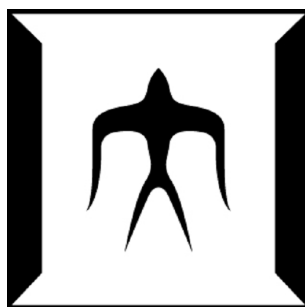


論文 / 著書情報
Article / Book Information

題目(和文)	
Title(English)	Robust Speech Recognition Based on Non-extensive Statistics
著者(和文)	ヒルマンパ ルデ ット
Author(English)	Hilman Pardede
出典(和文)	学位:博士(工学), 学位授与機関:東京工業大学, 報告番号:甲第9248号, 授与年月日:2013年6月30日, 学位の種別:課程博士, 審査員:篠田 浩一,佐藤 泰介,村田 剛志,杉山 将,藤井 敦
Citation(English)	Degree:Doctor (Engineering), Conferring organization: Tokyo Institute of Technology, Report number:甲第9248号, Conferred date:2013/6/30, Degree Type:Course doctor, Examiner:,,,,
学位種別(和文)	博士論文
Type(English)	Doctoral Thesis

Robust Speech Recognition Based on Non-extensive Statistics



Hilman Ferdinandus Pardede

Department of Computer Science

Graduate School of Information Science and Engineering

Tokyo Institute of Technology

Dissertation submitted to the Tokyo Institute of Technology
for the degree of Doctor of Engineering

2013

Abstract

Statistical speech recognition techniques have been able to achieve a good performance in clean environments. However, they are not robust against environmental noise. Environmental noises can be categorized into two: additive and convolutive noise. Most compensation methods to improve the robustness of speech recognition systems in noisy environments are derived in an extensive framework, which is based on Boltzmann-Gibbs statistics and Shannon entropy. Under this framework, it is assumed that noise and speech are uncorrelated. Accordingly, Gaussian distributions are used to model speech, noise, and noisy speech since they maximize Shannon entropy. However, these assumptions are generally not true. The use of a limited window in signal processing introduces a cross-term between speech and noise in the power spectral domain, hence they are nonadditive. In addition, it is well known that the short time speech spectra do not follow Gaussian distributions due to the complexity of a speech pattern. Therefore, noisy speech observed in real situations often follows a heavy-tailed distribution, not a Gaussian distribution.

In this thesis, we implement a non-extensive framework for robust speech recognition to tackle the problem of the cross-term and non-Gaussianity of noisy speech. We proposed two methods: q -log spectral mean normalization (q -LSMN) and q -spectral subtraction (q -SS). In q -LSMN, we utilize the q -logarithmic (q -log) function of the non-extensive statistics and introduce the q -log spectral domain, which is a non-extensive domain. We derive a normalization method, q -LSMN, which is an extension of log spectral mean normalization (LSMN) to this domain. By doing so, we not only remove convolutive noise but also reduce the cross-term. The recognition experiments on a synthesized noisy speech database, the Aurora-2 database,

showed that q -LSMN was consistently better than conventional normalization methods, such as CMN, LSMN, and MVN. Furthermore, q -LSMN was even more effective for speech when applied to a real noisy environment in the CENSREC-2 database. It significantly outperformed ETSI AFE front-end.

Our second method, q -SS, is a spectral subtraction method derived on the non-extensive framework. Spectral subtraction is a popular method to remove additive noise. We introduce a q -Gaussian distribution of non-extensive statistics, which is a heavy-tailed distribution, to represent the distribution of noisy speech and derive a new spectral subtraction method based on it. In our analysis, the q -Gaussian distributions fit the noisy speech spectra better than the Gaussian distributions do. Our speech recognition experiments showed that q -SS outperformed the conventional spectral subtraction method on the Aurora-2 and the CENSREC-2 tasks.

"The Lord God is my strength, and he will make my feet like hinds' feet,
and he will make me to walk upon mine high places." (Habakkuk 3:19)

Acknowledgements

I would like to express my sincere thanks to my supervisor Prof. Koichi Shinoda for giving me a chance to study in his laboratory and for his invaluable guidance and advice he has offered me during these years. I would also like to extend many thanks to A. Prof. Koji Iwano, for the support and help he provided during my Ph.D course, for all advise and time he spent discussing my work.

I also want to thank my colleagues in Shinoda Laboratory especially to Johan Rohdin and Ryan Price for helping me whenever necessary discussing my work and proofreading my writings. Thank you to other students of my laboratory for their friendship and support during my doctoral course. You all have been my family here during my stay in Japan.

I want to thank my family in Indonesia for their constant love and encouragement. To my girlfriend Ing, who encourages me in my difficult time during this study. You have been a great support even though we are separated by thousands of miles. Thank you for being patient with me even during the most difficult period.

I would like to thank The Ministry of Education, Culture, Sports, Science, and Technology (MEXT) of Japan, who grant me the scholarships that enable me to continue my study in Japan.

I dedicated this work to my Lord Jesus. You give me strength when I am weak, You raise me up when I am down. To You all glory and praise.

Declaration

I declare that this doctoral thesis is my own work and has not been submitted in any form for another degree or diploma at any university or other institution of tertiary education. Information derived from the published or unpublished work of others has been acknowledged in the text and a list of references is given.

The thesis work was conducted from 2009 to 2013 under the supervision of Prof. Koichi Shinoda at Tokyo Institute of Technology

Tokyo, April 23, 2013

Hilman F. Pardede

Contents

List of Figures	xi
List of Tables	xv
1 Introduction	1
2 Statistical speech recognition: an overview	9
2.1 Introduction	9
2.2 MFCC feature extraction	11
2.3 HMM-based statistical speech recognition	15
3 Noise robust speech recognition	23
3.1 The effect of noise on the speech features	24
3.2 Noise robust speech recognition methods	25
3.2.1 Model compensation approach	26
3.2.1.1 Parallel model combination (PMC)	26
3.2.1.2 Vector Taylor series (VTS) for model compensation . .	28
3.2.2 Feature enhancement	29
3.2.2.1 Spectral subtraction	29
3.2.2.2 MMSE-based feature enhancement	30
3.2.2.3 Feature normalization	33
3.2.2.4 Vector Taylor series (VTS) for feature enhancement . .	34
3.2.2.5 ETSI advanced front end (AFE)	35
3.2.3 Dealing with the cross-term	37
3.2.3.1 Non-linear spectral subtraction	38
3.2.3.2 MMSE-based methods using non Gaussian distributions	39

3.2.3.3	Phase sensitive VTS	39
3.3	Summary	40
4	Non-extensive statistics: a review	43
4.1	Introduction	43
4.2	Non-extensive functions	44
4.2.1	Definitions	44
4.2.2	Properties	45
4.3	Tsallis entropy	47
4.4	Tsallis distributions	47
4.4.1	Q -Gaussian distribution	48
4.4.2	Q -Exponential distribution	49
4.5	Implementations	50
4.6	Summary	51
5	Feature normalization based on non-extensive statistics	53
5.1	Introduction	53
5.2	Non-additivity in noisy speech features	54
5.3	Framework	56
5.4	Q -Log spectral mean normalization (Q -LSMN)	58
5.5	Spectral Subtraction	61
5.6	Experimental Setup	62
5.6.1	Databases	62
5.6.1.1	The Aurora-2 database	62
5.6.1.2	The CENSREC-2 database	63
5.6.2	ASR configuration	64
5.7	Experimental results and discussions	64
5.7.1	The effect of q -LSMN on noisy speech	64
5.7.2	Evaluation of q -LSMN without spectral subtraction	68
5.7.3	Evaluation of q -LSMN with spectral subtraction	70
5.8	Conclusion	73

