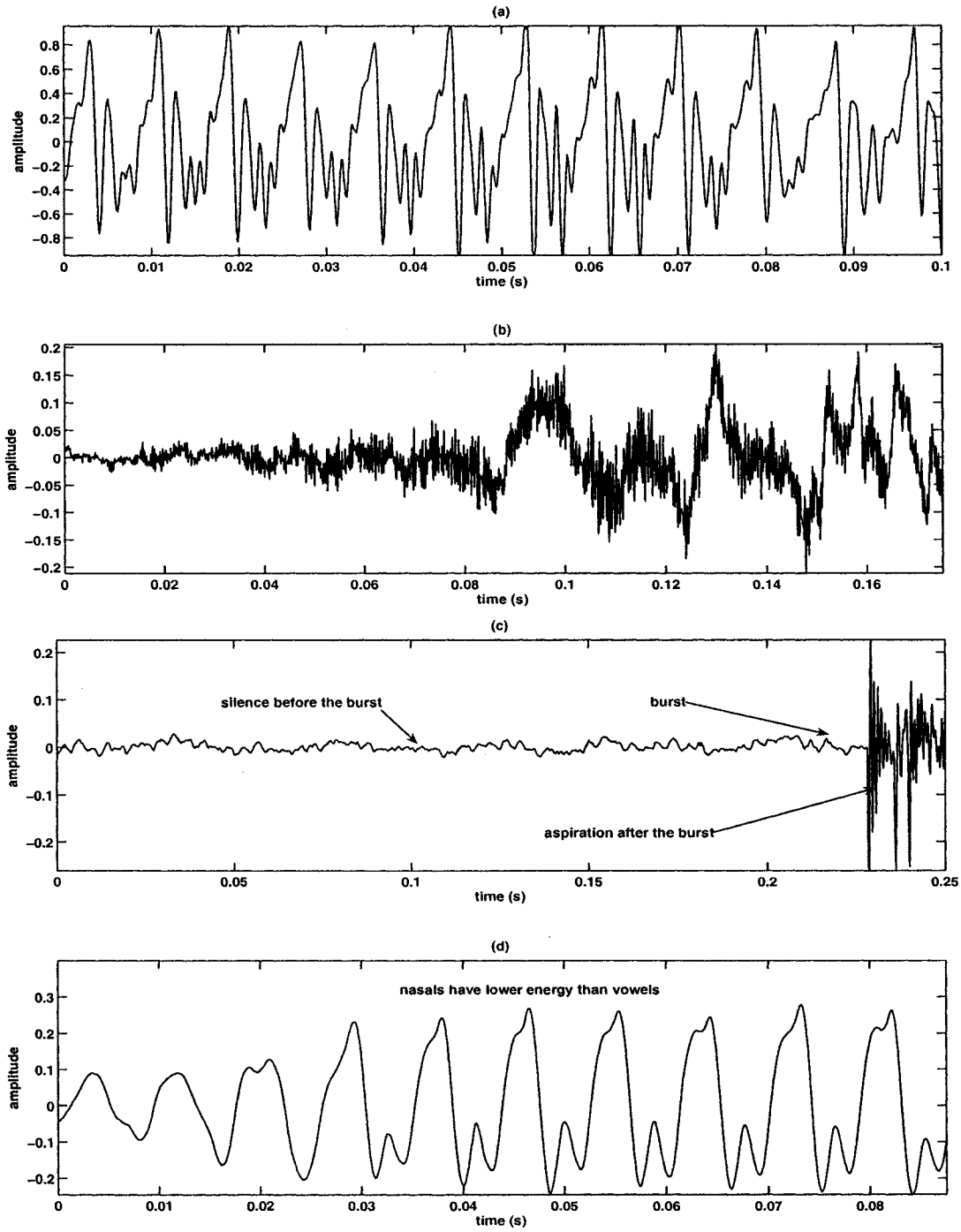

Figure 1: Illustrating waveforms of the main phoneme categories with their properties: (a) vowel /O/ - periodic and high energy, (b) fricative /S/ - noise-like and low energy, (c) stop /k/ - silence and aspiration, before and after the burst, respectively, and (d) nasal /n/ - periodic and lower energy than vowels.

2.1.1 Voicing

A very important property of all speech sounds is voicing which is associated with the functioning of a critical articulator: vocal cords. The vocal cords are present in the larynx (at the base of the vocal tract) and are capable of vibrating which produces periodic restrictions in the egressive airstream. In this manner, the vocal cords produce pulses of air which periodically excite the vocal tract and the speech hence produced is called voiced. The frequency at which the vocal cords vibrate is known as the fundamental frequency (F0) or pitch. On the other hand, unvoiced speech is produced when vocal cords are not vibrated, and an alternate narrow constriction in created elsewhere in the vocal tract. The articulatory aspects of voiced and unvoiced sounds have a clear impact on their acoustics where voiced sound signals show a periodic waveform and unvoiced sounds have a noise like appearance [27]. Figure 1 (a) and 1 (d) show the waveforms for two voiced sounds, /O/ (in c*o*ffee) and /n/ (in *n*ew). Similarly, Fig. 1 (b) and 1 (c) show waveforms for two unvoiced sounds, /S/ (in *sh*oe) and /k/ (in c*o*ffee). In voice activity detection, voiced sounds like the vowels are easy to detect owing to their high energy and prominent periodic structure. On the other hand, the similarity of the unvoiced sounds to noise or silence presents a challenge to most VADs, such as the energy based detectors which fail to recognize these sounds as speech [1].

2.1.2 Phonemes

Phonemes are the smallest unit of meaningful sound in a language [27]. They are contrastive units which are distinguishable from each other, and the combination of a sequence of phonemes forms a word. The sound produced when a phoneme is articulated is called a phone. The phonemes of the English language are shown in

Table 3: Phonemes of the English language

Category	Phoneme	Example word		Category	Phoneme	Example word
Vowels	/i/	h*ea*t		Nasals	/m/	*m*other
	/I/	*i*t			/n/	*n*o
	/e/	*a*te			/η/	r*ing*
	/ϵ/	b*e*t		Fricatives	/f/	*f*amily
	/ae/	h*a*t			/v/	*v*ery
	/u/	f*oo*l			/θ/	*th*ick
	/o/	*o*val			/δ/	*th*en
	/O/	f*ough*t			/s/	*s*lim
	/U/	p*u*ll			/z/	*z*oo
	/$\ni$/	p*u*tt			/S/	*sh*oe
	/a/	f*a*ther			/Z/	mea*s*ure
Stops	/p/	*p*an			/h/	*h*at
	/t/	*t*an		Glides	/y/	*y*ou
	/k/	*c*an			/w/	*w*ater
	/b/	*b*ig		Liquids	/l/	*l*ight
	/d/	*d*ig			/r/	*r*at
	/g/	*g*o				

Table 3. Phonemes are grouped into six major classes based on their articulatory origins and acoustical properties: vowels, nasals, glides, liquids, fricatives and stops. The waveforms for a vowel, fricative, stop and nasal are shown in Fig. 1 (a), (b), (c) and (d), respectively, illustrating some peculiar properties of each phoneme classes.

Each phoneme is associated with a unique articulatory configuration and acoustical characteristics. At a higher level, phonemes can be separated into two distinct groups, i.e., vowels and consonants. A pair of phonemes form a syllable and words are formed by combining syllables. A syllable consists of a dominant vowel sound which is either preceded or succeeded by a consonant. Figure 2 shows the hierarchial relationship between speech and phonemes along with the intermediate stages. It also shows the classification of phonemes into vowel and consonant groups.

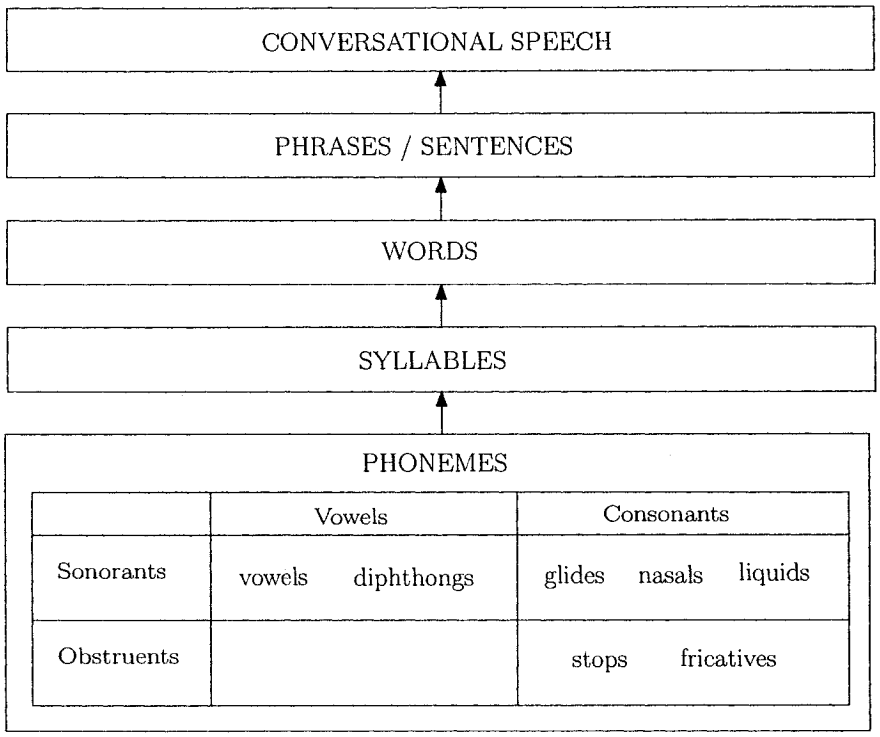

Figure 2: Hierarchial relationship between phonemes and conversational speech.

2.1.3 Articulatory Phonetics

At an articulatory level, sounds are distinguished based on their manner and place of articulation. Manner of articulation is concerned with the nature of airflow in production of sounds: the path that the airflow takes and the obstructions it faces in form of vocal tract constrictions [27]. Phonemes are grouped into a number of broad categories based on the manner of articulation:

- Vowels and diphthongs: They employ minimum or no constriction in the vocal tract and the airflow is largely unrestricted.
- Glides: Similar to vowels but employ slight vocal tract constrictions.
- Liquids: Also similar to vowels but employ the tongue to produce vocal tract constrictions.

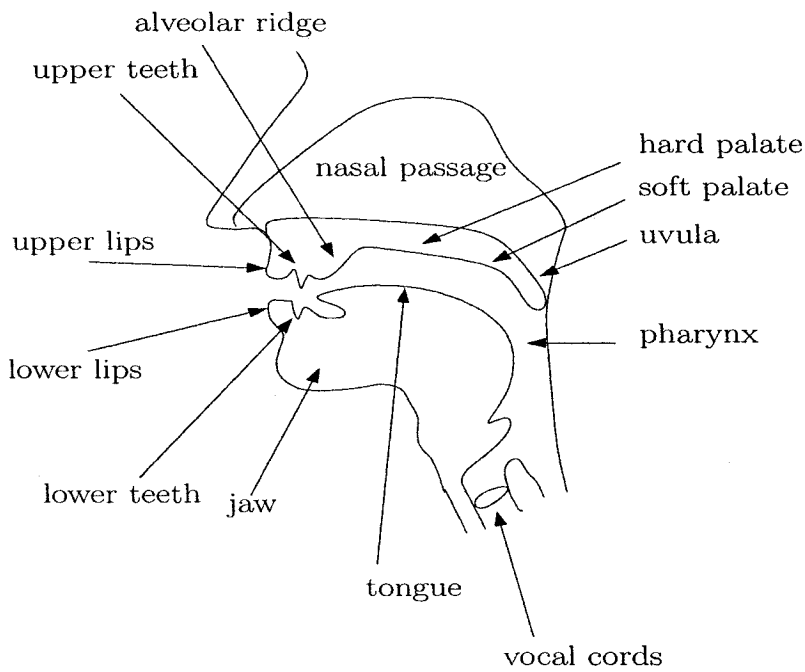

Figure 3: Major articulators and different places of articulation in the vocal tract [27].

- Nasals: Use velum as the vocal tract constriction which results in an unobstructed airflow through the nasal passage, and no or reduced airflow through the oral passage.
- Stops: Employ a closure and sudden release of a vocal tract constriction.
- Fricatives: Employ narrow vowel tract constriction which creates a noise like sound.

Vowels, diphthongs, liquids, glides and nasals employ voicing and are strong sounds which are also known as sonorants. Stops and fricatives use the vocal tract constriction as their primary source of excitation, are both voiced and unvoiced, and are also known as obstruents. Sonorants are high energy sounds which possess prominant waveforms and are easily detected by VADs. Obstruents are weak sounds with low energy and are difficult to pick up in a noisy backgrounds.

Table 4: Places of articulation.

Articulator	Example	Comments
Labials	/f/, /m/	the lips constrict or the lower lip touches the upper teeth.
Dental	/θ/	tongue touches the upper incisor teeth.
Alveolar	/n/,/s/,/t/	tongue touches the alveolar ridge.
Palatal	/S/, /Z/	tongue touches the hard palate.
Velar	/η/, /k/	tongue touches the soft palate.
Uvular	French /R/	tongue approaches the uvula.
Pharyngeal	Arabic /ha'/.	the pharynx is constricted.
Glottal	/h/	the vocal cords are constricted or closed.

The place of articulation is the location of the constriction in the vocal tract and it is critical in distinguishing between the phonemes of the same group. For instance, among the stop obstruents /p/, /t/ and /k/, the most important distinguishing factor is the place of articulation. Table 4 enumerates the different places of articulation employed by phonemes, and Fig. 3 shows the location of various articulators and places of articulation.

2.1.4 Acoustic Phonetics

Acoustic phonetics deals with the acoustical properties of phonemes and it is closely tied to audition and perception, where it is well known that the ear extracts information from the speech spectrum. The primary acoustic cues to phonemes is the dynamic behavior of the formant and spectral regions of energy. Formants are the peaks observed in the speech spectrum and are related to the position and movement of articulators [27].

Vowels are the strongest phonemes and show strong, pronounced formants. Most spectral cues to the vowels sounds lie in the lower frequency region of 0-3 kHz where they are identified and distinguished based on their formant locations. Figure 4 (a)

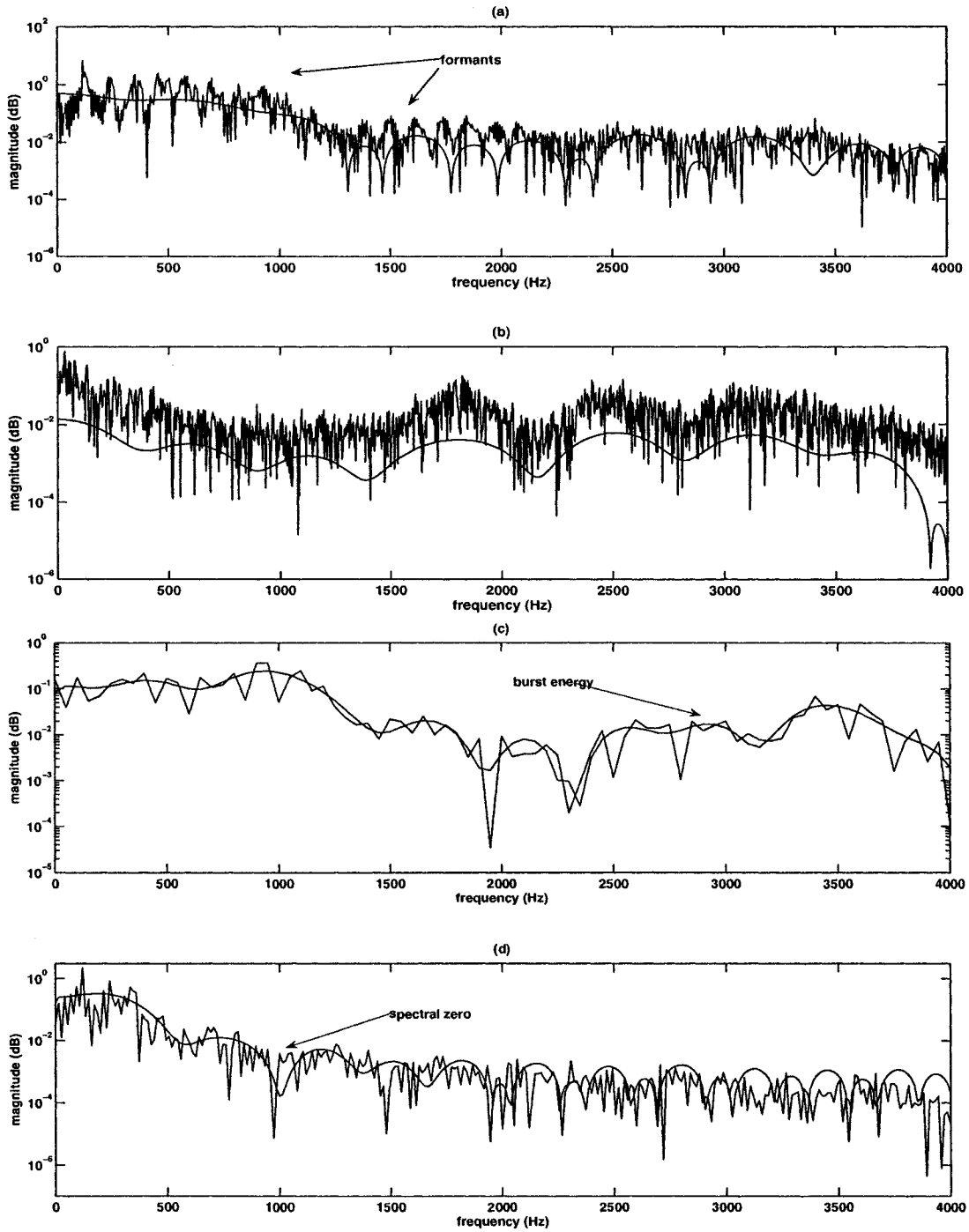

Figure 4: Illustrating the spectral properties of the main phoneme categories: (a) vowel /O/ - prominant formants, (b) fricative /S/ - diffused spectra, (c) stop /k/ - burst energy in mid frequencies, and (d) nasal /n/ - formants like vowel and a spectral zero around 800 Hz.

shows the spectrum of the vowel phoneme /O/, where the formants are highlighted. It is also seen that there is general roll-off in the spectrum with low frequencies possessing greater energy than the higher frequencies. In general glides, diphthongs and liquids show acoustical features which are similar to that of vowels. Nasals have lower energy than vowels, and this shows in their suppressed formants. Moreover, the nasal spectrum also shows the presence of a zero in the mid-frequency region, as shown in Fig. 4 (d). Fricatives are characterized by a concentration of energy in high frequencies and unvoiced fricatives mimic noise-like characteristics very closely. The noise-like spectrum of the phoneme /S/ is seen in Fig. 4 (b), where it is observed that the spectrum is relatively flat. Unlike other categories, stops are highly transient phonemes which are acoustically complex. For most stops, the closure portion of the waveform looks similar to silence or a signal of very low energy [27]. Figure 4 (c) shows the spectrum of a stop, /k/.

From Fig. 4, it is observed that the speech spectrum reflects the diversity of the speech sounds. As different phonemes are part of a conversation, the short-term cues to voice activity detection are also diverse, and well distributed in the speech spectrum. This property of speech is very important as it ensures that real-world noises, which tend to be colored, corrupt only some portions of the speech spectrum, and several voice activity cues remain well preserved in the relatively cleaner parts of the spectrum. In this thesis, the proposed primary detector exploits the above property where it estimates the clean and corrupted portions of the noisy speech spectrum, and utilizes the voice activity cues within the clean portions alone for detection.

2.1.5 Masking

A single tone is heard by the ear when its intensity exceeds a particular threshold. However, audition is complex when two sounds are played out simultaneously where one sound affects the perception of the other and this phenomenon is known as masking. The presence of a masker sound raises the threshold of perception for the maskee sound. In other words, it is harder for a listener to hear the maskee sound in presence of the masker sound. Moreover, it is also observed that lower frequency sounds always tend to mask higher frequency sounds, when played out simultaneously.

Masking is an important factor which introduces a nonlinearity into speech perception where the total response of any complex stimuli cannot be assumed to be the sum total of the individual responses [27]. Moreover, masking easily explains the difficulty humans as well as VADs face in detecting speech at low SNR, i.e., strong background noises easily mask weak phones.

2.1.6 Critical Band Phenomenon

The critical band (CB) phenomenon is commonly used to explain masking. The CBs are similar to bandpass filters whose responses map the behavior of the auditory neurons in the ear, and a set of 24 critical bands are usually used to model this behavior. The CB phenomenon suggests that among two competing sounds in a critical band, the sound with the greater energy (masker) dominates perception, and the energy of masker is indicative of the degree of masking.

Hence, using the CB phenomenon, the discussion on the effect of colored noise on speech in Sec. 2.1.4 can be restated in more effective terms, i.e., while colored noises mask the speech sounds in certain critical bands, speech is relatively clean in the other critical bands. Hence, humans must rely on the relatively clean critical

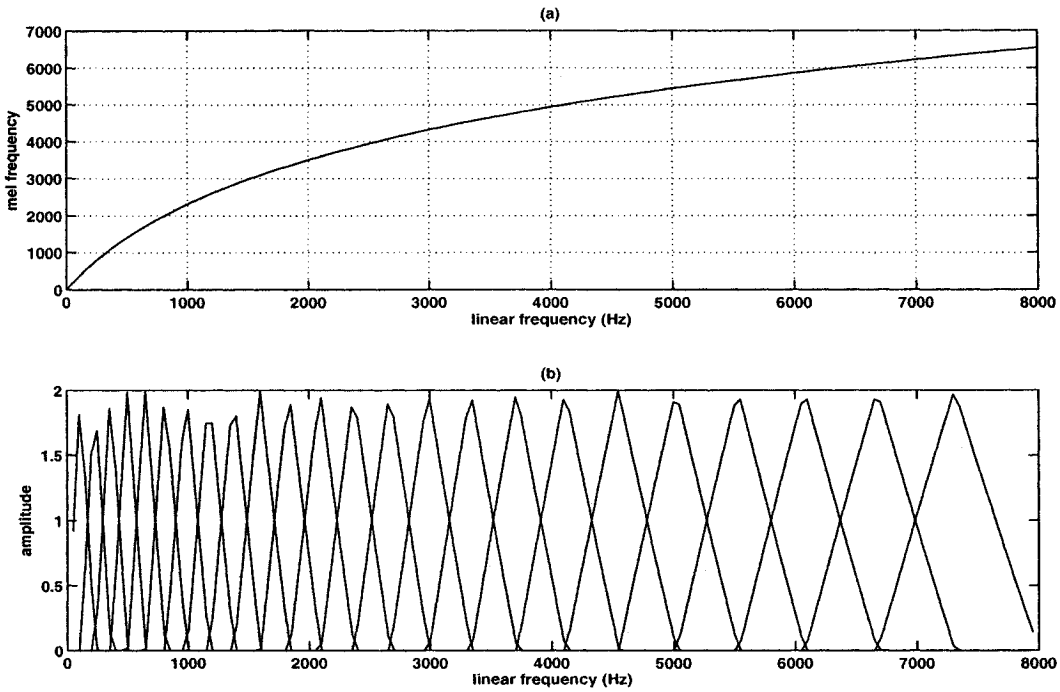

Figure 5: Illustrating the perceptual properties of speech: (a) the mapping between mel frequency and linear frequency, and (b) the structure of the triangular mel filter banks which mimic the critical bands.

bands to decipher the sound stimulus [21]. The fact that humans do a very good job of grasping different sounds at low SNRs implies that they exploit the diversity of acoustical cues present in the speech signal. As mentioned in Sec. 2.1.4, the proposed primary detector is inspired by this observation as it determines the corrupted and clean critical bands, and utilizes the speech cues within the clean critical bands to make a decision.