6	Q-Gaussian based spectral subtraction	75
6.1	Introduction	75
6.2	Spectral subtraction	77
6.2.1	Derivation	77
6.2.2	Nonlinear spectral subtraction (NSS)	78
6.3	Q -Spectral subtraction (q -SS)	79
6.3.1	Derivation	79
6.3.2	Relation to nonlinear spectral subtraction	80
6.3.3	Q -SS based on the optimum q -Gaussian distribution	81
6.3.4	Noise estimation	83
6.3.5	Flooring	84
6.4	Experimental setup	85
6.5	Experimental results and discussions	85
6.5.1	Q -Gaussian representation of noisy speech	85
6.5.2	Performance of q -SS	86
6.5.3	Comparison with NSS	91
6.6	Conclusions	95
7	Conclusions and future works	97
7.1	Conclusions	97
7.2	Future works	99
	References	103

List of Figures

1.1	Environmental distortion model with additive and convolutive noise . . .	2
2.1	The block diagram of an HMM-based speech recognition system	10
2.2	MFCC feature extraction	11
2.3	Mel-scale and the implementation of triangular filterbanks	14
2.4	An HMM with three true states. The initial and final states are used to tie with other HMM and do not model speech.	17
3.1	Parallel Model Combination	27
3.2	Noise reduction scheme in ETSI AFE	36
4.1	The q -logarithmic function of real variable x for different q -values. . . .	44
4.2	The q -exponential function of real variable x for different q -values. . . .	45
4.3	q -Gaussian distribution for several q	49
4.4	q -exponential distribution for several q and λ	50
5.1	Block diagram of the proposed front-end.	57
5.2	The log spectral distance between “filtered speech”, speech affected by convolutive noise, and the clean speech after applying q -LSMN.	65
5.3	The log spectral distance between the clean speech estimate and the clean speech with and without applying q -LSMN for the spectral valleys. The clean speech estimate is obtained after employing spectral subtraction and normalization on the noisy speech spectra. In this figure, babble noise is used as additive noise.	66

5.4	The log spectral distance between the clean speech estimate and the clean speech with and without applying q -LSMN for the spectral valleys. The clean speech estimate is obtained after employing spectral subtraction and normalization on the noisy speech spectra. In this figure, white noise is used as additive noise.	67
5.5	The word accuracies (%) of q -LSMN for filtered speech, i.e speech contaminated with convolutive noise only.	67
5.6	The word accuracies (%) of q -LSMN with and without spectral subtraction for the Aurora-2 task for different q values.	69
5.7	The word accuracies (%) of q -LSMN for various SNR conditions in the Aurora-2 task.	70
5.8	The optimum q -values for four types of noise in Test Set A of the Aurora-2 database for various SNR conditions.	71
5.9	The word accuracies (%) of q -LSMN with and without spectral subtraction for CENSREC-2 task with different q values.	72
5.10	The word accuracies (%) of q -LSMN for CENSREC-2 task for each condition evaluation with different q values.	73
6.1	The scatter plot between $\tilde{q}$, the q -value that fits the empirical distribution of noisy speech based on the MMSE criterion, and the SNR condition. The solid line is the relation between $\tilde{q}$ and the SNR based on linear regression method.	81
6.2	The optimal q -values for different SNR conditions.	85
6.3	Gaussian and q -Gaussian distributions fitted to the histogram of the speech data corrupted with subway noise at 0 dB SNR.	86
6.4	Performance of q -SS_c for $1 \leq q \leq 2$ for the Aurora-2 database. The average values over 0-20 dB of the SNR conditions of all the noise types are shown	87
6.5	Performance of q -SS_c for $1 \leq q \leq 2$ for the CENSREC-2 database. The average values over all the evaluation conditions are shown	87
6.6	Word Accuracy of q -SS_c for different SNR conditions of the Aurora-2 database. The averaged values over all the noise types are shown.	89

6.7	Word Accuracy of q -SS_c for each evaluation condition of the CENSREC-2 database.	89
6.8	Performance comparison of q -SS_c and NSS_c for $1 \leq q \leq 2$ and its equivalent α value for Aurora-2. The average values over 0-20 dB of the SNR conditions of all the noise types are shown.	92
6.9	Performance comparison of q -SS_c and NSS_c for $1 \leq q \leq 2$ and its equivalent α value for CENSREC-2. The average values over all the evaluation conditions are shown.	92
6.10	The comparison of the attenuation curve of q -SS_c and NSS_c for several values of q and their respective α value.	93

List of Tables

5.1	Word accuracy (%) of the Aurora-2 task when the cross-term is assumed to be zero.	68
5.2	The word accuracy (%) of the Aurora-2 task.	71
5.3	The word accuracy (%) of the CENSREC-2 task.	71
5.4	The word accuracy (%) of q -LSMN and the other methods after spectral subtraction for the Aurora-2 task.	72
5.5	The word accuracy (%) of q -LSMN and the other methods after spectral subtraction for the CENSREC-2 task.	73
6.1	Performance comparison (Word Accuracy (%)) of q -SS_c ($q = 1.88$) and q -SS_m for different types of noise and SNR conditions of the of Aurora-2 database	90
6.2	Performance comparison (Word Accuracy (%)) of q -SS_c ($q = 1.88$) and q -SS_m for each test set in Aurora-2.	90
6.3	Performance comparison (Word Accuracy (%)) of q -SS_c ($q = 1.87$) and q -SS_m for each evaluation condition in CENSREC-2.	90
6.4	The performance comparison (Word accuracy (%)) of several spectral subtraction methods for each SNR conditions of the Aurora-2 database. The average values over all noise types are shown.	94
6.5	The performance comparison (Word accuracy (%)) of several spectral subtraction methods for each evaluation condition of the CENSREC-2 database.	94

Chapter 1

Introduction

Currently, statistical approach is the state-of-the-art of the speech recognition technology. Advances in computer and communication technologies in recent decades enable us to utilize more resources in statistical speech recognition. Using such large resources, statistical speech recognition techniques have greatly improved. At present, word accuracies greater than 90% have been achieved for read speech in quiet environments [Lee, 1988; Woodland et al., 1994; Young, 1996].

Statistical speech recognition is based on information theory, which is an *extensive* framework based on Boltzmann-Gibbs statistics and Shannon entropy. In an extensive framework, it is assumed that subsystems derived from a system are independent from each other, and thus, the additivity between them holds. It is well known that exponential distributions, including Gaussian distributions, maximize Shannon entropy under appropriate constraints. For that reason, Gaussian distributions are often used to model speech in this framework.

Despite the great progress in the speech recognition technologies, many difficult problems in speech recognition still remain. The present speech recognition technologies are not robust against variability in speech signals. The sources of this variability are the speaker individuality, e.g. different speakers, speaking styles, age, and sex, and environmental noise, e.g. additive background noise, channel distortion, reverberation. The focus of this thesis is to improve the robustness of speech recognition against environmental noise. This field has been an active area of research for the last few decades.

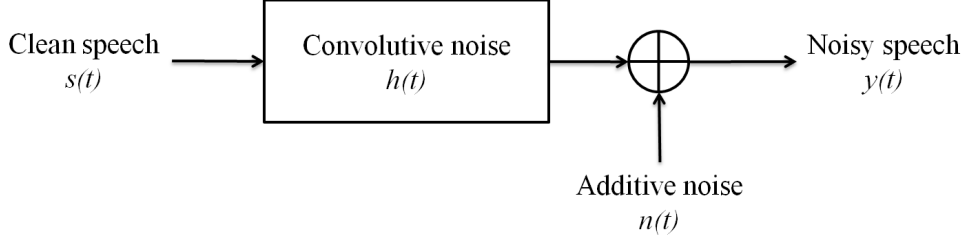


Figure 1.1: Environmental distortion model with additive and convolutive noise

Figure 1.1 illustrates the sources of environmental noise that affect the speech signals. Generally, these sources are classified into two categories: additive noise and convolutive noise. Examples of additive noise are street noise, train noise, computer fan, and the voice of other persons. Examples of convolutive noise are reverberation and channel distortions.

Common features used for speech recognition, such as Mel frequency cepstral coefficient (MFCC) and perceptual linear prediction (PLP), are derived from the short-time power spectra of speech. In short-time processing such as short-time Fourier transform (STFT), the speech signal is processed in blocks over which speech is assumed to be stationary. This block of speech is called a frame. The length of a frame is usually between 20-30 ms. In the time domain, the relation between noisy speech $y(t)$, clean speech $s(t)$, additive noise $n(t)$ and convolutive noise $h(t)$ can be written as the following:

$$y(t) = s(t) * h(t) + n(t). \quad (1.1)$$

Denoting $x(t) = s(t) * h(t)$, we can write (1.1) as the following:

$$y(t) = x(t) + n(t). \quad (1.2)$$

By taking STFT, we can represent (1.2) in the frequency domain as follows:

$$Y(m, k) = X(m, k) + N(m, k), \quad (1.3)$$

where k is the index of frequency bin (a total frequency components $K = 256$) and m

is the frame index and:

$$X(m, k) = |X(m, k)| \exp(j\theta_X(m, k)), \quad (1.4)$$

$$N(m, k) = |N(m, k)| \exp(j\theta_N(m, k)). \quad (1.5)$$

$|X(m, k)|$, $|N(m, k)|$ are the magnitude spectra, and $\theta_X(m, k)$, $\theta_N(m, k)$ are the phase spectra of filtered speech $x(t)$ and additive noise $n(t)$ respectively. From (1.3), we obtain the power spectral representation of noisy speech as follows:

$$\begin{aligned} |Y(m, k)|^2 &= |X(m, k) + N(m, k)|^2 \\ &= |X(m, k)|^2 + |N(m, k)|^2 + 2\text{Re}[X(m, k)N^*(m, k)], \end{aligned} \quad (1.6)$$

where $N^*(m, k)$ is the complex conjugate of $N(m, k)$. Substituting (1.4) and (1.5) into (1.6), we obtain:

$$\begin{aligned} |Y(m, k)|^2 &= |X(m, k)|^2 + |N(m, k)|^2 \\ &\quad + 2|X(m, k)||N(m, k)| \cos(\theta_X(m, k) - \theta_N(m, k)), \end{aligned} \quad (1.7)$$

where $\theta_X(m, k) - \theta_N(m, k)$ is the phase difference between $X(m, k)$ and $N(m, k)$. The last term of Eq. (1.7) is called a cross-term, which depends on the phase difference between speech and noise. Eq. (1.7) shows noise alters the power spectra of speech, hence distorts the speech features. This distortion causes a mismatch between quiet training conditions and recognition conditions, which are noisy. This mismatch significantly degrades the accuracy of speech recognition.

A number of methods have been developed to reduce this mismatch and improve the robustness of speech recognition in noisy environments. Their examples are spectral subtraction [Boll, 1979], cepstral mean normalization(CMN) [Furui, 1981], vector Taylor series (VTS) [Moreno et al., 1996], and parallel model combination (PMC) [Gales and Young, 1996]. Most compensation techniques are also derived on the extensive framework. Under this framework, it is assumed that noise and speech are uncorrelated. Therefore, the cross-term is ignored. Accordingly, the linear addition between speech and noise holds in the power spectral domain. Furthermore, Gaussian distributions are used to model speech and noise. Since speech and noise are assumed to follow

Gaussian distributions and are independent to each other, noisy speech is also assumed to follow a Gaussian distribution.

The extensive framework however, can be applied only to those systems which have a known structure (e.g., the number of subsystems) and relation between subsystems. It fails to explain some phenomena in *complex* systems. In a complex system, we do not know about the sub-systems and their relations. In such a system, the additive property does not hold. In addition, complex systems exhibit non-Gaussian distributions. Many of them have shown to follow heavy-tailed distributions.

A speech pattern is an example of complex systems [William A. Kretzschmar, 2009]. In a speech pattern, various long-term correlations exist among its different spectral components in complex ways in various time scales. Therefore, it is not surprising that the short-time speech spectra do not follow Gaussian distributions [Wilbur B. Davenport, 1952] but show heavy-tailed distributions instead. For this reason, it is very likely that the short-time spectra of noisy speech do not follow Gaussian distributions either. In addition, several studies have shown that the cross-term does exist in the short-time power spectra [Jeong and Williams, 1992; Kadambe and Boudreaux-Bartels, 1992]. This cross-term has been shown to significantly degrade the performance of speech recognition [Deng et al., 2004; Faubel et al., 2008]. These inaccurate assumptions hamper the performance of robust speech recognition methods. It is found that the same accuracy as when using clean speech cannot be obtained even when the noise spectra are accurately estimated [Deng et al., 2004; Evans et al., 2006; Zhu and Alwan, 2002].

To tackle these problems in the present framework of the extensive statistics, we need to factorize the elements of a speech pattern and to find the relation between them. Several studies have focussed on finding the long-term correlations among different spectral components of speech [Cohen, 2005; McAuley et al., 2005; Ming et al., 1996]. In robust speech recognition studies, there has been an increasing interest to include the cross-term on deriving robust speech recognition methods [Deng et al., 2004; Faubel et al., 2008; Li et al., 2009; Lu and Loizou, 2008; Zhu and Alwan, 2002]. For this purpose, the phase difference between speech and noise needs to be estimated. However, it is difficult to estimate the speech phase in a short window [Alsteris and Paliwal, 2007]. Nevertheless, the inclusion of the cross-term has shown to further improve the performance of robust speech recognition methods.

In this thesis, we propose an entirely different approach to solve these problems. We introduce a new framework, a non-extensive one, for robust speech recognition and try to recapture the problems in a new framework. This framework is adopted from a theory of non-extensive statistics in statistical mechanics.

Recently, a theory of non-extensive statistics has been introduced to explain several phenomena in complex systems [Tsallis, 1988]. This framework is a generalization of Boltzmann-Gibbs statistics. This framework relies on two non-extensive functions, q -exponential (q -exp) and q -logarithmic (q -log) functions. In this framework, the entropy is redefined, which is called Tsallis entropy. This entropy is a generalization of Shannon entropy and a non-extensive one. Also in this framework, heavy-tailed type distributions can be derived. One of them is called the q -Gaussian distribution, which is obtained by maximizing Tsallis entropy in a similar way Gaussian distribution can be derived from Shannon entropy. The q -Gaussian distribution is a generalization of the Gaussian distribution. The non-extensive framework has successfully represented many phenomena in complex systems in statistical mechanics, economics, finance, biology, astronomy and machine learning.

In this thesis, we propose two methods for robust speech recognition based on the non-extensive framework. The first method is q -log spectral mean normalization (q -LSMN), which is a feature normalization method. Feature normalization is a well established method to remove convolutive noise. In q -LSMN, we address the non-additivity problem caused by the existence of the cross-term in the speech features by implementing the q -log function. Our second method is q -spectral subtraction (q -SS), which is a spectral subtraction method. Spectral subtraction is a popular method for removing additive noise. In q -SS, we deal with non-Gaussianity of noisy speech spectra by implementing the q -Gaussian distribution to model the noisy speech spectra.