2.1.7 Mel scale

The critical bands show interesting peculiarities such as unequal bandwidths and asymmetrical filter shapes. The bandwidth of the critical bands increase with increasing frequency which indicates that the perceived frequency of a stimulus is different

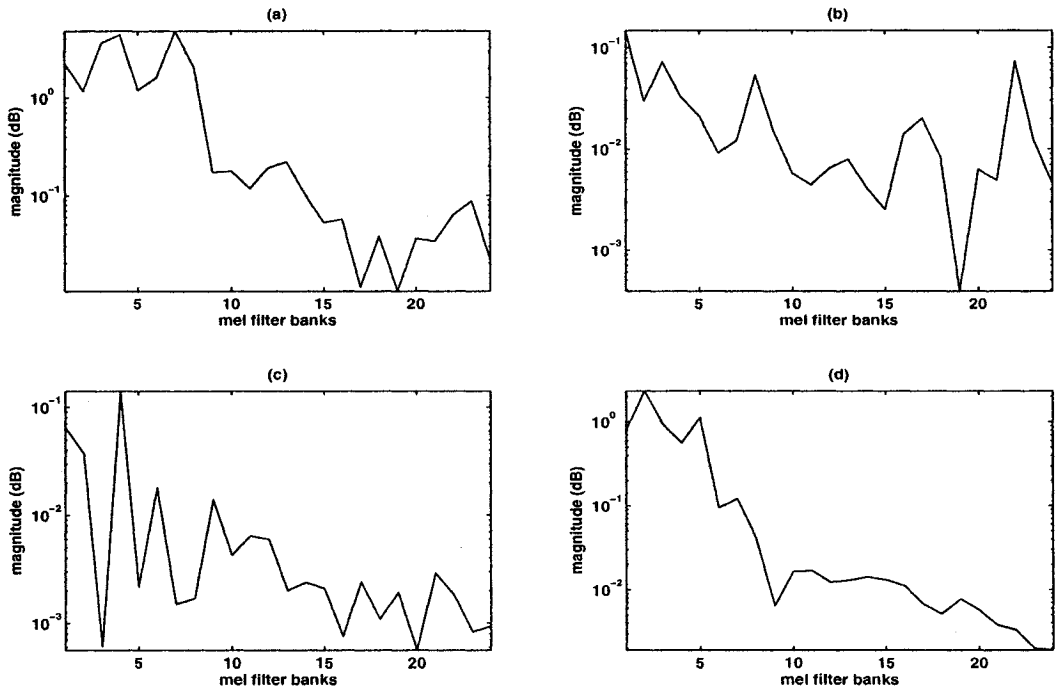

Figure 6: Illustrating the MFSC magnitude spectrum of (a) vowel /O/, (b) fricative /S/, (c) stop /k/, and (d) nasal /n/.

from the acoustical frequency. In order to accommodate this observation, a perceptual measure known as the mel scale is used to map the acoustical frequencies to the perceived frequencies, as shown in Fig. 5 (a).

The mel scale has been extensively used in speech processing applications, in the form of MFCCs. The mel frequency transformation is applied to the speech power spectra by using a set of triangular filters, as shown in Fig. 5 (b), where the increasing bandwidth of the triangular filter mimics the critical band structure. The mel filter banks transform the signal from a normal frequency domain to a perceptual frequency domain. In this thesis, we propose a new mel based speech feature for voice activity detection termed as the mel frequency spectral coefficients (MFSCs), where the speech spectrum is obtained by using the discrete cosine transform (DCT) rather than the Fourier transform. The MFSCs are obtained by projecting the DCT coefficients onto

the mel filter banks [21], i.e.,

$$\mathbf{F} = \mathbf{M}^T\mathbf{D}^T\mathbf{X} \tag{1}$$

where $\mathbf{F} = [f_1, f_2, ..., f_M]^T$ are the MFSCs, $\mathbf{M}$ is the mel filter transformation as shown in Fig. fig:melscale (b), $\mathbf{D}$ is the DCT transform and $\mathbf{X}$ is a frame of the input noisy speech signal. Here, the matrix product $\mathbf{DM}$ represents the perceptual frequency transformation. Fig. 6 shows the MFSC magnitude spectra of different speech sounds.

2.1.8 Context and Redundancy in Speech

Speech is a highly redundant signal where a wide variety of cues are available for humans to detect and recognize different speech sounds. For instance, speech can be detected and deciphered using coarticulation at acoustical level, lexical knowledge at word level, grammar at sentential level along with general knowledge, knowledge of the speaker and the conversation context. The high redundancy in speech implies that voice activity detection information is simultaneously available in the short-term and long-term statistics of the speech signal.

In this thesis, we are particularly interested in detecting the voice activity cues present at the utterance level, where it is well known that speech and pause periods are sustained over short durations of time. It may be useful to note that the term 'pause' refers to a period of non-speech in the conversation, and it may be silence or background noise in different circumstances. Typically, pauses in read speech are 1500ms, 530ms and 130 ms between paragraphs, sentences and phrases, respectively [28]. Pauses tend to be the longest and most frequent in conversations where the typical speech burst and pause last 300ms and 157ms, respectively [29] [30]. Hence, the speech and pause durations in a conversation offer critical voice activity cues

which could be used along with the cues in the short-term statistics to provide a robust VAD.

2.2 Existing Voice Activity Detectors

Numerous design strategies for voice activity detection exist in contemporary literature. These design approaches can be broadly classified into two groups, with one group comprising of the heuristically designed traditional VAD schemes, and the other representing the new breed of statistically modeled VAD systems. In this section, we first give a brief description of the general structure and operation of a VAD, and then review several VADs belonging to each of the above mentioned groups.

2.2.1 General VAD Structure and Operation

Voice activity detectors detect speech bursts and pauses in a conversation using some form of a speech pattern classification. At first, the speech signal is sliced into contiguous frames of an appropriate length, i.e., generally, a frame length of 10 ms to 20 ms is chosen in order to adhere to the real-time constraints of communication systems. Now, each frame is analysed for voice activity and subsequently classified as speech or pause. Typically, the decision is made by associating a real-valued parameter with each frame, and comparing it to a threshold [1]. In Fig. 7 (a), the operation of a simple energy based VAD is shown. The VAD divides the speech signal into frames and estimates a global noise energy threshold using pause only periods. Then, the decisions are taken by comparing the average energy of each speech frame with the noise threshold where speech or pause is declared if the average frame energy is greater or lesser than the threshold, respectively. In general, the above description

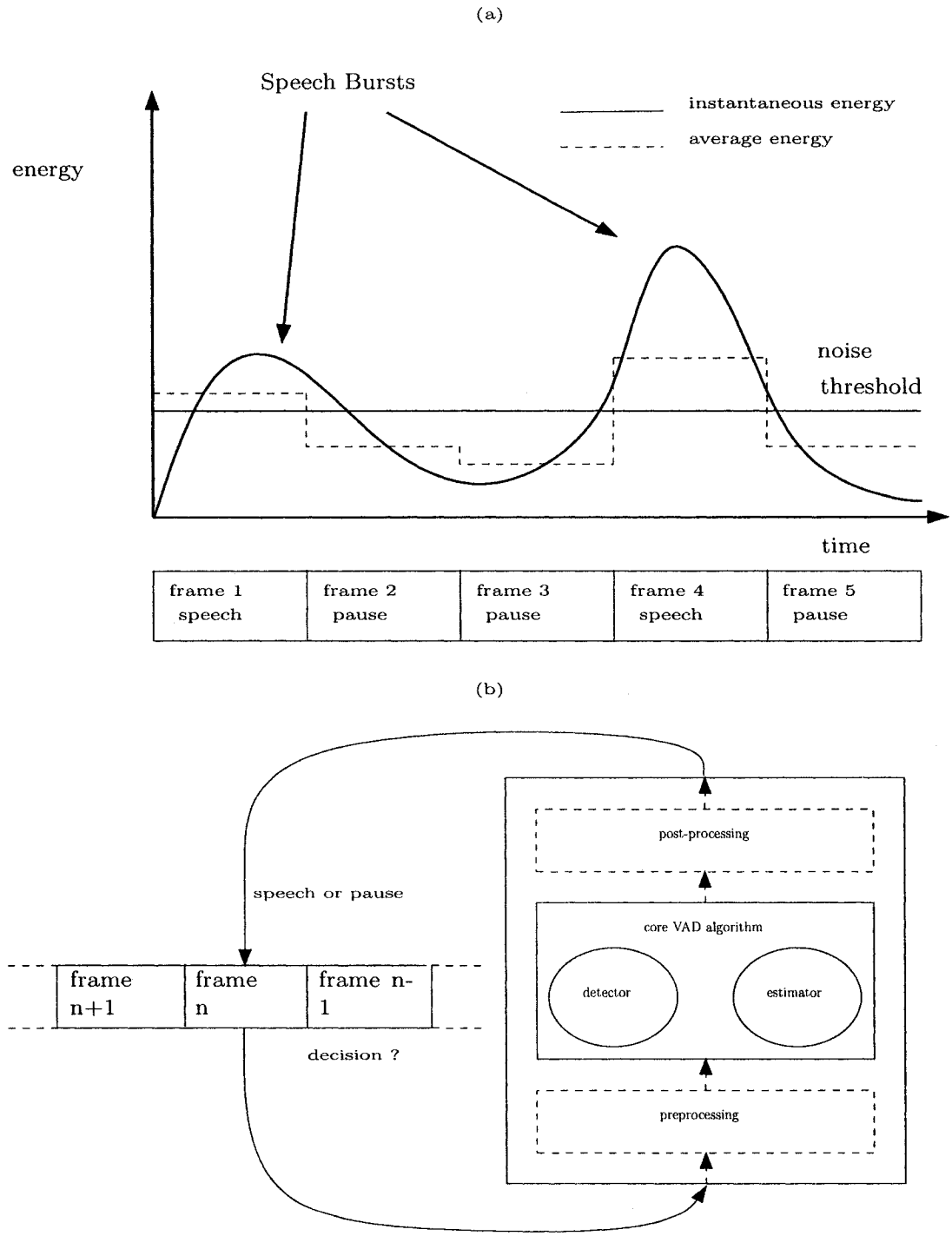

Figure 7: (a) Operation of a simple energy based VAD. (b) Three stage VAD system consisting of a preprocessing, core VAD algorithm and post-processing stage.

of the VAD operation is universal as most VADs use the parameter-threshold comparison for decision making and pause only periods for noise parameters estimation. However, in the evolution of VADs many parameters other than the signal energy such as the zero crossing rate (ZCR), linear prediction coefficients (LPC) etc. have been successfully used for detection.

As illustrated in Fig. 7 (b), most modern VADs follow a three stage acrhitecture which comprises of a core VAD algorithm sandwiched between a preprocessing and post-processing layer. The purpose of the preprocessing stage is to enhance the efficacy of the signal for detection by performing certain functions like framing, windowing, frequency-domain transformations etc. Similarly, the post-processing stage is generally used to reasses and correct the mistakes in the decisions of the core VAD algorithm by employing hang-over schemes. The pre and post-processors work towards boosting the performance of the core VAD algorithm which continues to be the central decision maker of the VAD.

2.2.2 VADs based on General Speech Features

Short-term energy is among the most commonly used feature in voice activity detection, where both the time and frequency domains have been used for the energy computation [1,31]. Generally, the frequency domain calculations give better performance over the time domain at the cost of additional computational complexity [32]. Most frequency domain algorithms divide the spectrum into sub-bands and calculate the energy separately for each sub-band. Thereafter, unlike the time domain algorithms which make a single VAD decision after the energy computation, the frequency domain algorithms make a series of intermediate decisions for each sub-band and finally combine the individual decisions to make a final VAD decision. For instance, Marzinzik and Kollmeier developed a highly complex energy based algorithm which

uses frequency band specific energies to track the power envelope dynamics of the speech signal, and make the VAD decisions [31]. Hence, in many ways the frequency domain algorithms exploit the CB phenomenon of audition which was discussed in Sec. 2.1.6. In general, energy based VADs are low complexity algorithms and easy to implement. On the other hand, most algorithms use heuristics to divide the spectra into sub-bands and to set up thresholds. Consequently, energy based VADs are prone to miss-detect weak, low energy phonemes.

Another popular feature used by many VADs is the zero crossings rate (ZCR) [1,2]. The ZCR based VADs exploit a well known fact that the rate of zero crossings for a speech signal falls within an established range. However, background noises may share the ZCR range of the speech signal and this is the major drawback of the ZCR VADs. The cepstrum also has been used as a feature by VAD algorithms [9, 33]. Cepstral VADs generally resort to a speech pattern classification by building templates for 'speech' and 'pause', which are later used for classification. Some other commonly used features are the linear predictive coefficients (LPC) [31], autocorrelation function [9], periodicity measure [34] and pitch [35]. It is also common among VAD developers to combine some of the aforementioned features to form a comprehensive or fusion systems [2,32]. Comprehensive or fusion VADs rely on the paradigmm that increasing the dimensionality of a feature leads to an improved performance of the detector.

Perhaps the most popular VADs to use general speech properties are the G.729 VAD and AMR VADs. The G.729 annex B uses a combination of line spectral frequencies (LSF), full band energy, low band energy and ZCR for detection [13]. Similarly, the AMR VAD algorithm 1 adopts a frequency band energy computation approach, where the SNR is estimated in as many as nine bands. The decision is then taken by comparing the SNR estimates with thresholds where the thresholds are adapted with noise. The AMR VAD algorithm 2 divides the speech frame into

two sub-frames and computes the channel power, voice metrics, and noise power for each sub-frame. The sub-frame is judged as speech, if the measured parameters exceed certain adaptable thresholds, and the overall frame is declared speech, if at least one sub-frame is speech [36].

In general, VAD systems based on speech properties tend to be heuristic in development and inconsistent in performance [16]. Particularly, these schemes fails in mobile environments like car, busy streets and other noisy public places, especially when the SNR is low. This is largely because the design technique for the traditional VADs makes it harder to tune the relevant VAD parameters in order to adapt to the non-stationarity of the speech and noise processes, which results in a underperforming algorithm [16]. On the other hand, the operation of the traditional VADs is straight-forward and intuitive, which reflects in their popularity.

2.2.3 Statistical VADs

Recently, there has been a growing interest in statistically modeled VAD systems. Statistical VADs are more tractable than the traditional VADs as they are packaged with tunable parameters which can be conveniently varied and set for desired performance. Moreover, the design technique of statistical VADs ensures a consistent performance across a wide variety of noises and SNRs. Most statistical VADs adopt the model proposed by Ephraim and Malah (EM, [22]) which assumes that the Fourier transform coefficients of speech and noise are statistically indepedent zero-mean Gaussian random variables [16, 21, 37]. Some authors have proposed deviations from the traditional EM model such as Chang and Kim, who use a Laplacian distribution [18], and Gazor and Zhang, who use a Laplacian-Gaussian model [19]. Further, some unconventional VADs such as those based on higher order statistics (HOS) in the LPC residual domain and the statistical chi-square test have also been proposed

recently [17,20].

Among the first VADs to use the EM model was the VAD proposed by Sohn and Sung, which uses the maximum likelihood criterion for estimating the unknown parameters [38]. This VAD was later refined by Sohn et. al., who devised a decision directed (DD) technique for the estimation of the unknown parameters [15]. Another improvement over the Sohn VAD was proposed by Cho and Kondoz, who employed a likelihood ratio smoothening technique which resulted in better detection performance [37]. More recently, an EM based VAD which uses a special SNR measure and threshold adaption technique was proposed by Davis and Nordholm [16]. Similarly, Ramirez et. al. have also proposed a EM based multiple observation VAD scheme which is claimed to have reduced the variance of the LRT and improved the performance of the detector. The above-mentioned EM VADs have shown good performance in comparison to the G.729 B and AMR VADs across various SNRs and background noises. Further, the other significant advantage of the EM VADs is their low complexity and simplicity in implementation where the EM VADs have significantly lower number of parameters to tune when compared to the G.729 B or AMR VAD. However, the EM VADs have failed to address some of the issues as listed below:

- So far, EM VADs have not explored the usage of speech features based on perceptual properties such as the MFCCs. The use of perceptual features has shown benefits in parallel fields such as speech recognition which motivates their use in VADs as well.

- The Bayesian form of the LRT has been the popular choice of implementation for the EM VADs. However, setting an appropriate value for the Bayesian threshold is difficult and not intuitive. This is because the prior probabilities

of the speech and pause hypothesis are usually unknown and certainly not universal, i.e., they change from one conversation to another. This makes the tuning of the EM VADs unintuitive and cumbersome.

- The binary hypothesis posed by the EM model can be viewed as a composite hypothesis testing problem where the prior SNR term acts as the free parameter. The EM VADs estimate the value of the prior SNR from the data itself, and use this value in the detector. However, all VADs ignore the prior information that is available for the prior SNR in terms of a general relationship, i.e., a high value of prior SNR is more likely to be associated with 'speech hypothesis' than 'pause hypothesis', and vice-versa.

In this thesis, we adopt the following strategy to tackle the above-mentioned issues:

- In order to incorporate the perceptual properties of speech, we propose a new variant of the mel based speech features known as the MFSC, and explore its use in voice activity detection.

- We analyze the recently proposed competitive Neyman-Pearson (CNP) approach towards detector design and show that unlike the Bayesian or NP approach, it is adept at modeling prior information into the detector. Further, we develop Bayesian, NP and CNP detectors, and compare their performances theoretically and using computer simulation.

- We avoid the difficulty of tuning the Bayesian detector by using the NP and CNP detectors, where the tunable parameters are probability terms.

2.2.4 Preprocessing

All VADs use a preprocessing stage which is generally used to perform primitive tasks such as dividing the signal into frames with or without overlapping, windowing the signal etc. The choice of the frame duration is largely dependent upon the application, where real-time applications like the VoIP systems use a short frame duration of 10ms or 20 ms and applications which do not have real-time constraints may use longer frame lengths [1]. Also, depending on the VAD algorithm being used, the preprocessing stage may transform the incoming time signal into the Fourier domain, estimate the periodogram [16] or compute the cepstrum [33]. The preprocessor may also be used to perform more complex functions such as in the VAD developed by Tucker where the speech signal is preprocessed to detect and remove any kind of periodic interference [34].

2.2.5 Post-processing

The post-processing stage generally employs a decision corrective stage which attempts to boost the VAD performance. The primary function of the post-processing schemes is to reduce the risk of misdetecting a low energy portion of speech at the beginning or end of an utterance as pause [16]. The post-processing schemes achieve lower misses by delaying the transitions of decisions from speech to pause and readily transiting from a pause to speech decision.

Among the common systems employed for post-processing are the binary Markov models which implement speech and pause as different states, and give soft voice activity decisions in terms of probabilities [19]. A similar scheme based on the first order Markov process was suggested by Sohn et. al. [15] in their statistical VAD

scheme. The hang over scheme by Davis and Nordholm is another interesting implementation of a post-processor where a state machine is traversed as per the decisions made by the VAD algorithm and the position on the state machine gives the final VAD decision [16].

Most post-processing schemes implicitly model the duration of speech and pause. However, Markov model based post-processors assume that speech/pause durations are geometrically distributed which is not a good model of duration. On the other hand, a recent study suggests that the Poisson density is a more accurate representation of the distribution of speech duration [39]. In this thesis, we attempt to develop a post-processor in form of a contextual detector which processes the durational information of speech and pause to render decisions. Unlike the conventional post-processors which process the durational information on a frame by frame basis, the contextual detector processes the durational information in a group of similar decisions which we call as the same state periods (SSPs). Hence, the contextual detector treats the entire speech burst or pause duration as one entity, and computes the likelihood of activity of the current frame based on the past SSP durations. Moreover, the handling of frame decisions at a SSP level gives the added advantage of enforcing well known facts that speech and pause durations have lower bounds [27].

Chapter 3

Proposed Voice Activity Detectors

A VAD scheme segments the input signal into contiguous frames and then classifies each frame as active or inactive. Most VAD algorithms seek voice activity cues within the frame in question itself - primary cue, i.e., the algorithms compute some form of short term statistics for detection. However, many speech events such as the acoustic similarity of an aspiration in a stop to pause and fricatives to noise, present ambiguous primary cues and lead to misdetections. Naturally, complete dependency on primary cues results in poor detection and obviates the need for a scheme which accommodates the contextual speech cues or long term statistics into the final VAD decision. This motivates the design of our statistical VAD scheme which detects and combines the voice activity information present in the short-term and the long-term statistics of the speech signal using the primary and contextual detectors.

In this chapter, we propose three primary detectors based on the PEM model, using the Bayesian, Neyman-Pearson (NP) and competitive NP (CNP) design approaches. Further, we analyze and compare the proposed primary detectors, and reveal the following new findings: (i) the sufficient statistics (SS) of the proposed primary detectors is a speech energy estimator that is a function of the prior SNR of

the noisy speech signal, (ii) the PEM model for voice activity detection can be viewed as a composite hypothesis testing problem with the prior SNR acting as the free parameter, (iii) the NP approach assumes that the hypothesis and the free parameter are uncorrelated, whereas the CNP approach exploits the correlation between the hypothesis and the free parameter, and models their interdependency into the detector design. For the voice activity detection problem, since an intuitive relationship between the hypothesis and the prior SNR (free parameter) exists, we assert that the CNP detector should perform better than the NP as it makes use of the partial prior information about the prior SNR. This assertion is confirmed by our simulation results which shows that the performance of the CNP detector is consistently better than that of the NP and Bayesian detectors at low SNRs.

Next, we propose a contextual detector which processes the long-term information in the speech signal using the durational information in speech bursts and pause periods. Unlike, traditional hang-over schemes, the contextual detector does not process the individual decisions of the primary detector, but groups the speech and pause decisions to form speech bursts and pause periods, respectively. Subsequently, the statistics of the speech bursts and pause periods are used to estimate the voice activity of a frame in the form of a LR. Finally, the contextual and primary LRs are combined to obtain the comprehensive VAD (CVAD) scheme which gives the final VAD decision. Particularly, the combination of the Bayesian, NP and CNP primary detectors with the contextual detector gives the CVAD-Bayesian, CVAD-NP and CVAD-CNP schemes, respectively.

The organization of this chapter is as follows: In Sec. 3.1, we discuss the architecture of the proposed CVAD scheme using a block diagram. In Sec. 3.2, we develop a Bayesian detector using the proposed PEM model, and show that the SS acts as a speech energy estimator. We also derive some expressions for the conditional

statistics of the SS based on which the NP and CNP detectors are developed in the later sections. In Sec. 3.3, we analyze the CNP and NP approaches and show the superiority of the CNP over the NP as a more generalized approach. In Sec. 3.4, we develop the NP and CNP detectors, and carry out a theoretical comparison between the proposed VADs in terms of the probabilities of false-alarm and miss-detection. Finally, in Sec. 3.5, we develop the contextual detection scheme and in Sec. 3.6, we combine the primary detectors with the contextual detector to form the CVADs.

3.1 Proposed VAD Scheme

The proposed CVAD scheme is illustrated in Fig. 8, where the functional relationship between the primary and contextual detector is described diagrammatically. The CVAD scheme first divides the input signal into frames F_i, where the frame F_{i+1} precedes F_i in time. Next, the primary detector gives a decision D_i for each frame F_i, and the CVAD system stores the decisions D_i along with the primary LR, $\Lambda^P(D_i)$. Note that the decision D_i for the frame F_i and its primary likelihood $\Lambda^P(D_i)$ are based on the short-term statistics of the speech signal alone. Further, the VAD system forms a decision history consisting of the most recent m past decisions, $D_{i+1}, D_{i+2}, ..., D_{i+m}$ for each decision D_i where the decision history directly translates into speech and pause durational information. Figure 8 shows the decision history of D_0 only. Now, the contextual detector processes the decision history of the frame F_i and generates a contextual likelihood $\Lambda^C(D_i)$ of voice activity. Finally, the CVAD scheme combines the primary $\Lambda^P(D_i)$ and contextual likelihoods $\Lambda^C(D_i)$ to obtain the overall likelihood $\Lambda(D_i)$ of voice activity. The final decision FD_i for the frame F_i is taken by comparing the overall likelihood $\Lambda(D_i)$ with a threshold.

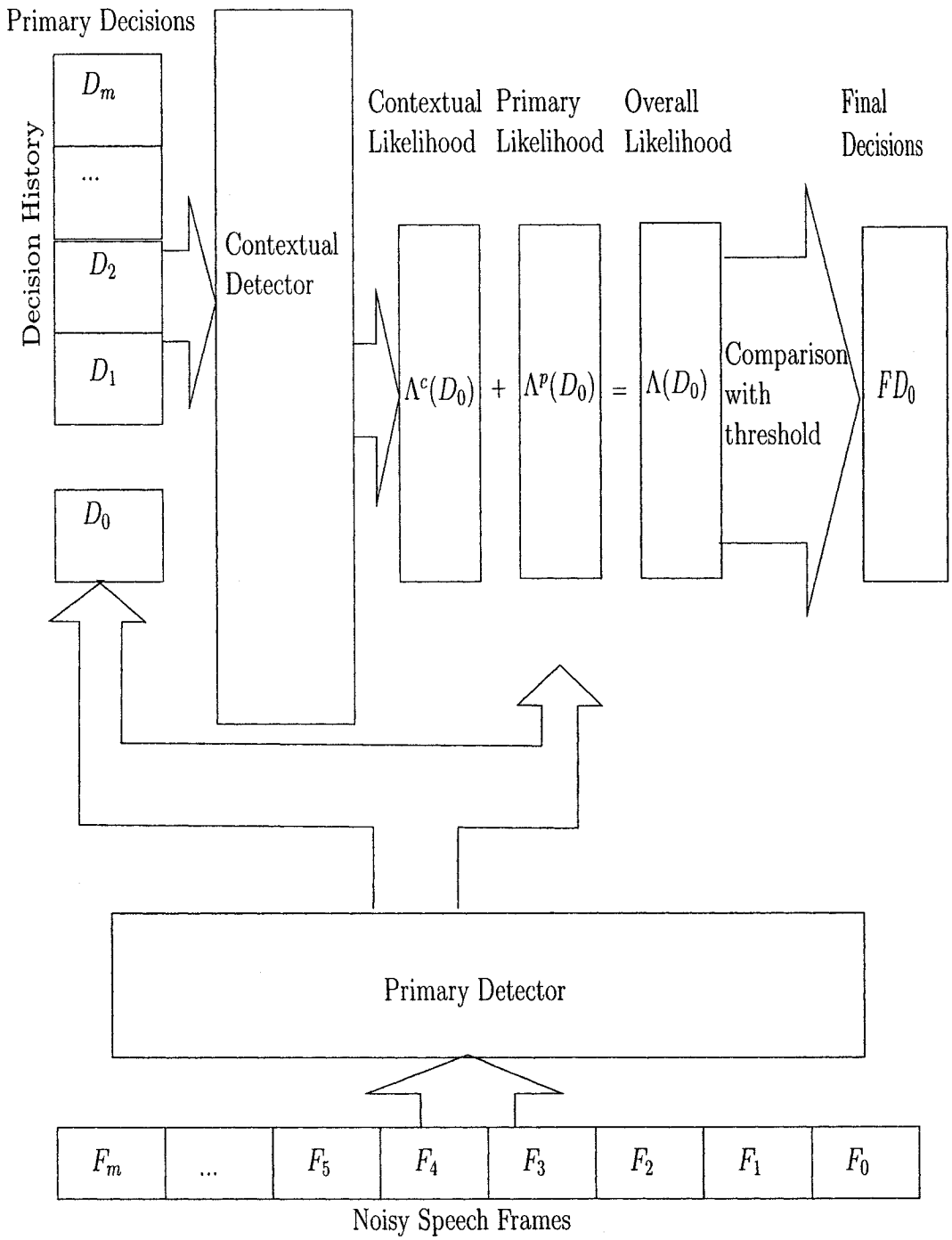

Figure 8: Block diagram illustrating the operation of the proposed CVAD scheme.

3.2 Bayesian Detector using PEM Model

The VAD problem can be modeled as a binary hypothesis, i.e.,

$$\begin{aligned} H_0 &: \text{pause}, \\ H_1 &: \text{speech}, \end{aligned} \tag{2}$$

where the input signal is first segmented into contiguous frames and then each frame is assigned to H_1 or H_0, if it is detected as speech or pause, respectively. In order to develop a Bayesian detector, we use the PEM model, which assumes that MFSCs are mutually statistically independent and zero-mean Gaussian random variables [21].

Let $\mathbf{S}$ and $\mathbf{N}$ denote respectively, one frame of speech and that of noise, i.e.,

$$\mathbf{S} = [s_1, s_2, ..., s_M], \tag{3}$$

$$\mathbf{N} = [n_1, n_2, ..., n_M], \tag{4}$$

where s_i and n_i are the i^{th} speech and noise MFSCs, respectively. Using the PEM model, the binary hypothesis problem in (2) can be rewritten in terms of the MFSCs of speech $\mathbf{S}$ and noise $\mathbf{N}$ as:

$$\begin{aligned} H_0 &: \mathbf{F} = \mathbf{N}, \\ H_1 &: \mathbf{F} = \mathbf{S} + \mathbf{N}, \end{aligned} \tag{5}$$

where $\mathbf{F}$ is the observation frame in the M-dimensional space $\mathbf{D}^{\mathbf{F}}$ as given by:

$$\mathbf{F} = [f_1, f_2, ..., f_M]. \tag{6}$$

Then, a likelihood ratio test (LRT) can be obtained using the definition of the likelihood function and the Gaussian probability density function as [23]:

$$\frac{\mathbf{F} \times (\mathbf{K_n}^{-1} - \mathbf{K_f}^{-1}) \times \mathbf{F}^T}{2} \underset{<H_0}{\overset{\geq H_1}{}} \ln(\eta) - \frac{1}{2} \times \ln \frac{|\mathbf{K_n}|}{|\mathbf{K_f}|}, \tag{7}$$

where η is the Bayesian threshold, $|.|$ the determinant, and $\mathbf{K_f}$ and $\mathbf{K_n}$ are the covariance matrix of the noisy speech $(\mathbf{S} + \mathbf{N})$ and that of noise $(\mathbf{N})$. The above equation represents a Bayesian detector where the left hand side (LHS) is the SS denoted by l, and the right hand side (RHS) is the overall threshold of the test which is denoted by γ.

In the following subsections, we show a new interpretation for the SS in (7) as a speech energy estimator where the value of the SS can be computed using the estimates of the prior SNR. Further, we establish that if the binary hypothesis in (5) is treated as a composite hypothesis problem [23], then the prior SNR estimates become free parameters of the hypothesis which is an important result for the design of the CNP detector. Lastly, we derive a few expressions pertaining to the SS which will be used later to design the CNP and NP detectors in Sec. 3.4.

3.2.1 Sufficient Statistics as a Speech Energy Estimator

In order to show that the SS in (7) is a speech energy estimator, we first apply a nonsingular transform matrix $\mathbf{Q}$ to simultaneously diagonalize $\mathbf{K_n}$ and $\mathbf{K_f}$, i.e.,

$$\mathbf{Q}^T \mathbf{K_n} \mathbf{Q} = \mathbf{I}, \tag{8}$$

$$\mathbf{Q}^T \mathbf{K_f} \mathbf{Q} = \mathbf{\Lambda}, \tag{9}$$

where $\mathbf{I}$ is the identity matrix, and $\mathbf{\Lambda}$ is a diagonal matrix whose i^{th} diagonal element is given by λ_i, i.e.,

$$\mathbf{\Lambda} = \begin{bmatrix} \lambda_1 & 0 & ... & 0 \\ 0 & \lambda_2 & ... & 0 \\ \vdots & & \ddots & \vdots \\ 0 & 0 & ... & \lambda_M \end{bmatrix} \tag{10}$$

It is well known that such a transform $\mathbf{Q}$ exists [40], and the technique to obtain $\mathbf{Q}$ is given in Appendix A. Now, let $\mathbf{Z} = [z_1, z_2, ..., z_M]$ be the transformed observation in space $\mathbf{D^Z}$, namely:

$$\mathbf{Z} = \mathbf{FQ}, \tag{11}$$

or

$$\mathbf{F} = \mathbf{ZQ}^{-1}. \tag{12}$$

Using (12) into the LHS of (7), we get an equivalent expression for the SS as:

$$l = \frac{\mathbf{Z} \times (\mathbf{Q}^{-1}{\mathbf{K_n}}^{-1}(\mathbf{Q}^{-1})^T - \mathbf{Q}^{-1}{\mathbf{K_f}}^{-1}(\mathbf{Q}^{-1})^T) \times \mathbf{Z}^T}{2}. \tag{13}$$

From (8) and (9), we get:

$$\mathbf{Q}^{-1}{\mathbf{K_n}}^{-1}(\mathbf{Q}^{-1})^T = \mathbf{I}, \tag{14}$$

$$\mathbf{Q}^{-1}{\mathbf{K_f}}^{-1}(\mathbf{Q}^{-1})^T = \mathbf{\Lambda}^{-1}. \tag{15}$$

and using these two identities, (13) can be simplified as,

$$l = \frac{\mathbf{Z} \times (\mathbf{I} - \mathbf{\Lambda}^{-1}) \times \mathbf{Z}^T}{2}, \tag{16}$$

$$= \frac{1}{2} \sum_{i=1}^{m} (1 - \frac{1}{\lambda_i}) z_i^2, \tag{17}$$

where z_i is the i^{th} element of the transformed observation vector $\mathbf{Z}$. In order to relate the SS with the prior SNR of the noisy speech, we use the definitions of the prior and posterior SNRs given by Ephraim and Malah [22], i.e., the i^{th} posterior SNR ($\gamma_i^{\mathbf{F}}$) is the ratio of the i^{th} noisy speech variance (σ_{fi}^2) to the i^{th} noise variance (σ_{ni}^2),

$$\gamma_i^{\mathbf{F}} = \frac{\sigma_{fi}^2}{\sigma_{ni}^2}, \tag{18}$$

while the i^{th} prior SNR ($\zeta_i^{\mathbf{F}}$) is defined by:

$$\zeta_i^{\mathbf{F}} = \gamma_i^{\mathbf{F}} - 1. \tag{19}$$

We now extend the above definitions of the prior and posterior SNR to $\mathbf{D}^{\mathbf{Z}}$ as:

$$\gamma_i^{\mathbf{Z}} = \lambda_i, \tag{20}$$

$$\zeta_i^{\mathbf{Z}} = \lambda_i - 1, \tag{21}$$

where $\gamma_i^{\mathbf{Z}}$ and $\zeta_i^{\mathbf{Z}}$ are the counterparts of $\gamma_i^{\mathbf{F}}$ and $\zeta_i^{\mathbf{F}}$, i.e., the posterior and prior SNR in $\mathbf{D}^{\mathbf{Z}}$. Using (20) and (21), the SS in (17) can be rewritten as:

$$l = \frac{1}{2} \sum_{i=1}^{M} (\frac{\zeta_i^{\mathbf{Z}}}{\zeta_i^{\mathbf{Z}} + 1}) z_i^2. \tag{22}$$

Now, we show that the SS is actually a speech energy estimator. Note that if the

speech component in the signal dominates along the i^{th} vector, namely, the i^{th} prior SNR ($\zeta_i^{\mathbf{Z}}$) is high, then,

$$\zeta_i^{\mathbf{Z}} >> 1, \tag{23}$$

or

$$(\frac{\zeta_i^{\mathbf{Z}}}{\zeta_i^{\mathbf{Z}}+1}) \rightarrow 1, \tag{24}$$

implying that the instantaneous signal energy (z_i^2) along the i^{th} vector is fully passed by the detector. On the other hand, if noise is more dominant along the i^{th} vector, then

$$\zeta_i^{\mathbf{Z}} \rightarrow 0, \tag{25}$$

or

$$(\frac{\zeta_i^{\mathbf{Z}}}{\zeta_i^{\mathbf{Z}}+1}) \rightarrow 0, \tag{26}$$

implying that the corresponding instantaneous energy component is removed from the detector. In general, the SS determines the proportion of speech along the i^{th} vector on the basis of the estimate of the i^{th} prior SNR $\zeta_i^{\mathbf{Z}}$ and weighs the signal energy of the noisy speech appropriately to obtain an estimate of the speech energy along the i^{th} dimension. Hence, the role of the SS as a speech energy estimator is justified.

The above result is very interesting as it gives a common platform to compare the PEM based VAD with the traditional energy based VAD schemes. It is clear from the above analysis that the PEM VAD attempts to divide the signal subspace into speech and noise subspaces. Subsequently, it relies heavily on the speech subspace for detection while downplaying the role of the noise subspace. This is similar in function to the traditional frequency-domain based energy VADs which split the speech spectrum into sub-bands. However, unlike the traditional scheme which employs heuristics to

divide the spectra, the PEM VAD learns the subspaces from the data itself. Hence, the PEM VAD is adaptable to different noises and SNRs, obviating its superiority over the traditional scheme.

3.2.2 Properties of the Sufficient Statistics

In this subsection, we derive certain conditional statistics of the SS which are necessary in the development and analysis of the proposed NP and CNP detectors. Particularly, we derive the conditional means, $E[l|H_0]$ and $E[l|H_1]$, and conditional variance, $Var[l|H_0]$ of the SS which will be used later in computing the probability of false-alarm (P_f) and that of miss-detection (P_m) for deriving the NP detector as well as determining the theoretical performance of the proposed detectors. We also deduce a formula for the normalized distance term d defined in [23] as:

$$d = \sqrt{\frac{(E[l|H_1] - E[l|H_0])^2}{Var[l|H_0]}}, \tag{27}$$

which will be used in the computation of the CNP threshold. In order to derive the above mentioned statistical quantities, we first present the following theorem:

Theorem 3.2.1. *Let the M-dimensional random vector* $\mathbf{F}$ *be drawn from one of two different zero-mean Gaussian distributions, where* H_0 *and* H_1 *are the events that* $\mathbf{F}$ *is chosen from the first and the second Gaussian distributions, respectively, i.e.,*

$$H_0 : \mathbf{F} \sim \mathcal{N}(\mathbf{0}, \mathbf{K_n}), \tag{28}$$

$$H_1 : \mathbf{F} \sim \mathcal{N}(\mathbf{0}, \mathbf{K_f}). \tag{29}$$

Then the sufficient statistics,

$$l = \frac{\mathbf{F} \times (\mathbf{K_n}^{-1} - \mathbf{K_f}^{-1}) \times \mathbf{F}^T}{2}. \tag{30}$$

has the following properties:

(i) The conditional mean of l given H_0, $E[l|H_0]$ is given by:

$$E[l|H_0] = \frac{1}{2}\sum_{i=1}^{M}(\frac{\zeta_i^{\mathbf{F}}}{\gamma_i^{\mathbf{F}}}). \tag{31}$$

(ii) The conditional mean of l given H_1, $E[l|H_1]$ is given by:

$$E[l|H_1] = \frac{1}{2}\sum_{i=1}^{M}\zeta_i^{\mathbf{F}}. \tag{32}$$

(iii) The conditional variance of l given H_0, $Var[l|H_0]$ is given by:

$$Var[l|H_0] = \frac{1}{2}\sum_{i=1}^{M}(\frac{\zeta_i^{\mathbf{F}}}{\gamma_i^{\mathbf{F}}})^2. \tag{33}$$

(iv) The value of the statistical quantity d defined in (27) is given by:

$$d = \sqrt{\frac{(\sum_{i=1}^{M}\frac{(\zeta_i^{\mathbf{F}})^2}{\gamma_i^{\mathbf{F}}})^2}{2\sum_{i=1}^{M}(\frac{\zeta_i^{\mathbf{F}}}{\gamma_i^{\mathbf{F}}})^2}} \tag{34}$$

where $\gamma_i^{\mathbf{F}}$ and $\zeta_i^{\mathbf{F}}$ are the posterior and prior SNR as defined in (18) and (19), respectively.

The proof for Theorem 3.2.1 is given in Appendix B. The distance measure d given by (27) is indicative of the separability of the conditional probability distributions,

$p(l|H_0)$ and $p(l|H_1)$, in terms of the first and second order conditional statistics of the SS, and depends purely on the value of the prior SNRs as seen from (34). As expected, the higher the prior SNR, the larger the value of d, and the easier the task of voice activity detection is.

3.2.3 Behavior of the Bayesian Threshold γ

The behavior of the threshold γ with changing prior SNR needs to be established in order to comprehend the LRT completely. Using (8), (9), (20) and (21) the threshold γ in (7) can be simplified as:

$$\gamma = \ln(\eta) + \frac{1}{2}\ln\frac{|\mathbf{K_f}|}{|\mathbf{K_n}|}, \tag{35}$$

$$= \ln(\eta) + \frac{1}{2}\ln\frac{|\mathbf{\Lambda}|}{|\mathbf{I}|}, \tag{36}$$

$$= \ln(\eta) + \frac{1}{2}\sum_{i=1}^{M}\ln(\zeta_i^{\mathbf{Z}} + 1), \tag{37}$$

Now, it is observed that like the SS, the threshold term too is a function of the prior SNR which clearly shows that the prior SNR is the free parameter of the composite hypothesis. The threshold γ in (37) consists of two terms with distinct behaviors, i.e.,

$$\gamma' = \underbrace{\ln\eta}_{\text{fixed}} + \underbrace{\frac{1}{2}\sum_{i=1}^{M}\ln(\zeta_i^{\mathbf{Z}} + 1)}_{\text{variable}}. \tag{38}$$

where the variable term of the threshold γ consists of a contribution from each of the M dimensions. Using (31) and (32), it is easy to show that the variable component of the threshold along each dimension is always bounded by the conditional means of

SS [see Appendix C for details], i.e.,

$$E[l|H_0] \leq \frac{1}{2}\sum_{i=1}^{M} \ln(\zeta_i^{\mathbf{z}} + 1) \leq E[l|H_1]. \tag{39}$$

Now, it is easy to see that fixed part $\ln(\eta)$ of γ acts as a bias term by pushing the threshold towards one hypothesis and biasing the detector in favor of the other hypothesis. However, it is useful to note that for a given η, γ cannot selectively bias the detector towards different hypothesis for different values of the prior SNR. As shown later, this is the main reason for the inability of the Bayesian detector to incorporate prior information about the prior SNR.

3.3 On the Neyman-Pearson and the Competitive Neyman-Pearson Approaches

In this section, we analyze the CNP and NP approaches [23,26], and show that unlike the NP, the CNP approach exploits the prior information about the free parameter (the prior SNR) of the composite hypothesis in the detector design, making the CNP approach a very useful design technique for voice activity detectors. The partial prior information about the prior SNR (the free parameter) can be interpreted as a general relation between the prior SNR and the hypothesis, i.e., higher values of prior SNR are more likely to be associated with H_1 than H_0 and vice-versa. This observation can be justified through the following analysis.

We first briefly review the composite hypothesis problem, and the NP and CNP approaches. Define a prior SNR vector ζ as:

$$\zeta = [\zeta_1, \zeta_2, ...\zeta_M] \in \mathbf{\Psi_i}. \tag{40}$$

where $\mathbf{\Psi_i}$ is the parameter set consisting of prior SNR values generated when H_i is true. Let $\mathbf{\Psi} = \mathbf{\Psi_0} \cup \mathbf{\Psi_1}$ be the parameter space. It is interesting to note that for voice activity detection, an exclusive partition of the space $\mathbf{\Psi}$ into $\mathbf{\Psi_0}$ and $\mathbf{\Psi_1}$ is in general not obtainable, and it is reasonable to assume a complete overlap between $\mathbf{\Psi_0}$ and $\mathbf{\Psi_1}$. This observation is important as it renders the generalized LRT (GLRT) approach towards hypothesis testing meaningless in voice activity detection [26]. The overlap between $\mathbf{\Psi_0}$ and $\mathbf{\Psi_1}$ is mainly a result of the non-stationarity of the underlying speech and noise processes.