Conventional feature normalization methods are usually performed either in the cepstral domain [Atal, 1974; Furui, 1981] or the log spectral domain [Avendano and Hermansky, 1997]. In the conventional approaches, the logarithmic function is utilized to transform a convolutive relation between speech and convolutive noise in the time domain into an additive one in the log spectral or cepstral domains. However, the cross-term introduces a nonlinearity between speech and convolutive noise. It cannot be removed by a simple linear normalization method. In q -LSMN, we utilize the q -log function and its properties. We introduce a new domain, the q -log spectral domain,

to model non-additivity in noisy speech features. We derive a feature normalization technique in this domain, which is an extension of the log spectral mean normalization (LSMN) [Avendano and Hermansky, 1997] to the q -log spectral domain. Our method not only normalizes speech features but also removes the cross-term. Our experiments on a synthesized noisy database, Aurora-2, and a real noisy environment database, CENSREC-2, confirmed the effectiveness of q -LSMN compared to the traditional feature normalization methods.

In spectral subtraction, it is assumed that noise and speech are uncorrelated and follow Gaussian distributions. Thus, noisy speech also follows a Gaussian distribution. Spectral subtraction can be derived by maximizing the likelihood of noisy speech with respect to its variance. In q -SS, we assume that noisy speech follows the q -Gaussian distribution. Then, we derive q -SS in a similar way as spectral subtraction is derived. We found that the q -Gaussian distribution fits the noisy speech distribution better than the Gaussian distribution does. Our speech recognition experiments confirmed the effectiveness of q -SS compared to the conventional spectral subtraction method on the Aurora-2 and the CENSREC-2 tasks.

The remainder of this thesis is organized as follows:

In Chapter 2, we present the fundamental theory of statistical speech recognition. This chapter includes the extraction process for Mel frequency cepstral coefficients (MFCCs), the acoustic modeling using hidden Markov model (HMM), and the decoding using Viterbi algorithm.

In Chapter 3, we review some previous studies of robust speech recognition related to our studies. First, we describe several popular methods in robust speech recognition which are derived in the extensive framework in more detail. Then we explain several variants that take into account the correlation between noise and speech. This chapter highlights the difference between the actual relation between speech and noise, that is including their correlation, and the usual relation used on deriving robust speech recognition methods which ignores the correlation between speech and noise.

In Chapter 4, we briefly explain the theory of non-extensive statistics. This chapter includes the non-extensive functions and their properties, Tsallis entropy, and the q -Gaussian distribution. This chapter emphasizes the generalization of the Boltzmann-Gibbs statistics using non-extensive statistics.

In Chapter 5, we explain and derive our first method, q -LSMN. We show that the correlation between speech and noise can be captured in the new domain, the q -log spectral domain. Then, we present our evaluation on speech recognition task to confirm the effectiveness of our method.

In Chapter 6, we derive our second method, q -SS. We show that a nonlinear spectral subtraction, which is a variant of spectral subtraction, can naturally be obtained using the non-extensive framework. We also present our experiments on speech recognition to evaluate our method and confirm its effectiveness.

In Chapter 7, we summarize the major outcomes of this research. We also discuss several possible directions for future research.

Chapter 2

Statistical speech recognition: an overview

2.1 Introduction

The objective of an automatic speech recognition (ASR) system is to convert a speech signal into a sequence of phonemes, syllabus, or words. There are several approaches to perform speech recognition in the literature. The popular approaches are dynamic time warping (DTW) [Myers and Rabiner, 1981; Pawate et al., 1987; Silverman and Morgan, 1990], Hidden Markov Model (HMM) [Rabiner, 1989], and artificial neural network (ANN) [Iso and Watanabe, 1990; Waibel et al., 1989].

DTW is a feature-matching method that basically tries to match “time-alignment” of the test features with the sets of reference using a dynamic programming procedure. Due to large variations of the speech signals (e.g. speaker effect, environment, dialect, etc), this approach requires a huge storage and a very high computational cost of searching for larger vocabularies task. Therefore, it is only suitable for small vocabularies systems.

HMM is a statistical pattern recognition approach. There are usually two phases in the statistical pattern recognition approach: the training phase and the testing phase. In the training phase, the system is trained using a large amount of representative data to find the model for each of the speech units (phonemes or words). In the testing phase, the speech signal that wants to be recognized is matched with all speech units’ models and the model that produces the best match is then chosen as the recognized

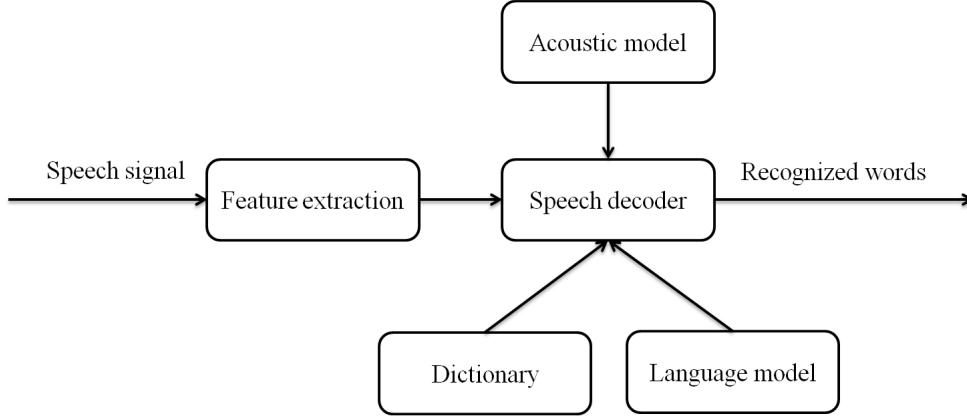


Figure 2.1: The block diagram of an HMM-based speech recognition system

speech.

In ANN, multi-layer perceptrons (MLP) are used to model the relation between the input speech signals and the desired output (phoneme or word) of the network. However, the dynamics of the speech signals are more difficult to be modeled by ANN than by HMM. This causes the recognition accuracies of the ANN-based systems are usually worse than those of the HMM-based systems. Recently, a new generation of ANN, deep belief network (DBN) [Dahl et al., 2012; Mohamed et al., 2012], has gained an increasing interest among researchers.

Currently, the HMM is the state-of-the-art method for ASR systems. It is also implemented in this thesis. Fig. 2.1 shows the basic structure of an HMM-based speech recognition system. It consists of five modules: feature extraction, acoustic model, language model, dictionary (vocabulary) and speech decoder (searching algorithm). In the feature extraction, the speech signal is transformed into some discriminative representation of speech signals. This representation is called the speech feature. In this thesis, we use Mel frequency cepstral coefficient (MFCC) [Davis and Mermelstein, 1980] as the speech feature. The acoustic model is the HMM representation for each sub-unit of speech, which is trained using the speech data. The language model contains all allowed combinations of words and the probability for each combination. In speech decoder, a recognition hypothesis given a sequence of features is produced.

In the remaining of this chapter, we briefly review each part of speech recognition system except the language model. In this thesis, we implement a digit recognition

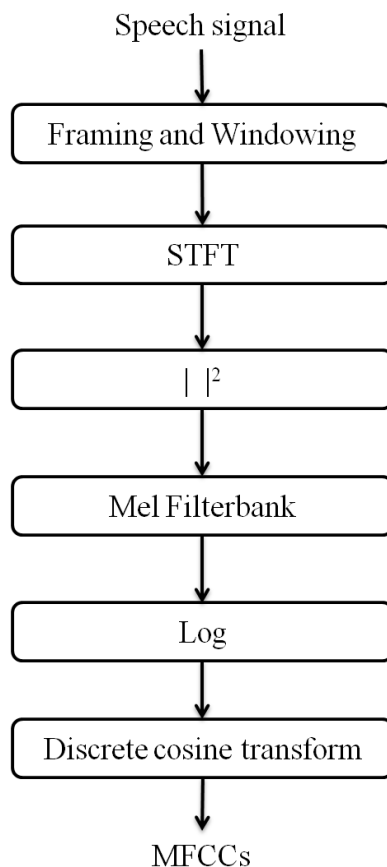


Figure 2.2: MFCC feature extraction

system and language model is not needed for this task. In Section 2.2, we explain the extraction process to obtain the MFCCs. In Section 2.3, we explain the HMM-based speech recognition system in more details. This section covers the acoustic models training, the decoding process and evaluation process.