The two types of errors in binary hypothesis testing are measured by P_f and P_m [23], i.e.,

$$\begin{aligned} P_f(\gamma) &\triangleq p(F \in H_1|H_0), \\ &= \int_{\gamma}^{\infty} p(l|H_0)dl, \end{aligned} \tag{41}$$

$$\begin{aligned} P_m(\gamma) &\triangleq p(F \in H_0|H_1), \\ &= \int_{-\infty}^{\gamma} p(l|H_1)dl, \end{aligned} \tag{42}$$

where $p(l|H_0)$ and $p(l|H_1)$ are the conditional pdfs of the SS. The total error probability P_e is given as a combination of P_f and P_m, i.e.,

$$P_e(\gamma) \triangleq P(H_0)P_f(\gamma) + P(H_1)P_m(\gamma). \tag{43}$$

Now, the NP approach determines a γ which minimizes P_m while constraining P_f by a constant upper bound, i.e.,

$$\begin{aligned} &\min_{\gamma} \ P_m(\gamma), \\ &P_f(\gamma) \leq \lambda, \end{aligned} \tag{44}$$

where λ is a constant. On the other hand, the competitive Neyman-Pearson (CNP) approach also minimizes P_m but constrains P_f with a variable upper bound which is a function of ζ [26], i.e.,

$$\begin{aligned} &\min_{\gamma} \; P_m(\gamma), \\ &P_f(\gamma) \leq \lambda(\zeta). \end{aligned} \tag{45}$$

Now, we analyze the NP and CNP approaches by expressing the prior probabilities $P(H_i)$ in terms of the prior SNR as:

$$P(H_i) = \int_{\Psi_i} P(H_i|\zeta)p(\zeta)d\zeta,$$

where $p(\zeta)$ is the pdf of ζ. Using the Bayes rule, the above expression can be rewritten as:

$$P(H_i) = \int_{\Psi_i} p(\zeta|H_i)P(H_i)d\zeta.$$

Using the above expression along with (41) and (42), into (43), we get:

$$\begin{aligned} P_e(\gamma) &= \sum_{i=0,1} \int_{\Psi_i} p(\zeta|H_i)P(H_i)d\zeta \int_{L_i} p(l|H_i,\zeta)dl, \\ &= \sum_{i=0,1} P(H_i) \int_{L_i}\int_{\Psi_i} p(l|H_i,\zeta)p(\zeta|H_i)d\zeta dl. \end{aligned} \tag{46}$$

where $L_1 = \{l : \infty > l \geq \gamma\}$, $L_0 = \{l : -\infty < l < \gamma\}$.Now, if the probability law governing the generation of ζ from the source, $p(\zeta|H_i)$, is completely known, then the composite hypothesis problem can be easily reduced to a simple hypothesis problem. However, in voice activity detection $p(\zeta|H_i)$ is not explicitly known and a straightforward design for hypothesis testing is not possible as the parameter ζ cannot be removed via the integration in (46). Alternatively, one could work with a

conditional error term which can be easily obtained from (46), i.e.,

$$\begin{aligned} P_e(\gamma) &= \sum_{i=0,1} P(H_i) \int_{L_i} \int_{\Psi_i} p(l|H_i,\zeta) p(\zeta|H_i) \frac{p(\zeta)}{p(\zeta)} d\zeta dl, \\ &= \int_{\Psi_i} \{ \sum_{i=0,1} P(H_i) \int_{L_i} p(l|H_i,\zeta) \frac{p(\zeta|H_i)}{p(\zeta)} dl \} p(\zeta) d\zeta, \end{aligned}$$

where the term inside the parenthesis $\{.\}$ represents the conditional error term $P_e(\gamma|\zeta)$, i.e.,

$$\begin{aligned} P_e(\gamma|\zeta) &= \sum_{i=0,1} P(H_i) \int_{L_i} p(l|H_i,\zeta) \frac{p(\zeta|H_i)}{p(\zeta)} dl, \\ &= \sum_{i=0,1} P(H_i) \frac{p(\zeta|H_i)}{p(\zeta)} \int_{L_i} p(l|H_i,\zeta) dl, \\ &= \frac{P(H_1) p(\zeta|H_1) P_m(\gamma|\zeta)}{p(\zeta)} + \frac{P(H_0) p(\zeta|H_0) P_f(\gamma|\zeta)}{p(\zeta)}. \end{aligned}$$

In the RHS of the above expression, the first and second terms represent the contributions to the overall error due to the conditional miss-detection and false-alarm, respectively. If we minimize the error due to the first error term while constraining the second error term by a constant value λ, we get:

$$\min_{\gamma} \frac{P(H_1) p(\zeta|H_1) P_m(\gamma|\zeta)}{p(\zeta)}, \tag{47}$$

$$\frac{P(H_0) p(\zeta|H_0) P_f(\gamma|\zeta)}{p(\zeta)} \leq \lambda. \tag{48}$$

Clearly, the terms in (47) which do not contain γ can be removed from the minimization. Further, the inequality (48) can be rewritten as:

$$P_f(\gamma|\zeta) \leq \frac{\lambda p(\zeta)}{P(H_0) p(\zeta|H_0)}, \tag{49}$$

Thus, the constrained minimization problem described in (47) and (48) can be rewritten in a simplified form as:

$$\min_{\gamma} \; P_m(\gamma|\zeta), \tag{50}$$

$$P_f(\gamma|\zeta) \leq \lambda' \frac{p(\zeta)}{p(\zeta|H_0)}, \tag{51}$$

where $\lambda' = \frac{\lambda}{P(H_0)}$ is a constant. Now, if ζ and H_i are uncorrelated, then $p(\zeta|H_0) = p(\zeta)$ and the RHS of (51) reduces to λ' only, yielding a constraint condition similar to the NP approach given in (44). On the other hand, if there is a correlation between ζ and H_i, then the RHS of (51) becomes a function of ζ, i.e., $\lambda'(\zeta)$, resulting in a constraint similar to the CNP approach given in (45). Hence, it is easily seen that the CNP approach uses the correlation between the parameter and the hypothesis by setting an upper bound for P_f as a function of the parameter itself. The ability of the CNP approach to model the prior information about the free parameter (whenever the information exists) into the detector design must lead to a better performance over the NP detector.

Just as P_f tunes the NP detector, it can be seen from (45) that the functional relationship between P_f and ζ is the tuning parameter of the CNP detector. In other words, different functional relationships between P_f and ζ give different operating points for the CNP detector. Therefore, implementing the CNP detector via NP is cumbersome as all the different functional relationships need to be determined and stored. This difficulty can be avoided by following an alternative strategy to design the CNP detector as described below. Let us define a probability term P_a related to P_f as:

$$P_a(\gamma'|\zeta) \;\triangleq\; \int_{\gamma'}^{\infty} p(l|H_0, \zeta) dl,$$

$$= \int_{\gamma'}^{\gamma} p(l|H_0,\zeta)dl + \int_{\gamma}^{\infty} p(l|H_0,\zeta)dl,$$
$$= \int_{\gamma'}^{\gamma} p(l|H_0,\zeta)dl + P_f(\gamma|\zeta), \quad (52)$$

In the above expression, if P_a is kept constant, then $P_f \leq P_a \ \forall\zeta$ when $\gamma' \leq \gamma$, and $P_f > P_a \ \forall\zeta$ where $\gamma' > \gamma$. Hence, a designer can appropriately vary P_f by suitably adjusting the distance between γ' and γ. If one maintains $\gamma' \leq \gamma$, then P_f is always bounded above and similarly, one can set lower bounds for P_f (or equivalently upper bounds on P_m) by maintaining $\gamma' > \gamma$. Hence, using this strategy the design of a CNP detector boils down to determining an appropriate curve γ', where P_a becomes the only tunable parameter of the detector. It is worth mentioning that the results developed in this section are not tied to voice activity detection but can be applied to the general composite hypothesis testing problem.

3.4 NP and CNP Detectors

We would first like to investigate the probability of false-alarm, P_f for the proposed Bayesian detector in (7), based on which the new NP and CNP detectors are developed. We also derive an expression for P_m, and then use both P_f and P_m to evaluate the theoretical performance of the Bayesian, NP and CNP detectors.

3.4.1 Probability of False-Alarm P_f

In order to obtain P_f, we need to first determine the conditional probability density $p(l|H_0)$ which is presented in the following theorem:

Theorem 3.4.1. *Let l be the SS given by (17). If z_i is a normal random variable $\forall$ $i = 1, 2, ..., M$, then the pdf of l is Gaussian for a large M.*

Proof. The random variable l given by (17) can be rewritten as:

$$l = \frac{1}{2}\sum_{i=1}^{m}(1 - \frac{1}{\lambda_i})z_i^2 = \frac{1}{2}\sum_{i=1}^{m}\kappa_i z_i^2 \tag{53}$$

where $\kappa_i = 1 - \frac{1}{\lambda_i}$ is the i^{th} weight factor. As z_i is normal, the distribution of z_i^2 is chi-square (χ^2) with one degree of freedom. Further, since $z_1, z_2, ..., z_m$ are independent and identically distributed (i.i.d), the variables $z_1^2, z_2^2, ..., z_m^2$ are i.i.d too. Now, l is a weighted sum of many i.i.d random variables z_i^2. Therefore, using the central limit theorem, l can be expected to approach the Gaussian distribution provided that M is sufficiently large. □

Using the Theorem 3.4.1, the conditional probability of SS, $p(l|H_0)$, is Gaussian for a large frame size M since the observations z_i are normal given the hypothesis H_0. Note that, the mean ($E[l|H_0]$) and variance ($Var[l|H_0]$) are given by (31) and (33), respectively. Therefore, the expression for P_f in (41) can be written as:

$$\begin{aligned} P_f &= \frac{1}{\sqrt{2\pi Var[l|H_0]}}\int_{\gamma}^{\infty}\exp\left(-\frac{(l - E[l|H_0])^2}{2Var[l|H_0]}\right)dl, \\ &= 1 - erf\left(\frac{\gamma - E[l|H_0]}{\sqrt{Var[l|H_0]}}\right). \end{aligned} \tag{54}$$

where $erf(.)$ is the standard error function defined as:

$$erf(x) \triangleq \int_{-\infty}^{x}\frac{\exp(-x^2/2)}{\sqrt{2\pi}}. \tag{55}$$

3.4.2 Neyman-Pearson Detector

Using (54), the NP threshold γ_{NP} can be determined by constraining $P_f = \alpha$ and solving for γ:

$$\gamma = \sqrt{Var[l|H_0]}erf^{-1}(1-\alpha) + E[l|H_0]. \tag{56}$$

The RHS of (56) represents γ_{NP} where $erf^{-1}(.)$ is the inverse of the standard error function. The NP detector can be easily developed by substituting (56) into (7), i.e.,

$$l \underset{<H_0}{\overset{\geq H_1}{}} \gamma_{NP}, \tag{57}$$

where P_f or α is the tunable parameter of the detector.

3.4.3 Competitive Neyman-Pearson Detector

We now develop the CNP detector using the method described in Sec. 3.3 for which the following expression is proposed for γ',

$$\gamma' = \sqrt{Var[l|H_0]}(\frac{\ln \eta}{S(\bar{\zeta}) \times d} + \frac{d}{2}) + E[l|H_0], \tag{58}$$

where d is the normalized distance given by (34) and $S(\bar{\zeta})$ is given by:

$$S(\bar{\zeta}) = 2 - \frac{2}{1+\exp(-\bar{\zeta})}. \tag{59}$$

It is interesting to note that $S(\bar{\zeta})$ is similar to the sigmoid function where the parameter $\bar{\zeta}$ represents the average prior SNR as given by:

$$\bar{\zeta} = \frac{1}{M}\sum_{i=1}^{M}\zeta_i^{\mathbf{F}}. \tag{60}$$

The expression for the curve γ' in (58) consists of two distinct terms which impact the curve in different ways. These terms are obtained by simplifying (58):

$$\gamma' = \underbrace{\frac{\sqrt{Var[l|H_0]}\ln(\eta)}{S(\bar{\zeta})\times d}}_{\text{bias term}} + \underbrace{\frac{\sqrt{Var[l|H_0]}d}{2} + E[l|H_0]}_{\text{optimizing term}}. \tag{61}$$

The optimizing term in (61) ensures that the curve γ' is equidistant from the conditional likelihoods $E[l|H_0]$ and $E[l|H_1]$. This can be shown very easily by simplifying the expression for the optimizing term using the definition of d in (27):

$$\frac{\sqrt{Var[l|H_0]}\times d}{2} + E[l|H_0] = \frac{(E[l|H_1]-E[l|H_0])}{\sqrt{Var[l|H_0]}}\frac{\sqrt{Var[l|H_0]}}{2} + E[l|H_0] \tag{62}$$

$$= \frac{(E[l|H_1]-E[l|H_0])}{2} + E[l|H_0] \tag{63}$$

$$= \frac{(E[l|H_1]+E[l|H_0])}{2} \tag{64}$$

where the term in (64) lies exactly in between the two mean-likelihoods $E[l|H_0]$ and $E[l|H_1]$, irrespective of the value of the prior SNR. The first term in (61) is the biasing term which controls the bias of the detector towards the hypothesis.

Using (58) into (52) and setting $P_a = \alpha$, we get,

$$\alpha = 1 - erf\left(\frac{\ln(\eta)}{S(\bar{\zeta})\times d} + \frac{d}{2}\right), \tag{65}$$

which is equivalent to:

$$\ln(\eta) = d \times S(\bar{\zeta})(erf^{-1}(1-\alpha) - \frac{d}{2}). \tag{66}$$

By subtracting the term $\frac{1}{2}\ln\frac{|\mathbf{K_n}|}{|\mathbf{K_f}|}$ from both sides of the above equation, we get:

$$\ln(\eta) - \frac{1}{2}\ln\frac{|\mathbf{K_n}|}{|\mathbf{K_f}|} = d \times S(\bar{\zeta})(erf^{-1}(1-\alpha) - \frac{d}{2}) - \frac{1}{2}\ln\frac{|\mathbf{K_n}|}{|\mathbf{K_f}|},$$

that is,

$$\gamma = \gamma_{CNP}. \tag{67}$$

Now, the CNP detector is given by:

$$l \underset{H_0}{\overset{H_1}{\gtrless}} \gamma_{CNP}. \tag{68}$$

Note that there is only one tunable parameter P_a (or α) of the detector.

3.4.4 Probability of Miss-Detection P_m

In order to compute P_m in (42), we need to determine $p(l|H_1)$. By defining a transform $\mathbf{Q}'$ which simultaneously diagonalizes $\mathbf{K_f}$ and $\mathbf{K_n}$, i.e.,

$$\mathbf{Q}'^T\mathbf{K_n}\mathbf{Q}' = \mathbf{\Lambda}', \tag{69}$$

$$\mathbf{Q}'^T\mathbf{K_f}\mathbf{Q}' = \mathbf{I}, \tag{70}$$

where $\mathbf{\Lambda}'$ is a diagonal matrix whose i^{th} diagonal element is given by λ_i', and following the steps similar to those in (11) to (17), we obtain a scalar form for the SS as:

$$l = \frac{1}{2}\sum_{i=1}^{M}(\frac{1}{\lambda_i'} - 1)z_i'^2, \tag{71}$$

where z_i' is the i^{th} element of the transformed observation $\mathbf{Z}' = \mathbf{Q}'\mathbf{F}$. Now, using Theorem 3.4.1, $p(l|H_1)$ is Gaussian as the element z_i' is normal given hypothesis H_1, and the mean ($E[l|H_1]$) is given by (32). It can be shown that the variance ($Var[l|H_1]$) is given by [see appendix D for details]:

$$Var[l|H_1] = \frac{1}{2}\sum_{i=1}^{M}(\zeta_i^{\mathbf{F}})^2. \tag{72}$$

Using the above results and following the steps in obtaining (54), a closed form expression for P_m can be achieved:

$$P_m = erf(\frac{\gamma - E[l|H_1]}{\sqrt{Var[l|H_1]}}). \tag{73}$$

3.4.5 Comparison of the Bayesian, NP and CNP Detectors

The theoretical plots of P_m and P_f with prior SNR for the proposed Bayesian, NP and CNP detectors are shown in Fig. 9 (a)-(b), (c)-(d) and (e)-(h), respectively, where the average prior SNR $\bar{\zeta}$ is set to -5 and 5 dB. Ideally, the VAD should favor H_0 for low values of prior SNR, or it should have a high P_m and low P_f. Similarly, the detector should favor H_1 for high values of prior SNR, corresponding to high and low values of P_f and P_m, respectively. This ideal behavior reflects the partial prior knowledge about the prior SNR. From Fig. 9, it can be seen that all detectors achieve

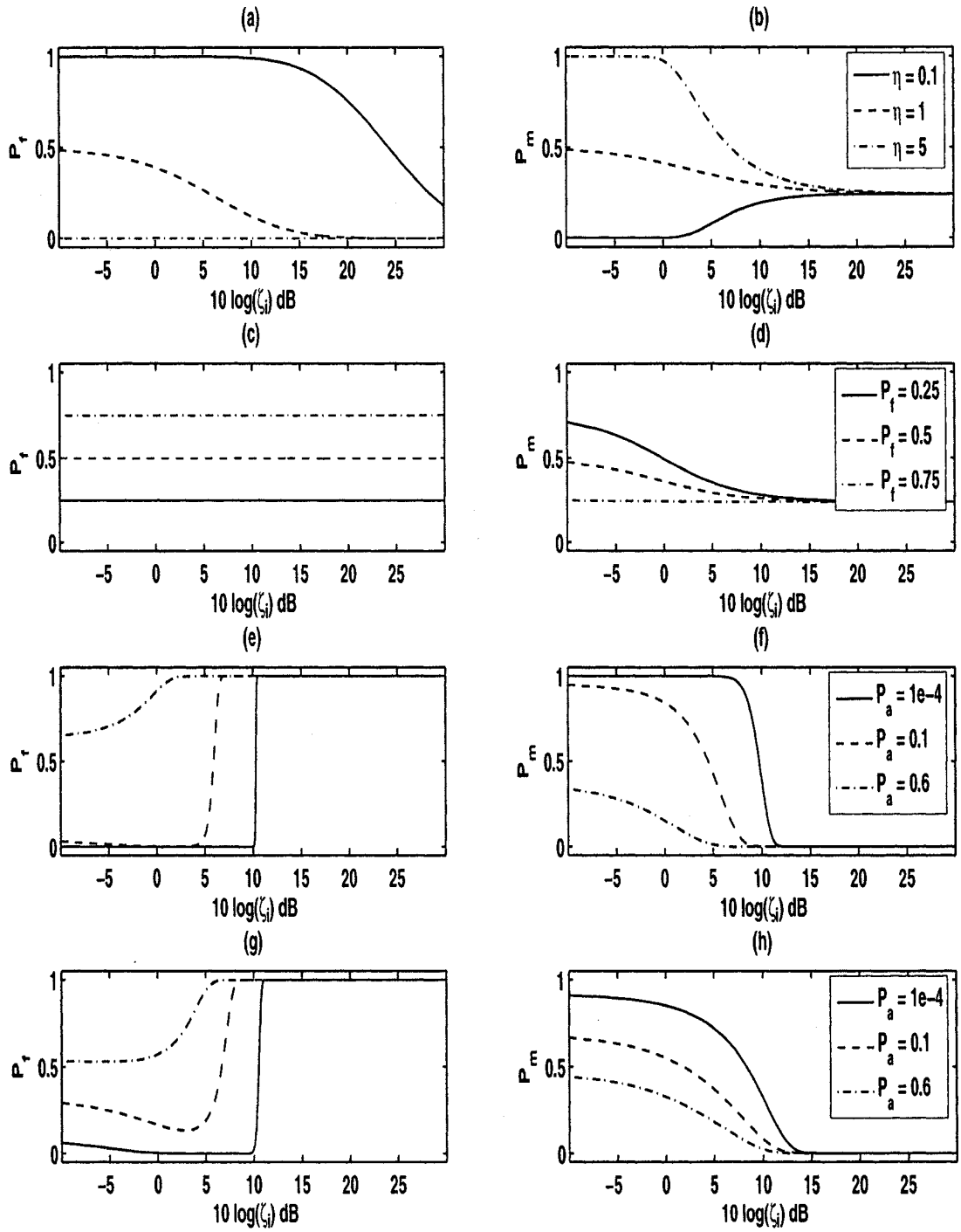

Figure 9: Variations of P_f and P_m with ζ_i: (a)&(b) Bayesian, (c)&(d) NP, (e)&(f) CNP with $\overline{\zeta} = -5dB$, and (g)&(h) CNP with $\overline{\zeta} = 5dB$ detectors.

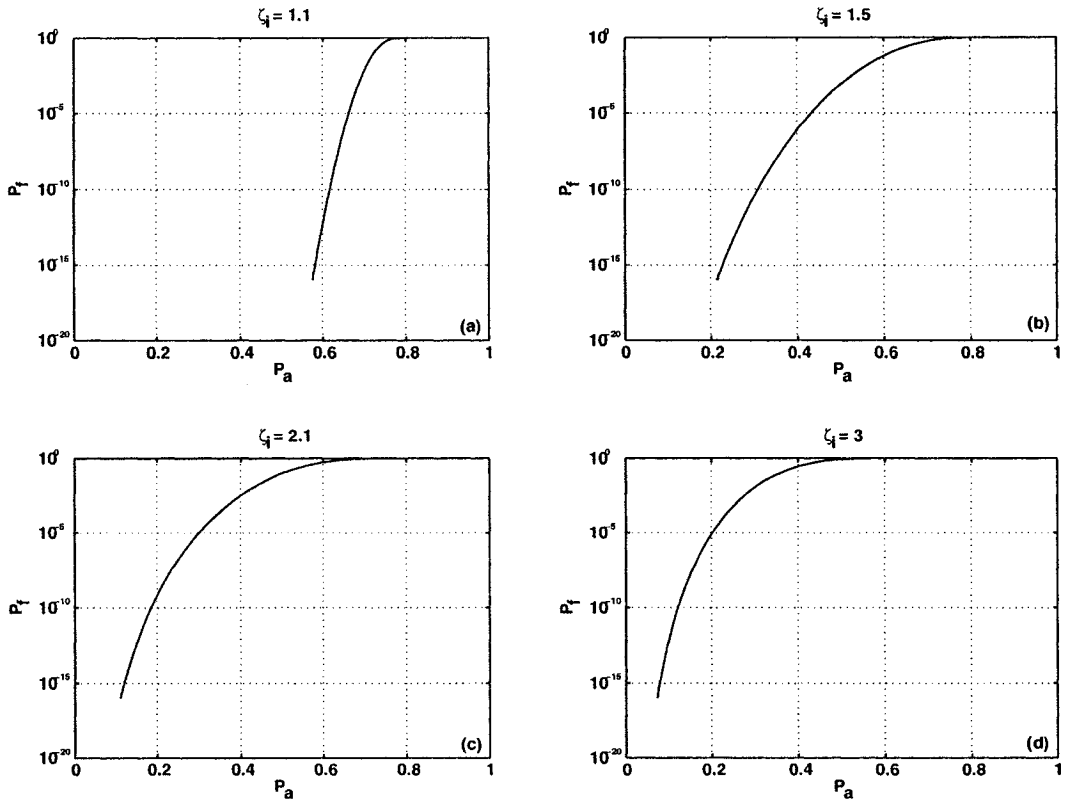

Figure 10: Illustrating the relationship between P_a and P_f for the CNP detector using different values of prior SNR: (a) $\zeta_i = 1.1$, (b) $\zeta_i = 1.5$, (c) $\zeta_i = 2.1$ and (d) $\zeta_i = 3$.

the ideal behavior for P_m but only the CNP detector achieves the ideal behavior for P_f.

In order to divulge the relationship between the P_a and P_f in the CNP detector, we plot the variation of P_f with P_a for different values of prior SNR in Fig. 10 (a), (b), (c) and (d). It is seen that the value of P_f is extremely small for small values of P_a for all prior SNRs [1]. As the value of P_a increases, P_f increases and finally exceeds the value of P_a. At low prior SNR, P_f crosses P_a at a relatively higher value when compared to the case of higher prior SNR, where P_f crosses P_a at a lower value of P_a. In general, for a particular P_a which is chosen as the operating value in the CNP

[1]The P_f vs. P_a curves end abruptly for all prior SNRs at low values of P_a due to level of numerical precision available in MATLAB.

detector, P_a acts as an upper bound to P_f at low prior SNR and lower bound at high prior SNR.

3.5 Contextual Detector

The proposed contextual detection scheme estimates the activity of a frame via the durational information present in the primary decisions of its neighboring frames. The real-time constraints on the VAD system forces the usage of past frames alone for this purpose. More specifically, the dynamic nature of speech suggests that a frame's activity is influenced by a short finite past. In order to process the contextual information, a set of m immediate past decisions (with respect to the current frame) is formed and termed as $\mathbf{B_D}$, i.e.,

$$\mathbf{B_D} = \{D_1, D_2, ..., D_m\} \tag{74}$$

where D_1 is the decision preceding the current frame's decision D_0, and D_m is the oldest decision in the set. It is easily seen that a group of successive speech decisions constitute a speech burst and successive pause decisions constitute a pause period in the noisy speech signal. It is our intention to determine the likelihood of speech activity in the i^{th} frame F_i based on neighboring speech bursts and pause periods in $\mathbf{B_D}$.

The process of contextual detection is demonstrated via an example in Fig. 11. In the example, the primary detector takes a decision D_i for the frame F_i where D_0 represents the current primary decision. In our illustration, a set of 13 past decisions with respect to D_0 constitute the set $\mathbf{B_D}$. Now, unlike contemporary hang-over schemes, we do not process the decisions directly, but aggregate them into a more intuitive and useful form, i.e., we define a group of successive similar decisions as a

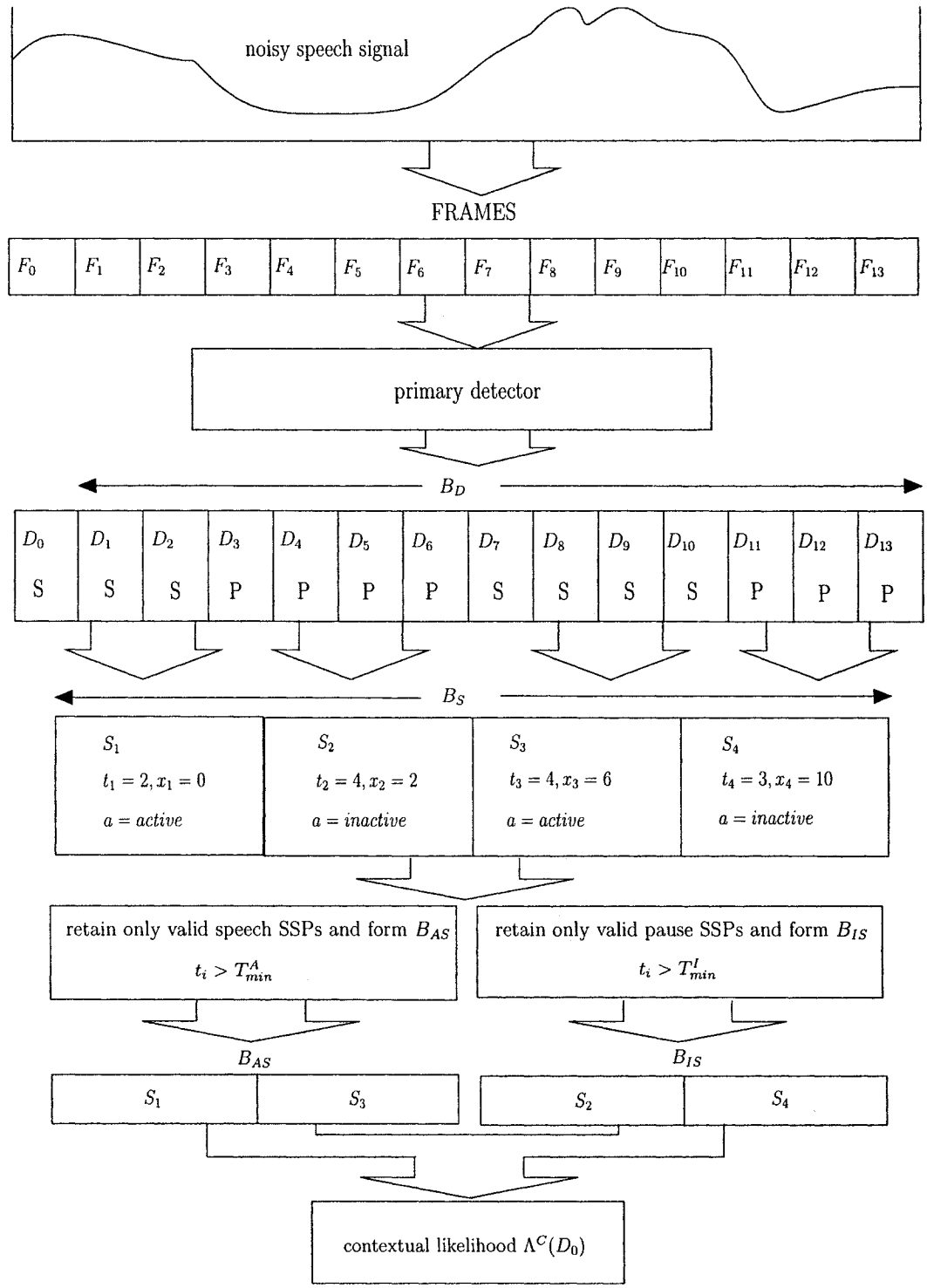

Figure 11: Illustrating the operation of the contextual detector.

'same state period' (SSP). Hence, all decisions in $\mathbf{B_D}$ can be grouped into speech and pause SSPs. For instance, in the example shown, S_2 is formed by grouping the four successive pause decisions and S_3 is a group of four successive speech decisions. Thus, from the decisions in $\mathbf{B_D}$, a set of SSPs can be formed which we denote by $\mathbf{B_S}$, i.e.,

$$\mathbf{B_S} = \{S_1, S_2, ..., S_n\}, (n \leq m) \tag{75}$$

where S_i is the i^{th} SSP. In Fig. 11, four SSPs are formed with two SSPs each of speech and pause.

Every SSP is characterized by duration, relative time position with respect to the current frame F_0 and its state of activity. The duration of the SSP is the 'number of frames' which are a part of that SSP itself, and the time position of a SSP is the minimum time distance between that SSP and the current frame F_0 measured in 'number of frames'. We now define three sets - duration $\mathbf{T}$, position $\mathbf{X}$ and activity $\mathbf{A}$ as:

$$\mathbf{T} = \{t : 1 \leq t \leq m\}, \tag{76}$$

$$\mathbf{X} = \{x : 0 \leq x \leq m-1\}, \tag{77}$$

$$\mathbf{A} = \{a : a = \textit{active} \text{ or } \textit{inactive}\}. \tag{78}$$

Hence, a convenient representation for the SSP is obtained in form of an ordered triplet:

$$S_i \equiv (t_i, x_i, a_i), \tag{79}$$

where $t_i \in \mathbf{T}$, $x_i \in \mathbf{X}$ and $a_i \in \mathbf{A}$. In Fig. 11, the duration, time position and activity of each SSP is shown. As it is expected that the decisions of the primary detector contain errors, we use the knowledge that speech/pause durations have lower limits

to prune $\mathbf{B_S}$, i.e., the contextual scheme validates all the SSPs in $\mathbf{B_S}$ by checking if the duration of the speech and pause SSP exceeds the minimum limit T^A_{min} and T^I_{min}, respectively. If a speech SSP duration is lower than T^A_{min} or a pause SSP duration is lower than T^I_{min}, then that SSP is invalidated and removed from the computation of contextual LR.

Further, by defining $\mathbf{B_{AS}}$ and $\mathbf{B_{IS}}$ as sets which contain the speech and pause SSPs alone, respectively, i.e.,

$$\mathbf{B_{AS}} = \{S_i : S_i \in B_S \text{ and } a_i = \text{active}\}, \tag{80}$$

$$\mathbf{B_{IS}} = \{S_i : S_i \in B_S \text{ and } a_i = \text{inactive}\}, \tag{81}$$

we form a partition over $\mathbf{B_S}$. Now, the valid speech and pause SSPs in $\mathbf{B_S}$ are divided between the $\mathbf{B_{AS}}$ and $\mathbf{B_{IS}}$, respectively. For instance, in the illustration, $S_1, S_3 \in \mathbf{B_{AS}}$ and $S_2, S_4 \in \mathbf{B_{IS}}$ assuming that all the SSPs are valid. Now, $\mathbf{B_{AS}}$ and $\mathbf{B_{IS}}$ are used to compute the contextual likelihood of activity for the frame F_0, where the contextual observation space is defined as the set of all possible SSPs, which can be expressed as a cartesian product of sets $\mathbf{T}$, $\mathbf{X}$ and $\mathbf{A}$. If the contextual observation space is a probability space $\mathbf{O^D}$, then each SSP S_i is an event in $\mathbf{O^D}$. Hence, the contextual LR of the decision D_0 based on observing $\mathbf{B_S}$, given the hypothesis in (2) can be written as:

$$\Lambda^C(D) = \frac{P(\mathbf{B_S}|H_1)}{P(\mathbf{B_S}|H_0)}. \tag{82}$$

where $P(\mathbf{B_S}|H_0)$ and $P(\mathbf{B_S}|H_1)$ are the conditional probability mass functions (PMFs).

The above equation is rewritten in terms of $\mathbf{B_{AS}}$ and $\mathbf{B_{IS}}$ as:

$$\Lambda^C(D) = \frac{P(\mathbf{B_{AS}}|H_1) + P(\mathbf{B_{IS}}|H_1)}{P(\mathbf{B_{AS}}|H_0) + P(\mathbf{B_{IS}}|H_0)}, \tag{83}$$

which can be further simplified using Bayes rule as:

$$\Lambda^C(D) = \frac{P(H_1|\mathbf{B_{AS}})P(\mathbf{B_{AS}})\frac{1}{P(H_1)} + P(H_1|\mathbf{B_{IS}})P(\mathbf{B_{IS}})\frac{1}{P(H_1)}}{P(H_0|\mathbf{B_{AS}})P(\mathbf{B_{AS}})\frac{1}{P(H_0)} + P(H_0|\mathbf{B_{IS}})P(\mathbf{B_{IS}})\frac{1}{P(H_0)}}. \tag{84}$$

If we assume that the speech and noise SSPs alone can influence the activity and inactivity of the frame F_0, respectively then the conditional probabilities $P(H_1|\mathbf{B_{IS}})$ and $P(H_0|\mathbf{B_{AS}})$ become zero, and (84) can be simplified as:

$$\Lambda^C(D) = \frac{P(H_1|\mathbf{B_{AS}})P(\mathbf{B_{AS}})}{P(H_0|\mathbf{B_{IS}})P(\mathbf{B_{IS}})}\frac{P(H_0)}{P(H_1)}. \tag{85}$$

The term $\frac{P(H_0)}{P(H_1)}$ is a constant and can be omitted from the likelihood. From the definition of the sets $\mathbf{B_{AS}}$ and $\mathbf{B_{IS}}$ in (80) and (81), we rewrite the likelihood in (85) as:

$$\Lambda^C(D) = \frac{P(H_1|\bigcup_i S_i^A)P(\bigcup_i S_i^A)}{P(H_0|\bigcup_i S_i^I)P(\bigcup_i S_i^I)} \tag{86}$$

Using the definition of conditional probability, we simplify (86) as:

$$\Lambda^C(D) = \frac{P(H_1 \bigcap (\bigcup_i S_i^A))}{P(H_0 \bigcap (\bigcup_i S_i^I))}, \tag{87}$$

$$= \frac{P(\bigcup_i (H_1 \bigcap S_i^A))}{P(\bigcup_i (H_0 \bigcap S_i^I))}. \tag{88}$$

Since $\mathbf{B_{AS}}$ and $\mathbf{B_{IS}}$ are sets of ordered triplets S_i which are mutually exclusive events

in $\mathbf{O^D}$, we can further simplify (88) as:

$$\Lambda^C(D) = \frac{\sum_i P(H_1 \bigcap S_i^A)}{\sum_j P(H_0 \bigcap S_j^I)}, \tag{89}$$

$$= \frac{\sum_i P(H_1|S_i^A) \times P(S_i^A)}{\sum_j P(H_0|S_j^I) \times P(S_j^I)}, \tag{90}$$

$$= \frac{\sum_i f_p(x_i) \times f_{ad}(t_i)}{\sum_j f_p(x_j) \times f_{id}(t_j)}, \tag{91}$$

where $f_{ad}(t)$ and $f_{id}(t)$ are the PMF of speech burst and pause duration, respectively. Also, $f_p(x)$ represents the PMF of the positional influence of a SSP (due to its time distance from the current frame) on the current frame. Hence, the likelihood in (91) favors the hypothesis which has more probabilistic and time relevant observations of SSPs. In a recent study on speech durational modeling, Chien and Huang [39] have reported that the Poisson distribution gives a good fit to the speech duration histogram. Hence, we assume that $f_{ad}(t)$ is Poisson distributed, i.e.,

$$f_{ad}(t_i) = \frac{\beta_A^{t_i}}{t_i!} \exp -\beta_A \times \frac{1}{N_F^A}, \tag{92}$$

where β_A is the parameter of the distribution, and N_F^A is a normalizing factor introduced as $t_i \in T$. Similarly, the pause duration is also modeled as a Poission distribution with the parameter β_I and normalizing factor N_F^I, i.e.,

$$f_{id}(t_j) = \frac{\beta_I^{t_j}}{t_j!} \exp -\beta_I \times \frac{1}{N_F^I}, \tag{93}$$

Intuitively, the speech or pause SSPs closest to the current frame in time must have the largest impact on the decision. Hence, it is assumed that the positional

influence of the SSPs on the state of the current frame is geometrically distributed,

$$f_p(x) = (1-\beta_P)^{x-1}\beta_P \times \frac{1}{R_f}, \tag{94}$$

where β_P is the parameter of the distribution and R_f is a normalizing factor introduced as $i \in X$. Using (92), (93) and (94) into (91),we obtain the final form of the contextual likelihood, i.e.,

$$\Lambda^C(D;\beta_P,\beta_I,\beta_A) = \frac{\sum_i (1-\beta_P)^{x_i-1}\beta_P \times \frac{\beta_A^{t_i}}{t_i!}\exp -\beta_A \times \frac{1}{N_F^A}}{\sum_j (1-\beta_P)^{x_j-1}\beta_P \times \frac{\beta_I^{t_j}}{t_j!}\exp -\beta_I \times \frac{1}{N_F^I}} \tag{95}$$

3.6 Comprehensive Voice Activity Detector

The overall log likelihood of activity is given by the sum of primary and contextual LRs, assuming that the observations D and F are mutually independent, i.e.,

$$\ln(\Lambda(D,F)) = \ln(\Lambda^C(D)) + \ln(\Lambda^P(F)) \tag{96}$$

where $\Lambda(D,F)$ is the overall LR. While the contextual LR, $\Lambda^C(D)$ is given by (95), the primary LRs for the Bayesian, NP and CNP detectors can be obtained by rewritting (7), (57) and (68) as:

$$l + \frac{1}{2}\ln\frac{\mathbf{K_n}}{\mathbf{K_f}} \underset{H_0}{\overset{H_1}{\gtrless}} \ln\eta, \tag{97}$$

$$\frac{1}{Var[l|H_0]} \times (l - E[l|H_0]) \underset{H_0}{\overset{H_1}{\gtrless}} erf^{-1}(1-\alpha) \tag{98}$$

and

$$\frac{1}{d \times S(\bar{\zeta})}(l + \frac{1}{2}\ln\frac{\mathbf{K_n}}{\mathbf{K_f}}) + \frac{d}{2} \underset{H_0}{\overset{H_1}{\gtrless}}, \quad erf^{-1}(1-\alpha), \tag{99}$$

respectively, where the LHS of (97), (98) and (99) represent the primary LRs of Bayesian, NP, and CNP detectors, i.e.,

$$\Lambda_B^P(F) = l + \frac{1}{2}\ln\frac{\mathbf{K_n}}{\mathbf{K_f}}, \tag{100}$$

$$\Lambda_{NP}^P(F) = \frac{1}{Var[l|H_0]} \times (l - E[l|H_0]), \tag{101}$$

$$\Lambda_{CNP}^P(D) = \frac{1}{d \times S(\bar{\zeta})}(l + \frac{1}{2}\ln\frac{\mathbf{K_n}}{\mathbf{K_f}}) + \frac{d}{2}. \tag{102}$$

Now, the LRT for CVAD-Bayesian, CVAD-NP and CVAD-CNP can be easily obtained by replacing the primary LR in (97), (98) and (99) by the overall LR, i.e.,

$$\Lambda_B(D,F) = \Lambda^C(D) + \Lambda_B^P(F) \underset{<H_0}{\overset{\geq H_1}{}} \ln\eta. \tag{103}$$

$$\Lambda_{NP}(D,F) = \Lambda^C(D) + \Lambda_{NP}^P(F) \underset{<H_0}{\overset{\geq H_1}{}} erf^{-1}(1-\alpha), \tag{104}$$

$$\Lambda_{CNP}(D,F) = \Lambda^C(D) + \Lambda_{CNP}^P(F) \underset{<H_0}{\overset{\geq H_1}{}} erf^{-1}(1-\alpha). \tag{105}$$

The implementation of the proposed Bayesian, NP and CNP primary detectors is outlined in Algorithm 1. In steps (16) and (18), α and β are the parameters of the updating rule which are set such that the time constants of the update rules for speech and noise are 10ms and 0.5s, respectively [19]. Further, the implementation of the CVAD-CNP, CVAD-NP and CVAD-Bayesian detectors is given in Algorithms 2 and 3.

Algorithm 1 Proposed primary detectors

1: Initialize: $\mathbf{K_f}$, $\mathbf{K_n}$ to non-zero values.
2: Compute a M-point MFSC for a speech frame.
3: Compute $\gamma_i^F = \frac{\sigma_{fi}^2}{\sigma_{ni}^2}$ and $\zeta_i^F = \gamma_i^F - 1 \; \forall i = 1, 2..., M$.
4: **if** implementing the CNP **then**
5: Compute $\overline{\zeta}$ and $S(\overline{\zeta})$ using (60) and (59).
6: Compute d using (34) and γ_{CNP} using (67).
7: **else if** implementing the NP **then**
8: Compute $E[l|H_0]$ and $Var[l|H_0]$ using (31) and (33).
9: Compute γ_{NP} using (56).
10: **else if** implementing the Bayesian **then**
11: Compute γ using (37).
12: **end if**
13: Compute the SS using (18), (19) and (30), giving:

$$l = \frac{1}{2} \sum_{i=1}^{M} \frac{\zeta_i^F}{\gamma_i^F} \frac{1}{\sigma_n i^2} f_i^2.$$

14: Make a decision D_0 using (68) for CNP, (57) for NP or (7) for Bayesian detector.
15: **if** VAD decision is speech **then**
16: Update $\mathbf{K_f}$ using:

$$\mathbf{K_f}[j] \Leftarrow \alpha \mathbf{K_f}[j-1] + (1-\alpha)(\mathbf{F}[j] \times \mathbf{F}^T[j]),$$

17: **else**
18: Update $\mathbf{K_n}$ using:

$$\mathbf{K_n}[j] \Leftarrow \beta \mathbf{K_n}[j-1] + (1-\beta)(\mathbf{F}[j] \times \mathbf{F}^T[j]),$$

19: **end if**
20: Goto step 2 and repeat for next frame.

Algorithm 2 Proposed contextual and comprehensive VADs

Initialize $\mathbf{B_S}$.
Obtain the primary decision D_i for frame F_i

procedure UPDATE SSPS($\mathbf{B_S}$) ▷ **UPDATES SSPs IN $\mathbf{B_S}$**
    $add\_flag = 0$
    for $i = 1, 2, ..., n$ **do**
        if $i \neq 1$ **AND** $i \neq n$ **then**
            $x_i \Leftarrow x_i + 1$
        else
            if $i = 1$ **then**
                if D_i is speech & a_i = activity | D_i is pause & a_i = inactivity **then**
                    $t_i \Leftarrow t_i + 1$
                else
                    $x_i \Leftarrow x_i + 1$
                    form new SSP S_0 with D_0 as the only member.
                    set $add\_flag = 1$
                end if
            end if
            if $i = n$ **then**
                if $t_n = 1$ **then**
                    remove S_n from $\mathbf{B_S}$
                else
                    $t_i \Leftarrow t_i - 1$
                    $x_i \Leftarrow x_i + 1$
                end if
            end if
        end if
    end for
    if $add\_flag = 1$ **then**
        add S_0 as the first member of $\mathbf{B_S}$ ▷ **ADD NEW SSP TO $\mathbf{B_S}$**
        reset add_flag = 0
    end if
end procedure

Algorithm 3 Proposed contextual and comprehensive VADs: continued

procedure COMPUTE CONTEXTUAL LR($\mathbf{B_S}$)
  initialize $N_r, D_r = 0$
  for $i = 1, 2, ..., n$ **do**
    if a_i = activity **then**
      if $t_i > T^A_{min}$ **then**
        Compute $f_{ad}(t_i)$ using (92), and $f_p(x_i)$ using (94).
        $N_r \Leftarrow N_r + f_{ad}(t_i) f_p(x_i)$.
      end if
    else
      if $t_i > T^I_{min}$ **then**
        Compute $f_{id}(t_i)$ using (93), and $f_p(x_i)$ using (94).
        $D_r \Leftarrow D_r + f_{id}(t_i) f_p(x_i)$.
      end if
    end if
  end for
  Compute $\Lambda^C(D)$ using (95), i.e., $\Lambda^C(D) = \frac{N_r}{D_r}$.
end procedure

procedure COMPUTE OVERALL LR
  if implementing CVAD-CNP **then**
    Compute $\Lambda^P_{CNP}(F)$ using (102).
    Use (105) to make the final VAD decision FD_0
  else if implementing CVAD-NP **then**
    Compute $\Lambda^P_{NP}(F)$ using (101).
    Use (104) to make the final VAD decision FD_0
  else if implementing CVAD-Bayesian **then**
    Compute $\Lambda^P_B(F)$ using (100).
    Use (103) to make the final VAD decision FD_0
  end if
end procedure
Goto step 2 and repeat for next frame.