2.2 MFCC feature extraction

Figure 2.2 shows the block diagram of the MFCC feature extraction process. This feature is widely used as a feature for speech recognition. Perceptual linear prediction (PLP) [Hermansky, 1990] and linear predictive coding (LPC) [Makhoul, 1975] are other features that are commonly used for speech recognition.

Human speech signals are quasi-stationary signals. This means the properties of

the speech signals change slowly over time. For that reason, the speech signals are processed in blocks over which they are assumed to be stationary. This block of speech is called a frame. To reduce artifacts at the boundary of each frame, the neighboring frames are over-lapped by $1/2$ to $2/3$ of the length of the frame. The frame shift is the frame length minus the frame overlap. The length of a frame is usually between 20-30 ms and the frame shift is 10 ms for speech recognition tasks.

After being partitioned into frames, each frame is multiplied by a window function. The windowing process is needed because the framing process introduces a discontinuity which causes a spectral leakage. By windowing, the amplitude of the samples at the beginning and at the end of the window is reduced, hence the spectral leakage is spread among the frequency components. In this thesis, we implement Hamming window, which is defined by:

$$w(t) = \begin{cases} 0.54 - 0.46 \cos\left(\frac{2\pi t}{T-1}\right), & 0 \leq t \leq T-1 \\ 0, & \text{otherwise} \end{cases} \quad (2.1)$$

where T is the length of the frame.

Each frame is then transformed into the spectral domain by applying discrete Fourier transform (DFT). If $s = [s(0)s(1) \cdots s(t) \cdots s(T-1)]$ represents the m -th frame of speech, then k -th DFT coefficients $S(m, k)$ can be computed as follows:

$$S(m, k) = \sum_{t=0}^{T-1} s(t) \exp\left(j \frac{2\pi}{T} kt\right), \quad 0 \leq k \leq T-1 \quad (2.2)$$

DFT coefficients are complex numbers containing both magnitude and phase information:

$$S(m, k) = |S(m, k)| \exp(j\theta_S(m, k)), \quad (2.3)$$

where $|S(m, k)|$ is the magnitude spectrum and $\theta_S(m, k)$ is the phase spectrum. The power spectral representation of speech can be obtained by taking the square of the magnitude spectrum.

Implementing some knowledge about human perception system is found to be useful in speech recognition. In MFCC, the power spectra of speech are warped into Mel-scaled frequency resolution. It is a nonlinear scaling which approximates the human auditory

systems. It is formulated as [Molau et al., 2001]:

$$f_{\text{mel}}(f) = 2595 \log_{10} \left(1 + \frac{f}{700} \right), \quad (2.4)$$

where f is the index of the linear frequency scale (Hz). In the implementation, we use triangular filter-banks with an overlap of 50% between the neighboring filters as shown in Fig. 2.3. To implement the filter-bank, we first calculate the center frequency in the Mel domain:

$$f_{\text{center}}(l) = (l + 1)\Delta_{\text{mel}}, \quad 1 \leq l \leq L \quad (2.5)$$

where l is the index of the filterbank, L is the total number of the filter (in this thesis, $L = 23$) and:

$$\Delta_{\text{mel}} = \frac{f_{\text{max}}}{L + 1}, \quad (2.6)$$

where:

$$f_{\text{max}} = f_{\text{mel}}(f_s/2), \quad (2.7)$$

where f_s is the sampling frequency. After calculating the center frequencies in the Mel domain, we map them into indices of the DFT. This can be done by dividing $f_{\text{center}}(l)$ by $\Delta_f = \frac{f_s}{K}$ and then rounding them to the next integer:

$$f_{\text{center}}^{(l)}(k) = \text{round} \left(\frac{f_{\text{center}}(l)}{\Delta_f} \right) \quad (2.8)$$

After finding the center frequencies in the linear domain that correspond to the center frequencies in the Mel domain, we calculate the transfer function of each filter-bank as follows:

$$W_l(k) = \begin{cases} \frac{2(k - f_{\text{center}}^{(l-1)})}{(f_{\text{center}}^{(l+1)} - f_{\text{center}}^{(l-1)})(f_{\text{center}}^{(l)} - f_{\text{center}}^{(l-1)})}, & \text{if } f_{\text{center}}^{(l-1)} < n \leq f_{\text{center}}^{(l)} \\ \frac{2(f_{\text{center}}^{(l+1)} - k)}{(f_{\text{center}}^{(l+1)} - f_{\text{center}}^{(l-1)})(f_{\text{center}}^{(l+1)} - f_{\text{center}}^{(l)})}, & \text{if } f_{\text{center}}^{(l)} < n \leq f_{\text{center}}^{(l+1)} \\ 0, & \text{otherwise} \end{cases} \quad (2.9)$$

where $W_l(k)$ is the filter weight at frequency band k in the l -th bank. The Mel-scaled

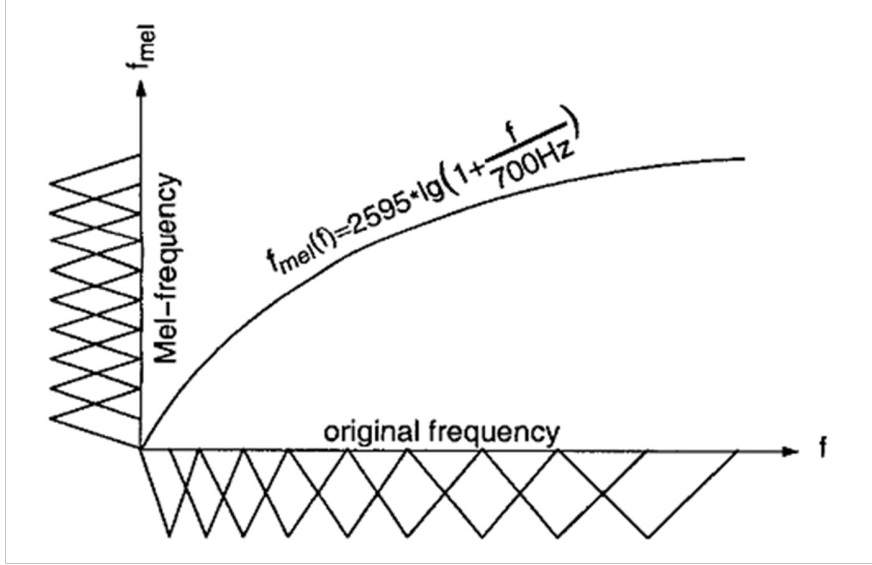


Figure 2.3: Mel-scale and the implementation of triangular filterbanks [Molau et al., 2001]

power spectra then are obtained using the following formula:

$$|\check{S}(l)|^2 = \sum_k W_l(k) |S(k)|^2. \quad 1 \leq l \leq L \quad (2.10)$$

After obtaining the Mel-power spectra, we take the natural logarithm (log) of them. This operation serves three purposes. First, the log function deals with the loudness nonlinearity since a humans perception of loudness is proportional to the log of the sound intensity. Second, the log function is a homomorphic operator [Oppenheim and Schafer, 1968]. Speech can be modeled as a time-dependent vocal tract filter convolve with a glottal excitation function. By taking the log of the speech spectra, the relation between the components of speech becomes linear. Third, the log function is used to reduce the dynamic range of the speech coefficients. Root function can also be used for this purpose [Alexandre et al., 1993; Lim, 1979] instead of the log function. The use of root function is found to be more robust against noise than the log function. This is because the root function is less sensitive than the log function to the change in the low regions of the power spectra, which are often affected by noise.