Chapter 4

Evaluation of the Proposed Voice Activity Detectors

IN this chapter, we present the simulation results of the proposed VAD schemes, and compare them with the AMR VADs. We start with the description of the simulation setup, followed by a discussion on the evaluation parameters used to measure the performance of the VADs, namely, the overall detection rate, speech and pause detection rate, receiver operating characteristics (ROC), activity burst corruption (ABC) parameter, and computational complexity. Finally, we compare the different VAD schemes using each of the above-mentioned evaluation parameters.

4.1 Simulation Setup

The proposed VAD schemes are tested using the SWITCHBOARD speech database where the transcribed speech files provided by the database are used to evaluate the performance of the VADs. The SWITCHBOARD database contains extracts from actual telephonic conversations from which our simulation uses 21 speech samples of

1 minute each with the percentage speech content of the files varying between 40% and 60%. In order to create noisy speech samples of -10, 0, 15 and 30 dB SNR, four types of noises: car, babble, F-16 cockpit and tank noise from the NOISEX database are used following the additive noise model.

For the simulation results presented in this chapter, the various parameters of the proposed VADs are set as follows. As in the AMR VADs, a frame length of 20 ms is chosen for all the proposed VAD schemes. The length of the decision history set $\mathbf{B_D}$ is chosen to be 500ms, i.e., it is equivalent to 25 noisy speech frames F_i. Further, the parameters α and β, used in the covariance matrix update rules are set to 0.6 and 0.02, respectively, and the parameters of the Poisson distributions, β_A and β_I are set to 10 and 17. Finally, the parameter of the geometric distribution, β_P is set to 0.5.

4.2 Evaluation Criteria

We compare the proposed VAD schemes with the AMR VADs in terms of their ability to correctly detect the speech and pause hypothesis. Towards this, we define the following objective parameters:

- *Overall Detection Rate (D)*: the ratio of correctly detected frames to the total number of frames in the given speech sample speech, expressed as a percentage. This parameter is complementary to the overall error probability (P_e).

- *Speech Detection Rate (S)*: the ratio of correctly detected speech frames to the total number of speech frames in the given speech sample speech, expressed as a percentage. This parameter is complementary to miss-detections probability (P_m).

- *Pause Detection Rate (P)*: the ratio of correctly detected pause frames to the

total number of pause frames in the given speech sample, expressed as a percentage. This parameter is complementary to false alarm probability (P_f).

In particular, we directly compare the overall detection rate (D) of the proposed VAD schemes with the AMR VADs, and demonstrate the speech (S) and pause (P) detection rates via the ROC curves. While the ROC completely characterizes the performance of a LRT, it is not sufficient to assess the subjective performance of the VADs. Subjective evaluation of a VAD is necessary to comprehend the perceptual effects of clippings on the quality and intelligibility of a speech sample. Herein, subjective measures such as the Comparison Mean Opinion Score ($CMOS$) are used to rate the perceptual quality of the processed speech samples. Unfortunately, CMOS scores are obtained via informal listening tests which can be a tedious and costly task. Alternatively, Beritelli et. al. have proposed a new subjective parameter termed as the Activity Burst Corruption (ABC), which can be calculated objectively. The ABC is based on a psychoacoustic auditory model, and the authors have also shown a strong correlation between the ABC and $CMOS$ [41]. In this chapter, we present the ABC scores of the proposed VADs and compare them with the AMR VADs. Further, we also study the effect of incorporating the proposed contextual detector scheme by comparing the performance of each primary detector with its CVAD counterpart, i.e., CNP with CVAD-CNP, NP with CVAD-NP and Bayesian with CVAD-Bayesian. The benefit of employing the contextual detector is judged by measuring the percentage change in overall detection rate, i.e.,

$$PI = 100 \times \frac{D_c - D_p}{D_p}, \tag{106}$$

where D_p and D_c are the overall detection rates of the primary detector and its CVAD counterpart, and PI represents the percentage change in detection. Finally,

we also report the computational complexity of the proposed VAD schemes, which is necessary to gauge the suitability of the VADs in real-time systems.

4.3 Overall Detection Rate

The overall detection rate (D) for the proposed primary VADs is compared with AMR VADs in Fig. 12 (a), (b), (c) and (d) for babble, car, F-16 cockpit and tank noises, respectively. As expected, the performance of the VADs is good at high SNR with all VADs consistently hitting the 90% detection mark. However, the real difference emerges at low to intermediate SNRs, where the superiority of the CNP detector over its Bayesian and NP counterparts is clearly seen. It is observed that the CNP detector achieves a performance improvement of close to 5-10% over the Bayesian and NP detectors in F-16 cockpit and tank noises. Further, CNP and NP detectors perform extremely well in babble noise where they outperform the AMR VADs by big margins. The only exception is the car noise where the detection rate of every VAD is very good across all SNRs, and the differences in performance are negligible.

At this point, it is useful to interpret the obtained results in terms of the spectral properties of the different noises used in our simulation. While the car and tank noises have a predominantly low-frequency spectrum, the F-16 cockpit and babble noises have a diffused spectrum where the spectral energy is well distributed. Moreover, the babble noise spectrum closely resembles the speech spectrum in terms of the energy concentrations. Therefore, unlike the car noise which affects a few primary cues of speech in low-frequency speech spectrum, the babble noise affects most primary cues across the entire speech spectrum. Since the proposed VADs and AMR VADs are frequency-band specific energy based methods, they rely on the clean frequency bands to make decisions and fail when the entire spectrum is corrupted. This observation

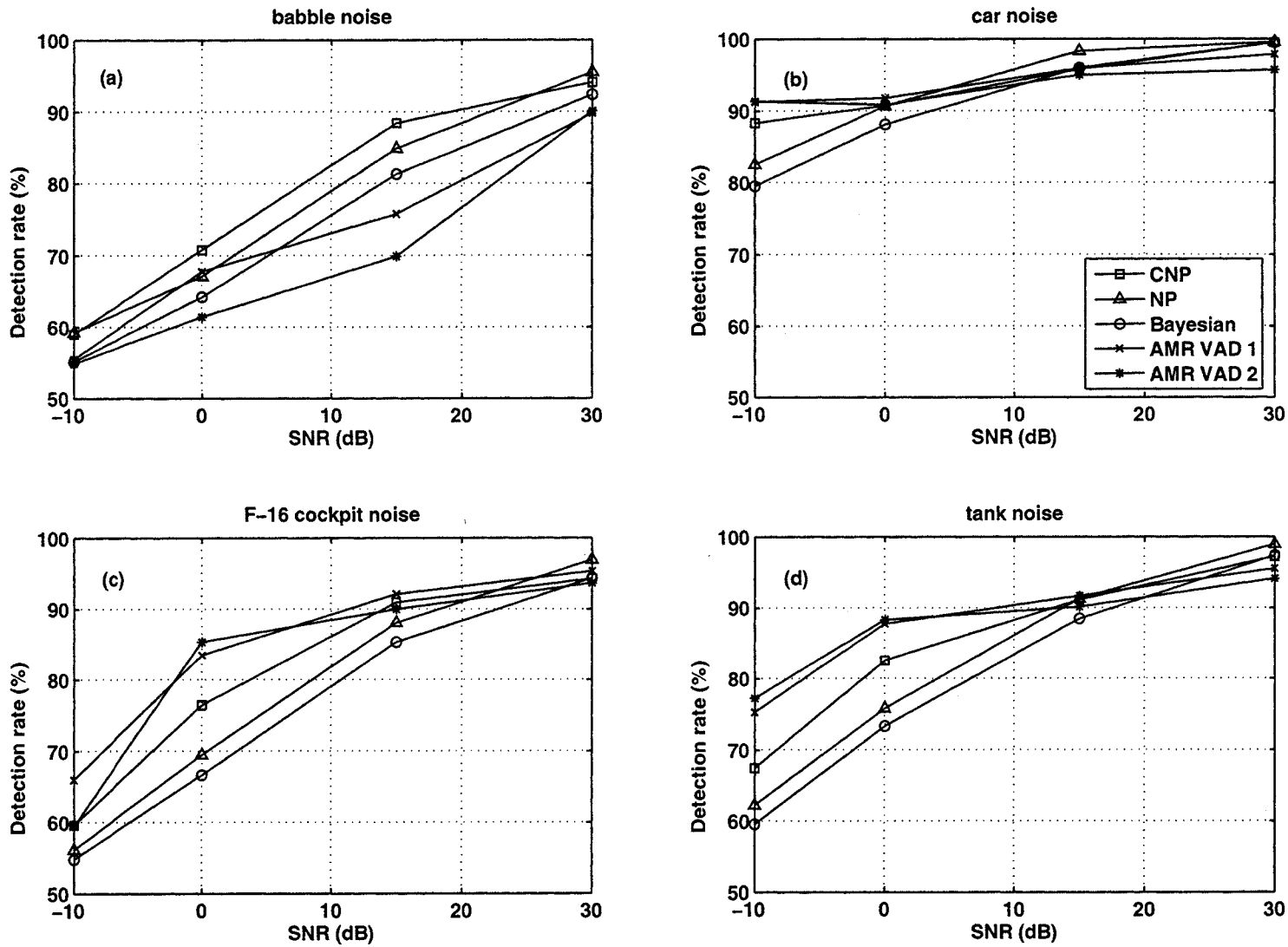

Figure 12: The overall detection rate (D) of the proposed primary detectors in (a) babble, (b) car, (c) F-16 cockpit and (d) tank noises.

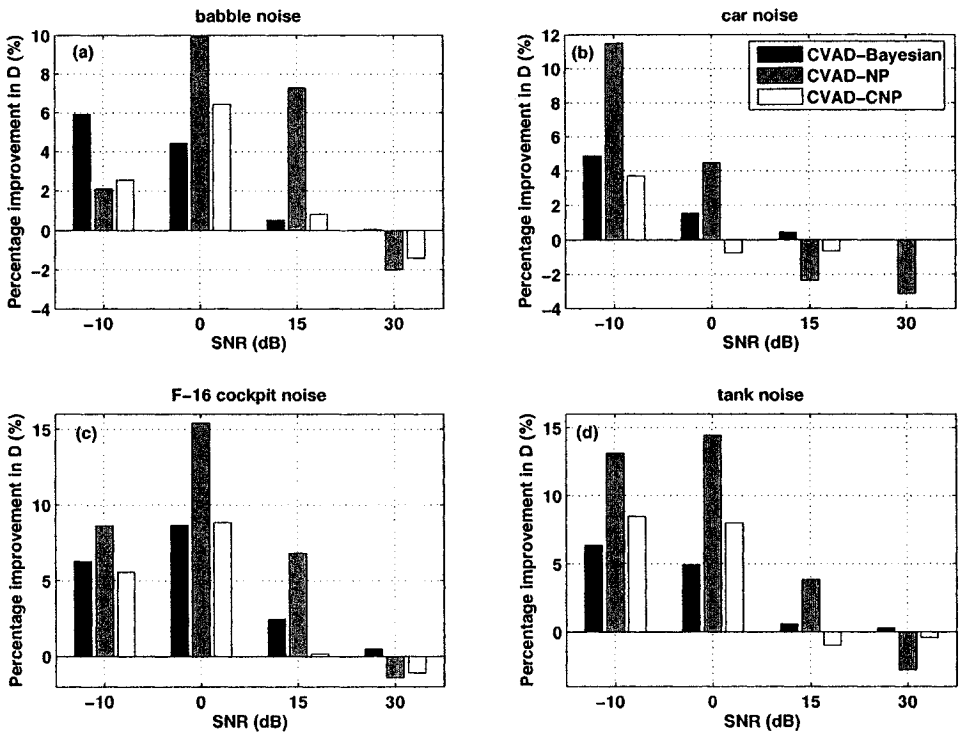

Figure 13: Performance Improvement (PI) parameter of the proposed CVAD schemes in (a) babble, (b) car, (c) F-16 cockpit and (d) tank noises.

explains the relatively sharper fall in detection with lowering of the SNR in babble and F-16 cockpit noises, than car and tank noises. Moreover, as the corrupting capability of the noises decrease with increasing SNR, the detector performance is seen to be good for all noises at higher SNRs.

The above explanation also extends to the utility of the contextual detector, i.e., incorporating contextual scheme in babble noise gives big gains as the secondary cues help in resolving the ambiguity posed by the primary cues. However, in the case of car noise where primary cues are not ambiguous, the contextual detector's value addition is limited. The improvement in performance as a result of employing the contextual detector scheme is shown in terms of the PI parameter in Fig. 13 (a),

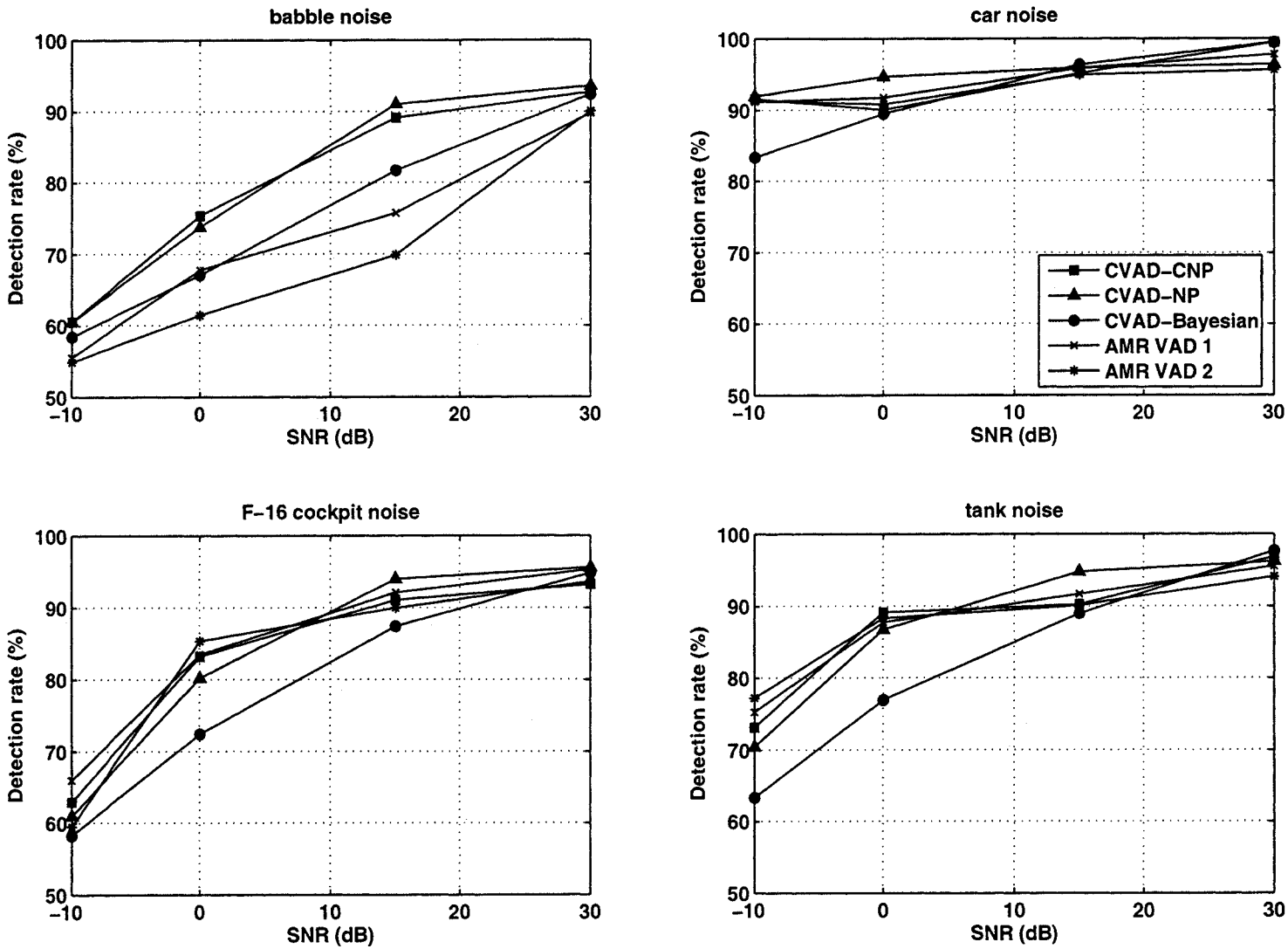

Figure 14: The overall detection rate (D) of the proposed comprehensive VADs in (a) babble, (b) car, (c) F-16 cockpit and (d) tank noises.

(b), (c) and (d) for babble, car, F-16 cockpit and tank noises, respectively. It is immediately evident that the biggest gains are obtained at low SNRs for all noises, and the gains diminish with increasing SNR. In fact, at high SNR, a small drop in detection rate is also observed in some cases. However, the drop is insignificant in all cases, as it coincides with a very high primary detection rate of 95% or above. On the other hand, the improvements at 0 dB SNR in F-16 cockpit, tank and babble noises are very impressive with gain margins of 10-15%. Finally, it is also seen that the CVAD-NP registers the biggest gains across all noises and SNRs.

Now, we compare the overall detection rate (D) of the CVADs to AMR VADs in Fig. 14 (a), (b), (c) and (d) for babble, car, F-16 cockpit and tank noises, respectively. It is easily seen that the CVAD-NP and CVAD-CNP closely match or better the performance of the AMR VADs consistently for all noises and SNRs. Again, the biggest gains over the AMR VADs are obtained in the case of the babble noise. Also, it is interesting to observe that while the CVAD-CNP and CVAD-NP schemes match the performance of the AMR VADs in the case of F-16 cockpit and tank noises, the performance of the CNP and NP primary detectors was inferior. This clearly indicates the benefits of incorporating the contextual detection scheme.

4.4 Receiver Operating Characteristics

The ROC plots the probability of detect ($P_d = 1 - P_m$) against the probability of false-alarm (P_f) by varying the tunable parameter of the detector [23]. In this manner, the ROC gives a complete picture of the detector in terms of its performance at different operating points. For the proposed VADs, the ROC curves are obtained by plotting the speech detection rate (S) against the complement of the pause detection rate (100 - P). It is useful to note that for a binary hypothesis, the straight line $P_d = P_f$ (or,

in this case, $S = 100 - P$) in the ROC is equivalent to random guessing. Hence, the higher the ROC curve from the line $S = 100 - P$, the better the performance of the detector is [23].

The ROC curves of the proposed detectors in babble, car, F-16 cockpit and tank noises are shown in Figs. 15, 16, 17 and 18 for -10, 0, 15 and 30 dB SNRs, respectively. For the babble and F-16 cockpit noises, it can be observed from Fig. 15 and 18 that the performance of all detectors is extremely good or poor in high or low SNR, respectively. This result is similar to the overall detection rate presented in the last section, as it shows that the task of voice activity detection is very easy at high SNR, and increasingly difficult with the lowering of the SNR. Comparing Figs. 15 and 18, it is seen that while the performance of the CVAD-NP is better than CVAD-CNP at 30 dB SNR for all noises, the CVAD-CNP does better than CVAD-NP at -10 dB SNR. Further, from Figs 15, 16, 17 and 18, it is also seen that the ROC curves for CVAD-CNP rise at a faster rate in comparison to CVAD-NP or CVAD-Bayesian, suggesting that the pause detection capability of CVAD-CNP is the best among the three CVADs. On the other hand, it can be seen that the CVAD-NP curve crosses the CVAD-CNP in most cases and, climbs a greater vertical distance, suggesting that the speech detection capability of CVAD-NP is better than CVAD-CNP or CVAD-Bayesian.

4.5 Activity Burst Corruption

The *ABC* parameter for the proposed VADs and AMR VADs is shown in Fig. 19 (a), (b), (c) and (d) for babble, car, F-16 cockpit and tank noises, respectively. It is useful to note that the *ABC* is closely related to speech detection rate but independent of overall detection rate. Hence, the *ABC* scores must not be judged in isolation, but

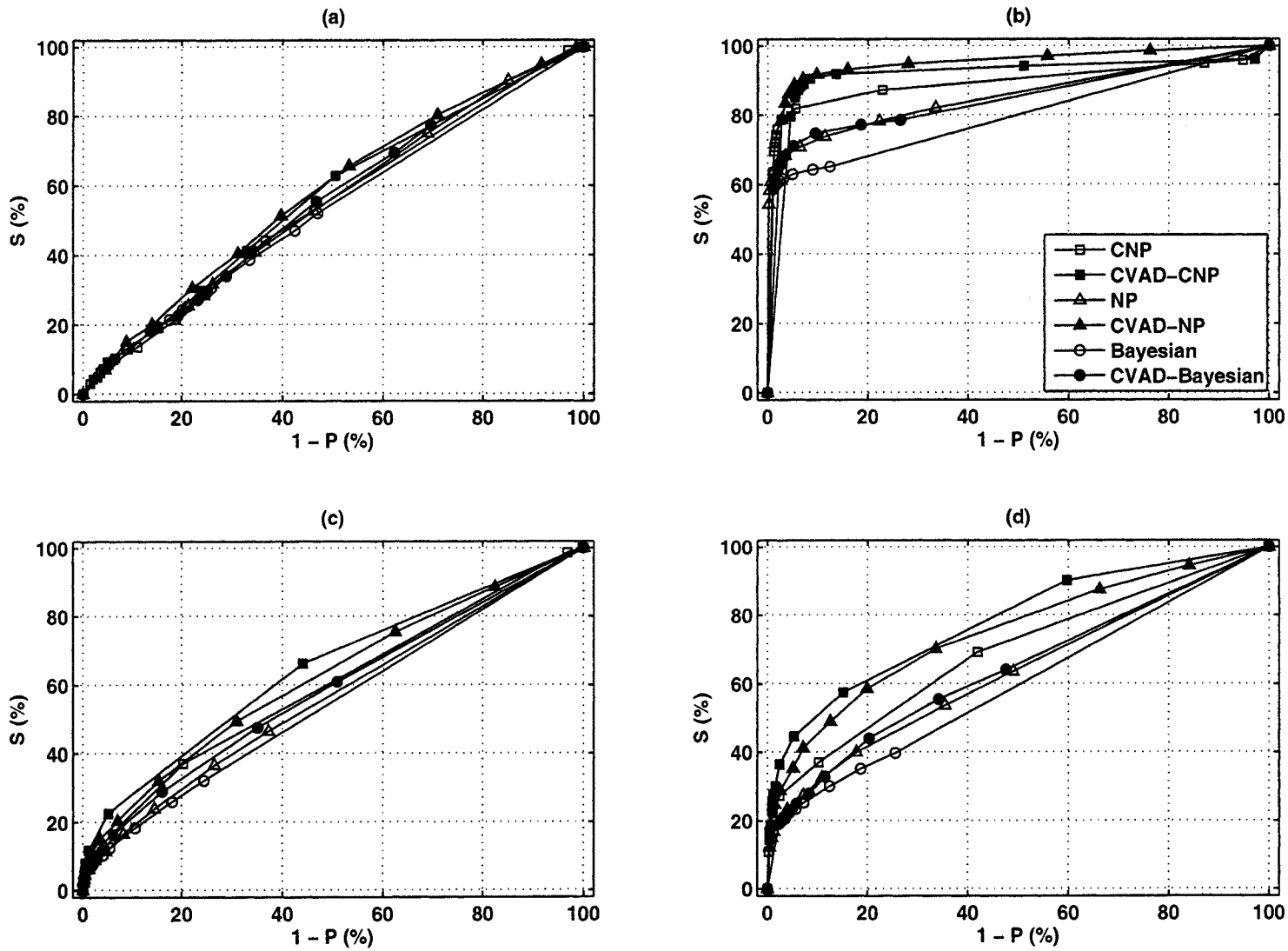

Figure 15: Receiver Operating Characteristics (ROC) of the proposed VADs at -10 dB SNR, in (a) babble, (b) car, (c) F-16 cockpit and (d) tank noises.

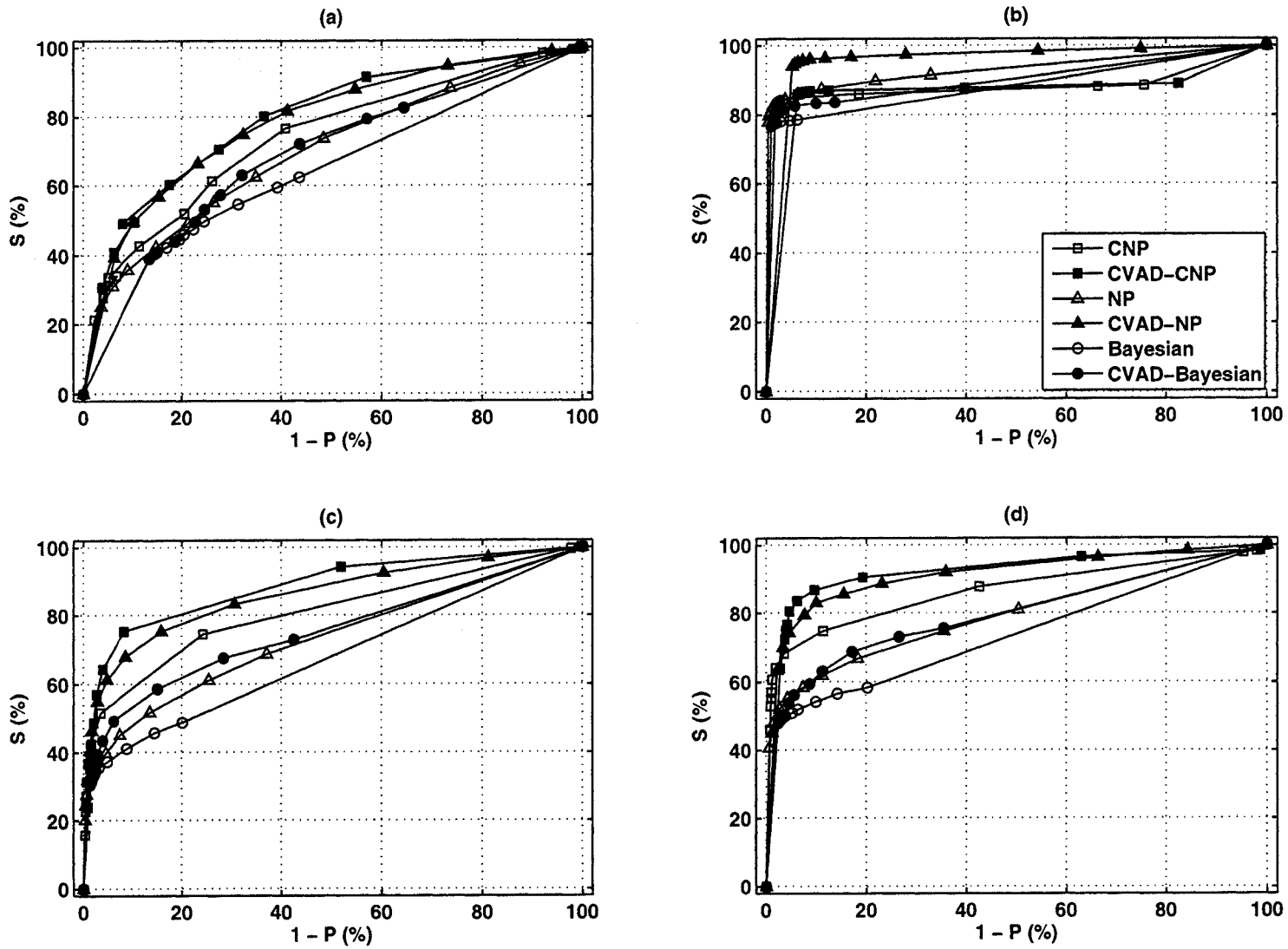

Figure 16: Receiver Operating Characteristics (ROC) of the proposed VADs at 0 dB SNR, in (a) babble, (b) car, (c) F-16 cockpit and (d) tank noises.

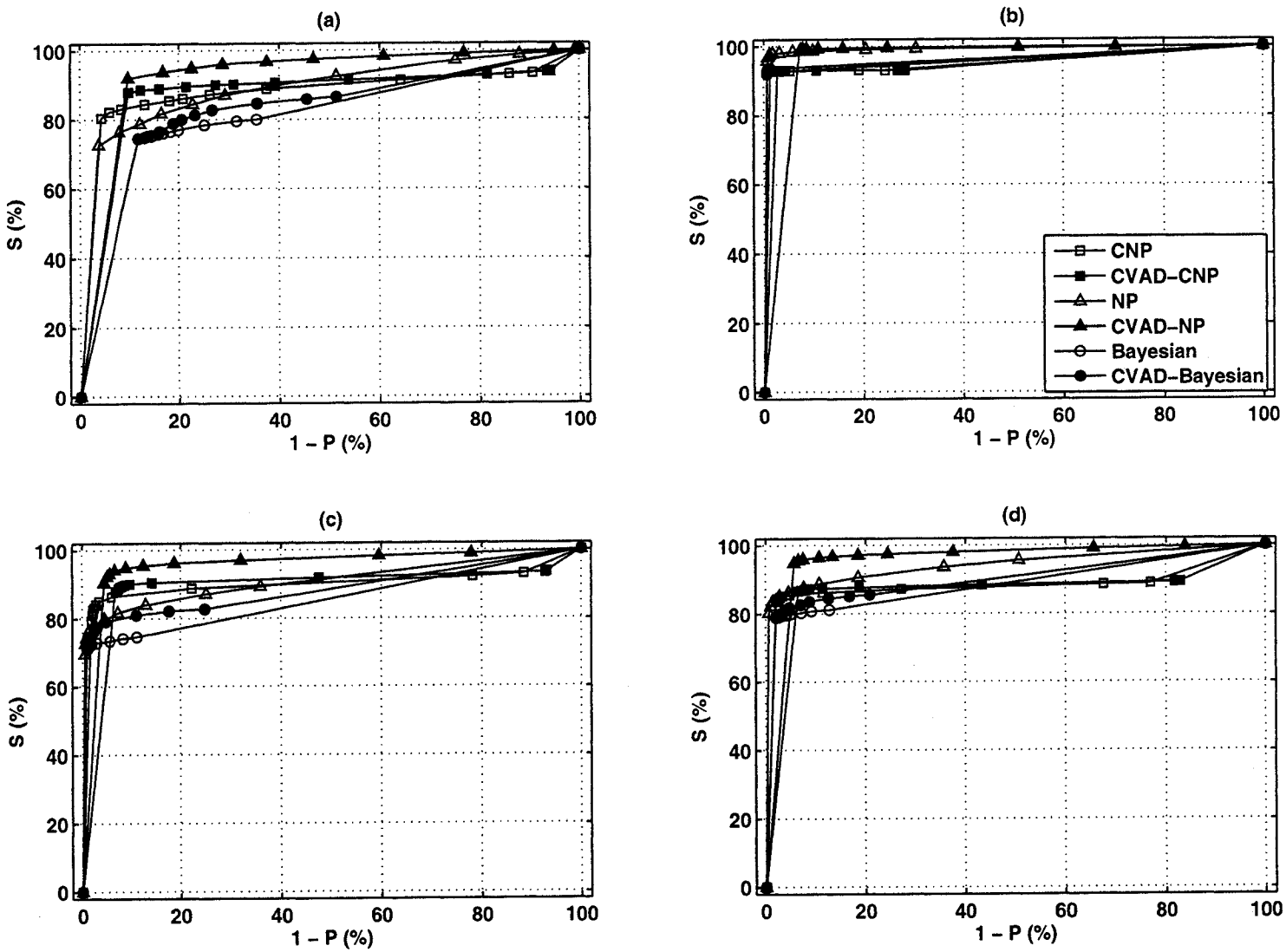

Figure 17: Receiver Operating Characteristics (ROC) of the proposed VADs at 15 dB SNR, in (a) babble, (b) car, (c) F-16 cockpit and (d) tank noises.

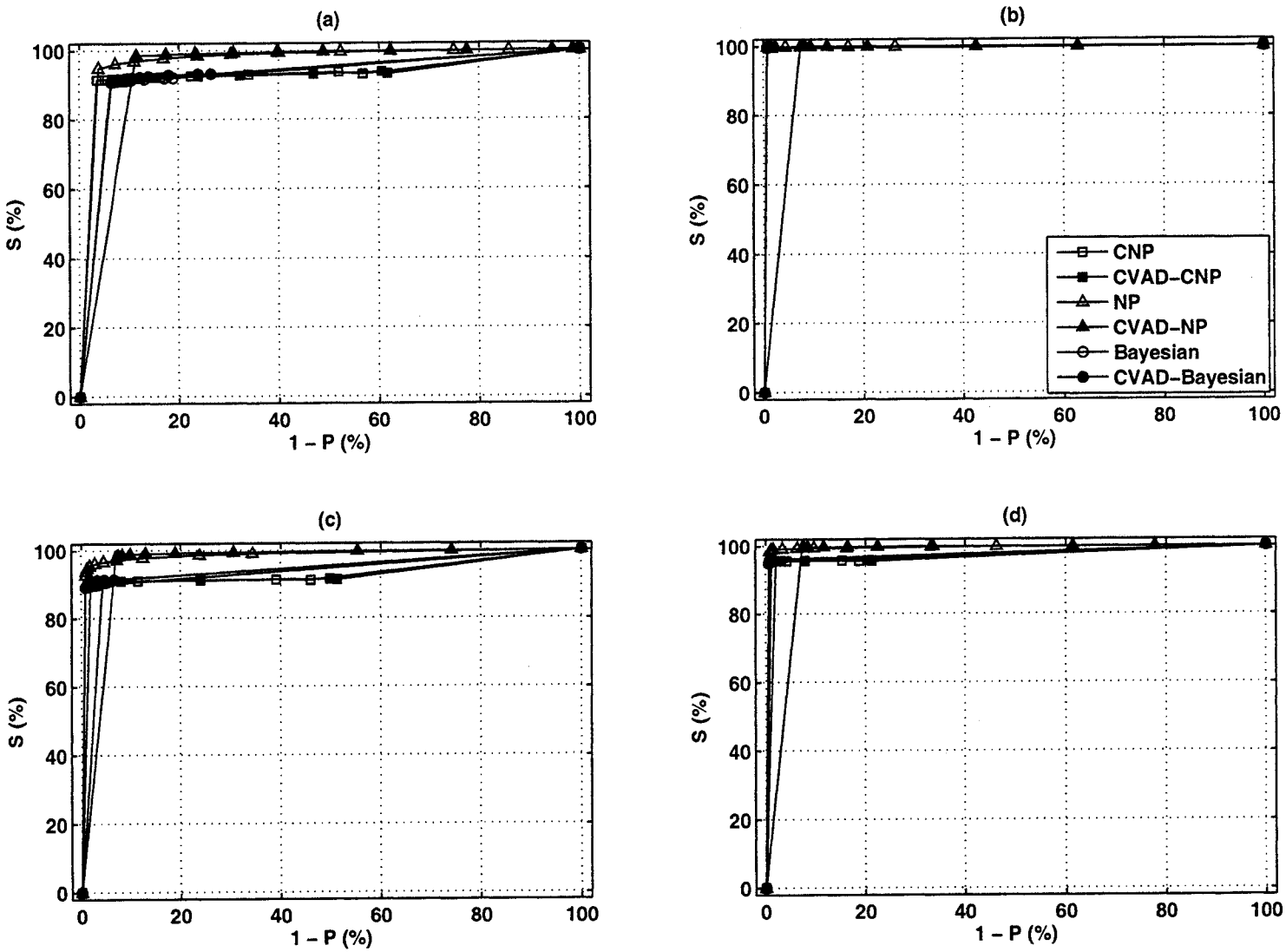

Figure 18: Receiver Operating Characteristics (ROC) of the proposed VADs at 30 dB SNR, in (a) babble, (b) car, (c) F-16 cockpit and (d) tank noises.

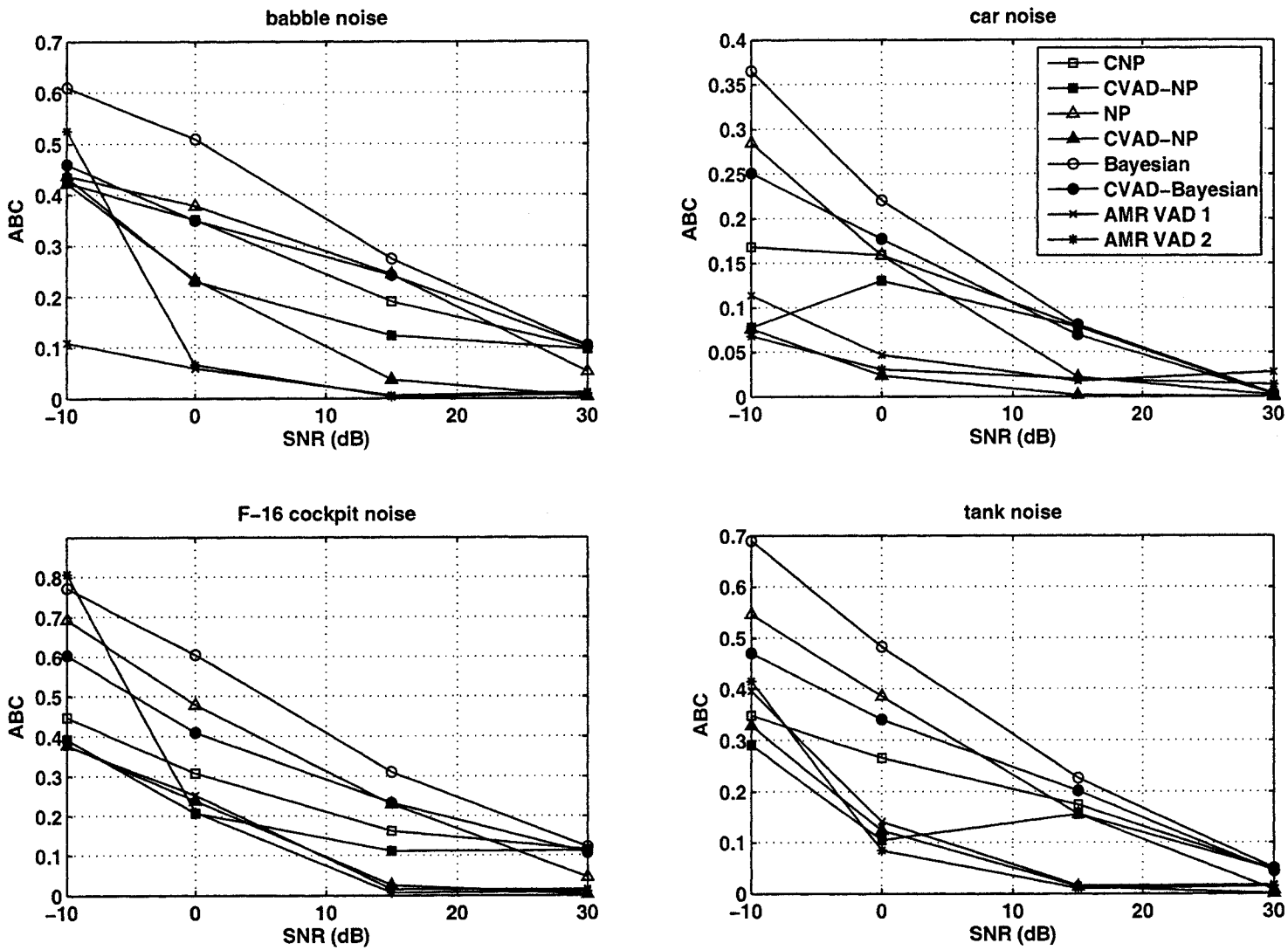

Figure 19: Illustrating the Activity Burst Corruption (ABC) parameter for the proposed VADs in (a) babble, (b) car, (c) F-16 cockpit and (d) tank noises.

in conjunction with overall detection rate.

As before, for all VADs, the fall in the ABC parameter as the SNR increases indicates the improvement in the speech perception quality. Also, the improvement in the ABC parameter while going from a primary to a comprehensive VAD scheme is also evident for all three primary VADs. The CVAD-NP achieves the best ABC values and is closely followed by the CVAD-CNP. Particularly, the CVAD-NP matches the ABC score of the AMR VADs. Further, the biggest improvement in the ABC score is achieved while moving from the NP to CVAD-NP detector. This result matches with the trends observed in the ROC curves presented in the last section, where it was concluded that the speech detection capability of the CVAD-NP is the best among the CVADs.

4.6 Computational Complexity

The computational complexity of the CVAD-CNP, CVAD-NP and CVAD-Bayesian is shown in terms of the number of mathematical operations and FLOPS (floating-point operations) in Table 5. The split up of the operations between the primary and contextual part of the detection scheme is also shown. The FLOPS for the various mathematical operations are obtained from Pinka's lightspeed Matlab toolbox [42], where the numbers reflect an implementation on the Intel's Pentium-4 architecture. The complexity calculations in the table are specific to the Algorithms 1 and 2, described in Chap. 3. Overall, the computational requirements of the proposed VAD schemes is very low, and roughly equivalent for all the three CVAD schemes.

Table 5: Computational complexity of the proposed CVAD-CNP, CVAD-NP and CVAD-Bayesian detectors.

Math Operation (Op)	Flops per Math Operation	CVAD-CNP		CVAD-NP		CVAD-Bayesian	
		Ops	Flops	Ops	Flops	Ops	Flops
		Contextual detector					
Multiplication	1	12	12	10	10	10	10
Division	8	2	16	2	16	1	8
Sub/Add	1	39	39	37	37	37	37
Logical	1	6	6	6	6	6	6
Comparison	1	85	85	85	85	85	85
Total			158		154		146
		Primary detector					
Multiplication	1	8447	8447	8420	8420	8394	8394
Division	8	73	584	72	576	72	576
Sub/Add	1	9659	9659	9610	9610	9586	9586
Logarithm	20	24	480	0	0	25	500
Exponential	40	1	40	0	0	0	0
Square Root	8	2	16	1	1	0	0
Comparison	1	1	1	1	1	1	1
Total			19227		18615		19057
Grand Total			19358		18769		19203

Chapter 5

Conclusion

5.1 Summary

IN this thesis, a few practical and low complexity VADs have been presented with the following new innovations: (i) the VADs incorporate perceptual preprocessing via the use of the MFSC which is a mel based speech feature, (ii) the context based information is exploited in dectection, (iii) the PEM (or EM) model is treated as a composite hypothesis with the prior SNR as the free parameter, where it was shown that the recently proposed CNP approach is adept at modeling the prior information about the free parameter into the detector design, (iv) it is revealed that the popular Bayesian and NP design approaches are not capable of incorporating prior information about the free parameter into the detector design and (v) finally, a new strategy to design the CNP detector is proposed which makes the tuning of the CNP detector less cumbersome.

The proposed VAD schemes have been rigorously evaluated based on the overall detection (D) percentages, ROC curves, psychoacoustic parameter ABC and computational complexity. Particularly, to evaluate the CNP approach, primary detectors

based on the Bayesian, NP and CNP approaches have been developed, where it is revealed that the CNP detector performs better than the NP or Bayesian detectors as it models the prior information about the free parameter, i.e., the prior SNR of the noisy speech, into the detector design. Further, the superiority of the CNP approach has been also demonstrated via the computer simulation where it was observed that the CNP outperforms the NP and Bayesian detectors, and compares well to the AMR VADs.

We have also shown the advantage of using contextual information for detection in low SNR and highlighted the important role of speech-pause duration as the source of contextual information. Our CVAD scheme show that the addition of the contextual detection enhances detection significantly. Moreover, it has been shown that the contextual detector has very low computational complexity and a modular design which allows for a convenient integration into existing LRT based VADs, making it an excellent alternative to the contemporary hang-over schemes. In particular, the CVAD-NP has a strong speech detection capability making it a good candidate for application in VoIP, mobile telephony etc. On the other hand, the CVAD-CNP has exhibited good pause detection which makes it ideal for speech enhancement systems.

5.2 Furture Work

The proposed NP and CNP detectors are members of the general family of CNP VADs. Although, the NP and CNP detectors have delivered robust performance, further reseach is warranted to discover other members of the CNP family which may outperform the proposed detectors. Further, in this thesis, we have extensively used the first and second-order statistics of the observation for detection with good results. However, recent studies [20,43] have shown the benefits of using higher-order

statistics (HOS) such as the skewness, kurtosis etc. in VADs. Hence, one could extend the proposed PEM model such that it incorporates HOS based speech features, and thereby evaluate the efficacy of HOS in the proposed detectors.

The proposed contextual detection scheme uses the Poisson and geometric densities to model the physical attributes of the speech bursts and pauses. The dynamics of conversations obviate the fact that speech bursts and pause periods are non-stationary. However, in this thesis, these parameters have been assigned fixed values which is certainly an unrealistic assumption. Hence, in order to further improve the efficacy of contextual detection, it is necessary to develop a technique by which the values of these parameters can be learnt online, and updated regularly to track non-stationarity.