The MFCCs then can be obtained by taking the Discrete Consine Transform (DCT) of the log mel-power spectra. The p -th coefficient of the MFCCs is calculated using the

following formula:

$$s(p) = \sum_{l=0}^{L-1} \log |\check{S}(l)|^2 \cos \left[\frac{\pi p(l - 0.5)}{P} \right], \quad 0 \leq p < P \quad (2.11)$$

where P is the total number of the cepstral coefficients. Usually, P varies from 24 to 40. In speech recognition task, only 12 or 13 of the first coefficients are used.

The features obtained from Eq. (2.11) are called static features because they only contain information in the local frame. Since the speech signals are time-dependent series, it is important to include the temporal information, which is not contained in static features. To obtain the temporal information, the time derivatives (the first delta) and accelerations (the second delta) are used [Picone, 1993]. The first delta coefficient is calculated as follow:

$$\Delta_p^m = \frac{\sum_{j=1}^J j(c_p^{m+j} - c_p^{m-j})}{2 \sum_{j=1}^J j^2}, \quad (2.12)$$

where Δ_p^m is delta coefficient for k -th coefficient for the m -th frame. The typical value for parameter J is 2 or 3. This value is used to describe the number of frames considered in the evaluation, which is $2J + 1$. The second delta is calculated in a similar manner as the first delta.

2.3 HMM-based statistical speech recognition

In statistical based approach, an ASR outputs the hypothesized word sequence, $\hat{\mathbf{W}}$ which has the highest probability given an observation speech sequence, which is generated from the features. It is mathematically written:

$$\hat{\mathbf{W}} = \underset{\mathbf{W}}{\operatorname{argmax}} P(\mathbf{W} | \mathbf{O}) = \underset{\mathbf{W}}{\operatorname{argmax}} \frac{P(\mathbf{W})P(\mathbf{O} | \mathbf{W})}{P(\mathbf{O})}, \quad (2.13)$$

where $P(\mathbf{W})$ is the prior probability of word sequence, $P(\mathbf{O} | \mathbf{W})$ is the probability of the observed speech sequence given the word sequence, $P(\mathbf{O})$ is the prior probability of the observed speech sequence, $\mathbf{W} = w_1, w_2, \dots, w_n$ denotes the word sequence, and $\mathbf{O}$ denotes the observation speech sequence:

$$\mathbf{O} = \mathbf{o}_1, \mathbf{o}_2, \dots, \mathbf{o}_t, \dots, \mathbf{o}_T, \quad (2.14)$$

where $\mathbf{o}_t$ is the observation at time t . Since $P(\mathbf{O})$ are not related to $\mathbf{W}$, Eq. (2.13) can be simplified as follows:

$$\hat{\mathbf{W}} = \underset{\mathbf{W}}{\operatorname{argmax}} P(\mathbf{W})P(\mathbf{O} | \mathbf{W}). \quad (2.15)$$

Therefore, the problem in ASR is to find suitable probabilistic models for speech and language. $P(\mathbf{O} | \mathbf{W})$ is given by the acoustic model, that is the HMM, and $P(\mathbf{W})$ is given by the language model.

In the HMM-based speech recognition, we assume that the observed sequence $\mathbf{O}$, which is generated from the speech features, is produced by a Markov model. Under this assumption, the speech signals can be transformed into piecewise stationary states. In speech recognition, only the acoustic features of the speech signal of interest are known, but which state sequence produces the observed sequence of features $\mathbf{O}$ is unknown, i.e. hidden. Therefore hidden Markov model is used.

In the acoustic model, each sub-unit of speech, i.e. word or phoneme, is modeled by an HMM to represent the unique characteristics of each sub-unit of speech. Fig. 2.4 illustrates 3-states left-to-right HMM. Initial and final states are non-emitting states, which are required to chain together multiple HMMs for continuous sub-word recognition. The rest of the states provide all the statistical information required for training the acoustic models.

An HMM is defined by the following parameters:

- A set of states $S = \{S_1, S_2, \dots, S_N\}$. The state at time t is denoted as s_t .
- A transition probability matrix $\mathbf{A} = \{a_{ij}\}$, where $\sum_{j=1}^N a_{ij} = 1$.
- A set of observation probability densities $\mathbf{B} = \{b_j(\mathbf{o}_t)\}$, which determines the probability of the observation $\mathbf{o}_t$ is generated from state j . In ASR implementations, this probability density is represented by a Gaussian mixture model (GMM):

$$b_j(\mathbf{o}_t) = \sum_{r=1}^R \gamma_{jr} \mathcal{N}(\mathbf{o}_t; \boldsymbol{\mu}_{jm}, \boldsymbol{\Sigma}_{jm}), \quad (2.16)$$

where r is the index of the mixture components, R is the total number of mixture components, γ_{jr} is the weight of the r^{th} mixture component. A GMM is comprised

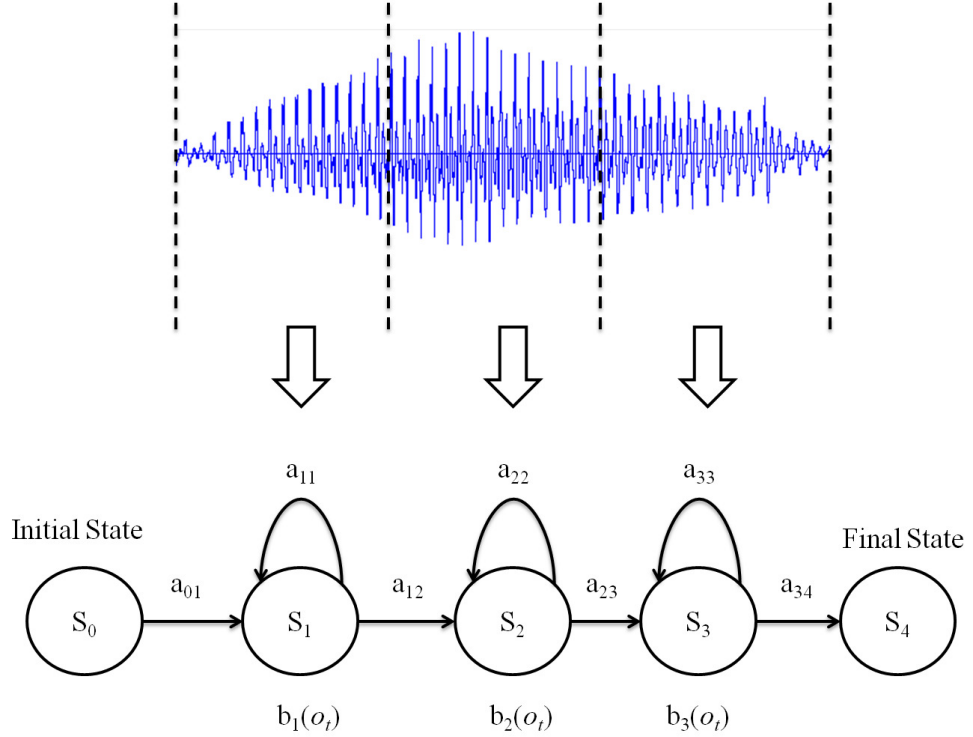


Figure 2.4: An HMM with three true states. The initial and final states are used to tie with other HMM and do not model speech.

of multivariate Gaussian probability density functions, which is formulated as the following:

$$\mathcal{N}(\mathbf{o}; \boldsymbol{\mu}, \boldsymbol{\Sigma}) = \frac{1}{\sqrt{(2\pi)^n |\boldsymbol{\Sigma}|}} \exp(-0.5(\mathbf{o} - \boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1}(\mathbf{o} - \boldsymbol{\mu})) \quad (2.17)$$

where $\boldsymbol{\mu}$ is the mean vector, $\boldsymbol{\Sigma}$ is the covariance matrix, and n is the dimensionality of feature vector.