Appendix A

Simultaneous diagonalization of two matrices

Let the eigenvalue decomposition (EVD) of $\mathbf{K_n}$ be given as:

$$\mathbf{A}^T\mathbf{K_n}\mathbf{A} = \mathbf{\Lambda_n} \tag{A-1}$$

where $\mathbf{\Lambda_n}$ is the diagonal eigenvalue matrix of $\mathbf{K_n}$ and $\mathbf{A}$ is the corresponding eigenvector matrix. Now, consider a transform $\mathbf{B}$ given by:

$$\mathbf{B} = \mathbf{A}\mathbf{\Lambda_n}^{-\frac{1}{2}}. \tag{A-2}$$

It is easily seen that the transformation $\mathbf{B}$ transforms the matrix $\mathbf{K_n}$ into an identity matrix, i.e.,

$$\mathbf{B}^T\mathbf{K_n}\mathbf{B} = \mathbf{I}. \tag{A-3}$$

Now, consider the EVD of matrix $\mathbf{B}^T\mathbf{K_f}\mathbf{B}$:

$$\mathbf{C}^T(\mathbf{B}^T\mathbf{K_f}\mathbf{B})\mathbf{C} = \mathbf{\Lambda} \tag{A-4}$$

where $\mathbf{\Lambda}$ is the diagonal eigenvalue matrix of $\mathbf{B}^T\mathbf{K_f}\mathbf{B}$ and $\mathbf{C}$ is the corresponding eigenvector matrix. Now, if we form a new transform $\mathbf{Q}$ such that:

$$\mathbf{Q} = \mathbf{BC}, \tag{A-5}$$

then the (A-4) can be rewritten as:

$$\mathbf{Q}^T\mathbf{K_f}\mathbf{Q} = \mathbf{\Lambda}. \tag{A-6}$$

Also, $\mathbf{Q}$ will diagonalize $\mathbf{K_n}$ to an identity matrix as the matrix $\mathbf{C}$ is orthonormal, i.e,

$$\mathbf{B}^T\mathbf{K_n}\mathbf{B} = \mathbf{I} \tag{A-7}$$

$$\mathbf{C}^T\mathbf{B}^T\mathbf{K_n}\mathbf{BC} = \mathbf{C}^T\mathbf{IC} \tag{A-8}$$

$$\mathbf{Q}^T\mathbf{K_n}\mathbf{Q} = \mathbf{I}. \tag{A-9}$$

Hence, (A-5) gives the transform $\mathbf{Q}$ which simultaneously diagonalize $\mathbf{K_f}$ and $\mathbf{K_n}$ as per (8) and (9).

Appendix B

Properties of the sufficient statistics

Proof of the Theorem 3.2.1. From (30) and (16), it is seen that the transformation $\mathbf{Q}$ preserves the SS. Hence, the statistics of SS can be equivalently computed in $\mathbf{D^F}$ and $\mathbf{D^Z}$. Before proceeding to the proofs, we develop three identities. From (8) and (9), we get:

$$\mathbf{Q}^{-1}{\mathbf{K_f}}^{-1}\mathbf{K_n}\mathbf{Q} = \mathbf{\Lambda}^{-1}\mathbf{Q}^T(\mathbf{Q}^T)^{-1}. \tag{B-1}$$

Taking the trace (tr) on both sides, we have:

$$tr(\mathbf{Q}^{-1}{\mathbf{K_f}}^{-1}\mathbf{K_n}\mathbf{Q}) = tr(\mathbf{\Lambda}^{-1}), \tag{B-2}$$

$$tr({\mathbf{K_f}}^{-1}\mathbf{K_n}) = tr(\mathbf{\Lambda}^{-1}), \tag{B-3}$$

Similarly, from (8) and (9), we also have:

$$tr(\mathbf{K_f}{\mathbf{K_n}}^{-1}) = tr(\mathbf{\Lambda}). \tag{B-4}$$

Multiplying (B-1) with itself yields:

$$\mathbf{Q}^{-1}\mathbf{K_f}^{-1}\mathbf{K_n}\mathbf{K_f}^{-1}\mathbf{K_n}\mathbf{Q} = \mathbf{\Lambda}^{-1}\mathbf{\Lambda}^{-1}, \tag{B-5}$$

$$tr(\mathbf{K_f}^{-1}\mathbf{K_n}\mathbf{K_f}^{-1}\mathbf{K_n}) = tr(\mathbf{\Lambda}^{-2}). \tag{B-6}$$

Now, we prove of the conditional statistics of the SS in Theorem 3.2.1 by using (B-3), (B-4) and (B-6).

(i) The conditional expectations $E[l|H_0]$ is now calculated using (17) as:

$$E[l|H_0] = E[\frac{1}{2}\sum_{i=1}^{M}(1-\frac{1}{\lambda_i})z_i^2|H_0]. \tag{B-7}$$

Interchanging summation and expectation, we get:

$$E[l|H_0] = \frac{1}{2}\sum_{i=1}^{M}(1-\frac{1}{\lambda_i})E[z_i^2|H_0]. \tag{B-8}$$

From the definition of the event H_0, we know that $E[z_i^2|H_0] = 1$. Therefore, the conditional mean becomes:

$$E[l|H_0] = \frac{1}{2}\sum_{i=1}^{M}(1-\frac{1}{\lambda_i}) = \frac{1}{2}tr(\mathbf{I}-\mathbf{\Lambda}^{-1}), \tag{B-9}$$

which can be simplified using (B-3) as:

$$E[l|H_0] = \frac{1}{2}tr(\mathbf{I}-\mathbf{K_f}^{-1}\mathbf{K_n}) = \frac{1}{2}\sum_{i=1}^{M}(\frac{\zeta_i^F}{\gamma_i^F}) \tag{B-10}$$

In obtaining the last expression in (B-10), we have neglected the non-diagonal elements of $\mathbf{K_f}$ and $\mathbf{K_n}$ since they tend to be insignificant compared to the diagonal ones as a results of applying the PEM model.

(ii) Following the steps in (i), the conditional expectation $E[l|H_1]$ is calculated as:

$$E[l|H_1] = E[\frac{1}{2}\sum_{i=1}^{M}(1-\frac{1}{\lambda_i})z_i^2|H_1], \tag{B-11}$$

$$= \frac{1}{2}\sum_{i=1}^{M}(\lambda_i - 1), \tag{B-12}$$

$$= \frac{1}{2}tr(\mathbf{\Lambda} - \mathbf{I}). \tag{B-13}$$

By using (B-4), the above expression becomes:

$$E[l|H_1] = \frac{1}{2}tr(\mathbf{K_f}\mathbf{K_n}^{-1} - \mathbf{I}) = \frac{1}{2}\sum_{i=1}^{M}\zeta_i^F \tag{B-14}$$

(iii) The conditional variance $Var[l|H_0]$ is computed using the identity:

$$Var[l|H_0] = E[l^2|H_0] - E^2[l|H_0] \tag{B-15}$$

where the term $E[l^2|H_0]$ can be written as:

$$E[l^2|H_0] = E[\frac{1}{2}\sum_{i=1}^{M}(1-\frac{1}{\lambda_i})z_i^2\frac{1}{2}\sum_{j=1}^{M}(1-\frac{1}{\lambda_j})z_j^2|H_0], \tag{B-16}$$

which can be rearranged as:

$$E[l^2|H_0] = \frac{1}{4}\sum_{i=1}^{M}\sum_{j=1,j\neq i}^{M}E[z_i^2 z_j^2|H_0](1-\frac{1}{\lambda_i})(1-\frac{1}{\lambda_j}) \tag{B-17}$$

$$+ \frac{1}{4}\sum_{i=j,i=1}^{M}E[z_i^4|H_0](1-\frac{1}{\lambda_i})^2. \tag{B-18}$$

The terms $E[z_i^2 z_j^2|H_0]$ and $E[z_i^4|H_0]$ are the fourth order moments of Gaussian random variables z_i and hence can be simplified as:

$$E[z_i^2 z_j^2|H_0] = E[z_i^2|H_0]E[z_j^2|H_0] = 1, \tag{B-19}$$

$$E[z_i^4|H_0] = 3E[z_i^2|H_0] = 3. \tag{B-20}$$

Using (B-19) and (B-20) in (B-18), we get:

$$E[l^2|H_0] = \frac{1}{4}\sum_{i=1}^{M}\sum_{j=1,j\neq i}^{M}(\frac{\zeta_i^Z}{\gamma_i^Z})(\frac{\zeta_j^Z}{\gamma_j^Z}) + \frac{3}{4}\sum_{i=j,i=1}^{M}(\frac{\zeta_i^Z}{\gamma_i^Z})^2. \tag{B-21}$$

Using (B-10), the term $E^2[l|H_0]$ can be written as:

$$E^2[l|H_0] = \frac{1}{4}\sum_{i=1}^{M}\sum_{j=1,j\neq i}^{M}(\frac{\zeta_i^Z}{\gamma_i^Z})(\frac{\zeta_j^Z}{\gamma_j^Z}) + \frac{1}{4}\sum_{i=j,i=1}^{M}(\frac{\zeta_i^Z}{\gamma_i^Z})^2. \tag{B-22}$$

Now, the conditional variance can be computed by using (B-21) and (B-22) into (B-15) as:

$$\begin{aligned} Var[l|H_0] &= \frac{1}{2}\sum_{i=1}^{M}(\frac{\zeta_i^Z}{\gamma_i^Z})^2 = \frac{1}{2}tr((\mathbf{I}-\mathbf{\Lambda}^{-1})(\mathbf{I}-\mathbf{\Lambda}^{-1})), \\ &= \frac{1}{2}tr((\mathbf{I}-2\mathbf{\Lambda}^{-1}+\mathbf{\Lambda}^{-2}). \end{aligned} \tag{B-23}$$

Using (B-3) and (B-6), we obtain:

$$Var[l|H_0] = \frac{1}{2}tr((\mathbf{I}-\mathbf{K_f}^{-1}\mathbf{K_n})^2) = \frac{1}{2}\sum_{i=1}^{M}(\frac{\zeta_i^F}{\gamma_i^F})^2 \tag{B-24}$$

(iv) Finally using (B-10), (B-14) and (B-24) in the definition of d in (27), we obtain:

$$d^2 = \frac{(E[l|H_1] - E[l|H_0])^2}{Var[l|H_0]} = \frac{(\sum_{i=1}^{M} \frac{(\zeta_i^F)^2}{\gamma_i^F})^2}{2\sum_{i=1}^{M}(\frac{\zeta_i^F}{\gamma_i^F})^2}$$

□

Appendix C

Bounds on the Bayesian threshold

By using (20) and (21), the term $\sum_{i=1}^{M} \ln(\zeta_i^{\mathbf{Z}} + 1)$ in (39) can be simplified as:

$$\sum_{i=1}^{M} \ln(\zeta_i^{\mathbf{Z}} + 1) = \sum_{i=1}^{M} \ln(\gamma_i^{\mathbf{Z}}) = \sum_{i=1}^{M} \ln \lambda_i. \tag{C-1}$$

Using a well known logarithm identity, i.e.,:

$$\ln(x) \leq (x - 1), \text{ where } x > 0. \tag{C-2}$$

we get

$$\ln(\lambda_i) \leq (\lambda_i - 1), \tag{C-3}$$

since $\lambda_i \geq 1$. Now, summing this result over $i = 1, 2, ..., M$, and multiplying both sides by $\frac{1}{2}$, we get the RHS of (39), i.e.,

$$\frac{1}{2} \sum_{i=1}^{M} \ln(\lambda_i) \leq \frac{1}{2} \sum_{i=1}^{M} (\lambda_i - 1). \tag{C-4}$$

Next, since the reciprocal of λ_i is strictly a non-zero positive number $\forall i$, we can use the log-identity to obtain:

$$\ln(\frac{1}{\lambda_i}) \leq \frac{1}{\lambda_i} - 1, \tag{C-5}$$

which is further simplified to get the LHS in (39):

$$\frac{1}{2}\sum_{i=1}^{M} \ln(\lambda_i) \;\; \geq \;\; \frac{1}{2}\sum_{i=1}^{M}(1 - \frac{1}{\lambda_i}). \tag{C-6}$$

By using (C-1), the inequalities in (C-4) and (C-6) can be written as:

$$\frac{1}{2}\sum_{i=1}^{M}(1 - \frac{1}{\lambda_i}) \leq \ln(\lambda_i) \leq (\lambda_i - 1). \tag{C-7}$$

Now, (C-7) is further simplified by using (B-9) and (B-12), i.e.,

$$E[l|H_0] \leq \ln(\lambda_i) \leq E[L|H_1], \tag{C-8}$$

which gives the desired result.

Appendix D

Conditional variance of the sufficient statistics

Using (71), the conditional mean $E[l|H_1]$ becomes:

$$E[l|H_1] = \frac{1}{2}\sum_{i=1}^{M}(\frac{1}{\lambda_i'} - 1). \tag{D-1}$$

Using (B-15), and following the steps as in (B-16) to (B-22), one can obtain the conditional variance $Var[l|H_1]$ below:

$$Var[l|H_1] = \frac{1}{2}\sum_{i=1}^{M}(\frac{1}{\lambda_i'} - 1)^2, \tag{D-2}$$

which can be rewritten in the vector form as:

$$Var[l|H_1] = \frac{1}{2}tr((\mathbf{\Lambda}'^{-1} - \mathbf{I})^2). \tag{D-3}$$

Using (69) and (70) and following the steps in (B-1) to (B-3), we get:

$$tr(\mathbf{\Lambda'}^{-1}) \;=\; tr(\mathbf{K_n}^{-1}\mathbf{K_f}). \tag{D-4}$$

Using (D-4) into (D-3), we obtain:

$$Var[l|H_1] = \frac{1}{2}tr((\mathbf{K_n}^{-1}\mathbf{K_f} - \mathbf{I})^2) = \frac{1}{2}\sum_{i=1}^{M}(\zeta_i^F)^2.$$

References

[1] Abhijeet Sangwan, Chiranth M.C., Rahul Sah, Vishal Gaurav, and R.V. Prasad, "Voice activity detection for VoIP - time and frequency domain solutions," in *Tenth annual IEEE Symposium on Multimedia, Communications and Signal Processing.* IEEE Bangalore Chapter, Nov 2001, pp. 20–23, IEEE Bangalore Chapter.

[2] S. Gokhun Tanyer and Hamza Ozer, "Voice activity detection in nonstationary noise," *IEEE Transactions on Speech and Audio Processing*, vol. 8, no. 4, July 2000.

[3] K. El-Maleh and P. Kabal, "Comparison of voice activity detection alogorithms for wireless personal communications systems," in *IEEE Canadian Conference on Electrical and Computer Engineering*, May 1997, pp. 470–373.

[4] J.H. James, Bing Chen, and Laurie Garrison, "Implementing VoIP: A voice transmission performance progress report," *IEEE Communications Magazine*, pp. 36–41, July 2004.

[5] Jane-Hwa Huang, Szu-Lin Su, and Jiann-Horng Chen, "Design and performance analysis for data transmission in GSM/GPRS system with voice activity detection," *IEEE Transaction on Vehicular Technology*, vol. 51, no. 4, pp. 648–656, July 2002.

[6] Majeed Abdulrahman, Ansar U.H. Sheikh, and David D. Falconer, "Decision feedback equalization for CDMA in indoor wireless communications," *IEEE Transaction on Selected Areas in Communications*, vol. 12, no. 6, pp. 698–706, May 1994.

[7] A.M. Kondoz and B.G. Evans, "A high quality voice coder with integrated echo canceller and voice activity detector for VSAT systems," in *3rd European Conference on Satellite-Communications (ECSC)*, Nov 1993, pp. 196–200.

[8] I.D. Lee, H.P. Stern, and S.A. Mahmoud, "A voice activity detection algorithm for communication systems with dynamically varying background acoustic noise," in *48th IEEE Conference on Vehicular Technology (VTC 98)*, May 1998, pp. 1214–1218.

[9] R.V. Prasad, H.S. Jamadagni, Abhijeet Sangwan, and Chiranth M.C., "VAD for VoIP using cepstrum," in *Sixth IEEE International Conference on High Speech Networks and Multimedia Communications*, Estoril, Portugal, July 2003, pp. 522–530.

[10] Francesco Beritelli, Salvatore Casale, and Salvatore Serrano, "A low-complexity speech-pause detection algorithm for communication in noisy environments," *European Transaction on Telecommunications*, vol. 15, pp. 33–38, 2004.

[11] Paul Newson and Mark R. Heath, "The capacity of a spead spectrum CDMA system for cellular mobile radio with consideration of system imperfections," *IEEE Transactions of Selected Areas in Communications*, vol. 12, no. 4, pp. 673–684, May 1994.

[12] Javier Ramirez, Jose C. Segura, Carmen Benitez, Luz Garcia, and Antonio Rubio, "Statistical voice activity detection using a multiple observation likelihood

ratio test," *IEEE Signal Processing Letters*, vol. 12, no. 10, pp. 689–692, Oct 2005.

[13] Adil Benyassine, Eyal Shlomot, Huan-Yu Su, Dominique Massaloux, Claude Lamblin, and Jean-Pierre Petit, "ITU-T recommendation G.729 annex B: A silence compression scheme for use with G.729 optimized for v.70 digital simultaneous voice and data applications," *IEEE Communications Magazine*, pp. 64–73, Sept. 1997.

[14] 3GPP SA 4, "Digital cellular telecommunications system (phase 2+); universal mobile telecommunications system (UMTS); adaptive multi-rate (AMR) speech codec; (3GPP TS 26.102 version 6.0.0 release 6)," Tech. Rep. TS 126 102, ETSI, Jan 2005.

[15] J. Sohn, N. Kim, and W. Sung, "A statistical model-based voice activity detector," *IEEE Signal Processing Letters*, vol. 6, no. 1, Jan 1999.

[16] Alan Davis and Sven Nordholm, "A low complexity statistical voice activity detector with performance comparisons to ITU-T/ETSI voice activity detectors," in *Fourth International Conference on Information, Communications and Signal Processing (ICICS)*, Singapore, Dec 2003, pp. 15–18.

[17] Beena Ahmed and W. Harvey Holmes, "A voice activity detector using chi-square test," in *International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, May 2004, pp. 625–628.

[18] Joon-Hyuk Chang and Nam Soo Kim, "Voice activity detection based on complex Laplacian model," *IEEE Signal Processing Letters*, vol. 39, no. 7, April 2003.

[19] Saeed Gazor and Wei Zhang, "A soft voice activity detector based on a Laplacian-Gaussian model," *IEEE Transaction on Speech and Audio Processing*, vol. 11, no. 5, pp. 498–505, Sept 2003.

[20] Elias Nemer, Rafik Goubran, and Samy Mohmoud, "Robust voice activity detection using higher-order statistics in the LPC residual domain," *IEEE Transactions on Speech and Audio Processing*, vol. 9, no. 3, Mar 2001.

[21] A. Sangwan, W.P. Zhu, and M.O. Ahmad, "Improved voice activity detection via contextual information and noise suppression," in *IEEE International Symposium on Circuits and Systems (ISCAS 05)*, May 2005, pp. 868–871.

[22] Y. Ephraim and D. Malah, "Speech enhancement using a minimum-mean square error short-time spectral amplitude estimator," *IEEE Transaction on Acoustics, Speech and Signal Processing*, vol. 32, no. 6, pp. 1109–1121, Dec 1984.

[23] Harry L. Van Trees, *Detection, Estimation and Modulation Theory, Part 1*, John Wiley and Sons Inc., 2001.

[24] A. Dogandzic and A. Nehorai, "Detecting a dipole source by MEG/EEG and generalized likelihood ratio tests," in *30th Asilomar Conference on Signals, Systems and Computers*, Nov 96, pp. 1196–1200.

[25] J.T. Neyhart, R.E. Eckert, R. Polikar, S. Mandayam, and M. Tseng, "A modified Neyman-Pearson technique for radiodense tissue estimation in digitized mammograms," in *2nd Conf. on Engineering in Medicine and Biology (EMBS)*, Oct. 2002, pp. 995–996.

[26] E. Levitan and N. Merhav, "A competitive Neyman-Pearson approach to universal hypothesis testing with applications," *IEEE Transaction on Information Theory*, vol. 48, no. 8, pp. 2215–2229, Aug. 02.

[27] Douglas O'Shaughnessy, *Speech Communications: Human and Machine*, Wiley-IEEE Press, 2 edition, Nov. 1999.

[28] Gunnar Fant, Anita Kruckenburg, and Joana Barbosa Ferreira, "Individual variations in pausing. a study in read speech," in *PHONUM 9*. Umea University, Department of Philiosophy and Linguistics, 2003, pp. 193–196.

[29] Sofia Gustafson-Capcova and Beata Megyesi, "A comparative study of pauses in dialogues and read speech," in *Eurospeech*, 2001, pp. 931–935.

[30] working party 15/1 Study group 15, "Experts group on very low bitrate video telephony," Tech. Rep., ITU telecommunications standardization sector, Feb 1997.

[31] Mark Marzinzik and Birger Kollmeier, "Speech pause detection for noise spectrum estimation by tracking power envelope dynamics," *IEEE Trans. on Speech and Audio Processing*, vol. 10, no. 2, Feb 2002.

[32] R.V. Prasad, Abhijeet Sangwan, H.S. Jamadagni, Chiranth M.C., Rahul Sah, and Vishal Gaurav, "Comparison of voice activity detection algorithms for VoIP," in *International symposium on computers and communications (ISCC)*, Taormina, Italy, July 2002, pp. 530–535.

[33] Sergei Skorik and Frederic Berthommier, "On a cepstrum-based speech detector robust to white noise," in *International Conference on Speech and Computer (SPECOM)*, St. Petersburg, Sept 2000.

[34] R. Tucker, "Voice activity detection using a periodicity measure," *Proceedings of the Institute of Electrical Engineers*, vol. 139, no. 4, pp. 377–380, Aug 1992.

[35] J.C. Junqua, B. Reaves, and B. Mak, "A study of endpoint detection algorithms in adverse conditions: Incidence on a DTW and HMM recognize," in *Eurospeech'91*, 1991, pp. 321–324.

[36] F. Beritelli, S. Casale, G. Ruggeri, and S. Serrano, "Performance evaluation and comparison of G.729/AMR/fuzzy voice activity detectors," *IEEE Signal Processing Letters*, vol. 9, no. 3, pp. 85–88, March 2002.

[37] Yong Duk Cho and Ahmet Kondoz, "Analysis and improvement of a statistical model-based voice activity detector," *IEEE Signal Processing Letters*, vol. 8, no. 10, Oct 2001.

[38] J. Sohn and W. Sung, "A voice activity detector employing soft decision based noise spectrum adaption," in *International Conference on Acoustics, Speech and Signal Processin*, May 1998, pp. 365–368.

[39] Jen-Tzung Chien and Chih-Hsien Huang, "Bayesian learning of speech duration models," *IEEE Transactions of Speech and Audio Processing*, vol. 11, no. 6, Nov 2003.

[40] S.B. Searle, *Matrix Algebra useful for Statistics*, New York: Wiley, 1982.

[41] F. Beritelli, S. Casale, and G. Ruggeri, "A psychoacoustic auditory model to evaluate the performance of a voice activity detector," in *International Conference on Signal Processing (ICSP 2000)*, Aug. 2000, pp. 807–810.

[42] Tom Minka, "The lightspeed matlab toolbox, version 2.0," http://research.microsoft.com/~minka/software/lightspeed/, Aug 2005.

[43] Ke Li, M.N.S Swamy, and M.O. Ahmad, "An improved voice activity detection using higher order statistics," *IEEE Transactions on Speech and Audio Processing*, vol. 13, no. 5, pp. 965–974, Sept. 2005.