- An initial state distribution $\pi = \{\pi_i\}$.

In short, an HMM, Λ , is described by a set of parameters $\Lambda = (\pi, \mathbf{A}, \mathbf{B})$.

In order to use HMMs for speech recognition, there are three problems need to be solved [Rabiner, 1989]:

- **The evaluation problem.** Given the observation sequence $\mathbf{O}$, and the model Λ , we want to calculate $P(\mathbf{O} | \Lambda)$, the probability of the observation given the

model Λ .

- **The decoding problem.** Given the observation sequence $\mathbf{O}$ and the model Λ , we want to find the most likely sequence $\mathbf{S} = [s_1, s_2, \dots, s_T]$,
- **The parameter estimation problem.** We want to adjust the parameters of Λ to maximize $P(\mathbf{O} \mid \Lambda)$.

The evaluation problem

The probability of an observation sequence $\mathbf{O}$ given the model Λ , $P(\mathbf{O} \mid \Lambda)$, can be obtained by summing the probabilities over all possible state sequences:

$$\begin{aligned}
 P(\mathbf{O} \mid \Lambda) &= \sum_{\text{all } \mathbf{S}} P(\mathbf{O}, \mathbf{S} \mid \Lambda) \\
 &= \sum_{\text{all } \mathbf{S}} P(\mathbf{S} \mid \Lambda) \cdot P(\mathbf{O}, \mathbf{S} \mid \Lambda) \\
 &= \sum_{\text{all } \mathbf{S}} \pi_{s_1} b_{s_1}(\mathbf{o}_1) a_{s_1, s_2} b_{s_2}(\mathbf{o}_2) \cdots a_{s_{T-1}, s_T} b_{s_T}(\mathbf{o}_T)
 \end{aligned} \tag{2.18}$$

The Baum-Welch (forward-backward) algorithm Baum and Eagon [1967] is usually used to efficiently compute Eq. (2.18).

The Baum-Welch algorithm works as follows. Let $\alpha_t(i) = P[\mathbf{o}_1 \mathbf{o}_2 \cdots \mathbf{o}_t, s_t = i \mid \Lambda]$ be the forward probability, which is the joint probability that the HMM at state i and time t having a partial observation sequence (the first t observations) and $\beta_t(i) = P[\mathbf{o}_{t+1} \mathbf{o}_{t+2} \cdots \mathbf{o}_T \mid s_t = i, \Lambda]$ be the backward probability, which is the conditional probability of observing the partial observation sequence (from time t to T), given that at time t and the HMM is in the state i . The $\alpha_t(i)$ can be solved inductively as follows:

1. Initialization:

$$\alpha_1(i) = \pi b_i(\mathbf{o}_1), \quad 1 \leq i \leq N \tag{2.19}$$

2. Induction

$$\alpha_{t+1}(i) = \left(\sum_{j=1}^N \alpha_{t-1}(j) a_{j,i} \right) b_i(\mathbf{o}_{t+1}), \quad 1 \leq t \leq T-1, \quad 1 \leq i \leq N \tag{2.20}$$

3. Final probability is given by:

$$P(\mathbf{O} \mid \Lambda) = \sum_{i=1}^N \alpha_T(i). \quad (2.21)$$

The $\beta_t(i)$ is calculated in a similar manner:

1. Initialization:

$$\beta_T(i) = 1, \quad 1 \leq i \leq N \quad (2.22)$$

2. Induction:

$$\beta_t(i) = \sum_{j=1}^N \beta_{t+1}(j) b_j(\mathbf{o}_{t+1}) a_{ij}, \quad t = T-1, T-2, \dots, 1, \quad 1 \leq i \leq N \quad (2.23)$$

3. Final probability is given by:

$$P(\mathbf{O} \mid \Lambda) = \sum_{i=1}^N \beta_1(i). \quad (2.24)$$

The decoding problem

We wish to find the “optimal” state sequence given the observation sequence. Viterbi algorithm [Forney, 1973] is usually used to solve this problem. This algorithm is similar with forward-backward procedure. Let:

$$\delta_t(i) = \max_{s_1, s_2, \dots, s_{t-1}} P(s_1 s_2 \dots, s_t = i, \mathbf{o}_1 \mathbf{o}_2 \dots \mathbf{o}_t \mid \Lambda), \quad (2.25)$$

be the best score along a single path, at time t , which accounts for the first t observations and ends at state S_i . By induction, we obtain:

$$\delta_{t+1}(j) = \left[\operatorname{argmax}_i \delta_t(i) a_{ij} \right] \cdot b_j(\mathbf{o}_{t+1}). \quad (2.26)$$

We define the array $\psi(j)$ to keep track the argument that maximizes Eq. (2.26). The complete Viterbi algorithm is describe as follows:

- Initialization:

$$\delta_1(i) = \pi_i b_i(\mathbf{o}_1), \quad 1 \leq i \leq N \quad (2.27)$$

$$\psi_1(i) = 0. \quad (2.28)$$

- Recursion:

$$\delta_t(i) = \max_{1 \leq j \leq N} [\delta_{t-1}(j) a_{ji}] b_i(\mathbf{o}_t), \quad 2 \leq t \leq T, \quad 1 \leq i \leq N \quad (2.29)$$

$$\psi_t(i) = \arg \max_{1 \leq j \leq N} [\delta_{t-1}(j) a_{ji}], \quad 2 \leq t \leq T, \quad 1 \leq i \leq N. \quad (2.30)$$

- Termination:

$$P^* = \max_{1 \leq i \leq N} [\delta_T(i)], \quad (2.31)$$

$$s_T^* = \arg \max_{1 \leq i \leq N} [\delta_T(i)]. \quad (2.32)$$

- Path backtracking:

$$s_t^* = \psi_{t+1}(s_{t+1}^*), \quad t = T-1, T-2, \dots, 1. \quad (2.33)$$

The parameter estimation problem

We wish to estimate the model parameters $\Lambda = (\pi, \mathbf{A}, \mathbf{B})$ that maximize $P(\mathbf{O} \mid \Lambda)$. The Baum-Welch algorithm based on Expectation-Maximization (EM) algorithm [Dempster et al., 1977] is usually used to solve this problem.

The EM algorithm is an algorithm to find the maximum likelihood estimation of the marginal likelihood by iteratively applying the following steps:

- Step 1: Initialization. Choose an initial model Λ .
- Step 2: Expectation step (E-step). In E-step, the expectation (likelihood) of cost function called auxiliary function is calculated. The auxiliary function $Q(\Lambda, \hat{\Lambda})$ is

defines as:

$$Q(\Lambda, \hat{\Lambda}) = \sum_{\text{all } \mathbf{S}} P(\mathbf{O}, \mathbf{S} \mid \Lambda) \log P(\mathbf{O}, \mathbf{S} \mid \hat{\Lambda}) \quad (2.34)$$

where $P(\mathbf{O}, \mathbf{S} \mid \Lambda)$ and $\log P(\mathbf{O}, \mathbf{S} \mid \hat{\Lambda})$ can be expresses as HMM parameters:

$$P(\mathbf{O}, \mathbf{S} \mid \Lambda) = \pi_{s_0} \prod_{t=1}^T a_{s_{t-1}s_t} b_{q_t}(\mathbf{o}_t) \quad (2.35)$$

$$\log P(\mathbf{O}, \mathbf{S} \mid \Lambda) = \log \pi_{s_0} + \sum_{t=1}^T \log a_{s_{t-1}s_t} + \sum_{t=1}^T \log b_{q_t}(\mathbf{o}_t) \quad (2.36)$$

Thus, Eq. (2.34) can be written:

$$Q(\Lambda, \hat{\Lambda}) = Q_{\pi}(\Lambda, \hat{\pi}) + \sum_{i=1}^N Q_{a_i}(\Lambda, \hat{\mathbf{a}}_i) + \sum_{i=1}^N Q_{b_i}(\Lambda, \hat{\mathbf{b}}_i), \quad (2.37)$$

where $\hat{\pi} = [\pi_1, \pi_2, \dots, \pi_N]$, $\hat{\mathbf{a}}_i = [a_{i1}, a_{i2}, \dots, a_{iN}]$ and $\hat{\mathbf{b}}_i$ is the vector that defines $b_i(\cdot)$ and:

$$Q_{\pi}(\Lambda, \hat{\pi}) = \sum_{i=1}^N P(\mathbf{O}, s_0 = i \mid \Lambda) \log \pi_i \quad (2.38)$$

$$Q_{a_i}(\Lambda, \hat{\mathbf{a}}_i) = \sum_{j=1}^N \sum_{t=1}^T P(\mathbf{O}, s_{t-1} = i, s_t = j \mid \Lambda) \log a_{ij} \quad (2.39)$$

$$Q_{b_i}(\Lambda, \hat{\mathbf{b}}_i) = \sum_{t=1}^T P(\mathbf{O}, s_t = i \mid \Lambda) \log b_i(\mathbf{o}_t) \quad (2.40)$$

- Step 3: Maximization step (M-step). Compute $\hat{\Lambda}$ to maximize the auxiliary function $Q(\Lambda, \hat{\Lambda})$. Since $Q(\Lambda, \hat{\Lambda})$ can be separated into three independent terms, we can maximize it over $\hat{\Lambda}$ by maximizing the individual term separately, subject to constraint:

$$\sum_{j=1}^N \pi_j = 1, \quad (2.41)$$

$$\sum_{j=1}^N a_{ij} = 1, \quad \forall j \quad (2.42)$$

and:

$$\sum_{r=1}^R b_i(r) = 1, \quad \forall i \quad (2.43)$$

where $b_i(\mathbf{o}_t = \mathbf{v}_r) = b_i(r)$. Then, the parameters of the re-estimate model $\hat{\Lambda} = (\hat{\pi}, \hat{\mathbf{A}}, \hat{\mathbf{B}})$ are calculated as:

$$\hat{\pi} = \frac{P(\mathbf{O}, s_t = i \mid \Lambda)}{P(\mathbf{O} \mid \Lambda)}, \quad (2.44)$$

$$\hat{a}_{ij} = \frac{\sum_{t=1}^T P(\mathbf{O}, s_{t-1} = i, s_t = j \mid \Lambda)}{\sum_{t=1}^T P(\mathbf{O}, s_{t-1} = i \mid \Lambda)}, \quad (2.45)$$

$$\hat{b}_i(r) = \frac{\sum_{t=1}^T P(\mathbf{O}, s_t = i \mid \Lambda) \delta(\mathbf{o}_t, \mathbf{v}_r)}{\sum_{t=1}^T P(\mathbf{O}, s_t = i \mid \Lambda)}, \quad (2.46)$$

where:

$$\delta(\mathbf{o}_t, \mathbf{v}_r) = \begin{cases} 1, & \text{if } \mathbf{o}_t = \mathbf{v}_r \\ 0 & \text{otherwise.} \end{cases} \quad (2.47)$$

- Step 4: Iteration. Set $\Lambda = \hat{\Lambda}$, and repeat from step 2 until converges

Chapter 3

Noise robust speech recognition

Various methods have been proposed in the past literature for improving the robustness of speech recognition in noisy environments. Generally, they can be categorized into two groups: feature enhancement and model compensation. In feature enhancement, the aim is to estimate the clean speech features in noisy speech by removing noise. Whereas, model compensation methods adapt the clean speech models to noisy conditions, by considering the noise statistics. They can usually achieve better performance than feature enhancement-based methods. On the other hand, they require higher computational cost and more data than feature enhancement-based methods. In addition, they should update the models each time a new type of noise is introduced. Despite their different approaches, most robust speech recognition methods assume that noise and speech are uncorrelated. In this chapter, we review several robust ASR methods which are based on this assumption.

This chapter is organized as follow. In Section 3.1, we review the effect of noise on the speech features. In Section 3.2, we explain several popular methods for robust speech recognition. We first review the consequences of assuming that noise and speech are uncorrelated. We next describe some previous studies of robust speech recognition methods which are derived under this assumption. We then explain several variants that take into account the correlation between noise and speech. We summarize this chapter in Section 3.3

3.1 The effect of noise on the speech features

In Chapter 1, we show the effect of noise on the power spectra of speech. In this section, we show the effect of noise on MFCCs. In the Mel spectral domain, Eq. (1.7) can be written:

$$|\check{Y}(m, l)|^2 = |\check{S}(m, l)|^2 |\check{H}(m, l)|^2 + |\check{N}(m, l)|^2 + 2\alpha(m, l) |\check{S}(m, l)| |\check{H}(m, l)| |\check{N}(m, l)|, \quad (3.1)$$

where:

$$|\check{Y}(m, l)|^2 = \sum_k W_l(k) |Y(m, k)|^2, \quad (3.2)$$

$$|\check{S}(m, l)|^2 = \sum_k W_l(k) |S(m, k)|^2, \quad (3.3)$$

$$|\check{N}(m, l)|^2 = \sum_k W_l(k) |N(m, k)|^2, \quad (3.4)$$

$$|\check{H}(m, l)|^2 = \sum_k W_l(k) |H(m, k)|^2, \quad (3.5)$$

and

$$\alpha(m, l) = \frac{\sum_k W_k(m, l) |S(m, k)| |H(m, k)| |N(m, k)| \cos \theta(m, k)}{|\check{S}(m, l)| |\check{H}(m, l)| |\check{N}(m, l)|}. \quad (3.6)$$

To improve the readability, we dropped the subscripts m , k , and l from now on. By taking the log and applying DCT matrix $\mathbf{C}$ to Eq. (3.1), we obtain the MFCCs of noisy speech as follows:

$$\mathbf{y} = \mathbf{s} + \mathbf{h} + \mathbf{C} \log \left\{ 1 + \exp [\mathbf{C}^{-1}(\mathbf{n} - \mathbf{s} - \mathbf{h})] + 2\alpha \exp \left[\frac{\mathbf{C}^{-1}(\mathbf{n} - \mathbf{s} - \mathbf{h})}{2} \right] \right\}, \quad (3.7)$$

where $\mathbf{y}$, $\mathbf{s}$, $\mathbf{h}$ and $\mathbf{n}$ are the MFCCs of the noisy speech, clean speech, convolutive noise and additive noise respectively, and α is the phase factor:

$$\alpha = \frac{\exp[\mathbf{y} - \mathbf{s} - \mathbf{h}] - \exp[\mathbf{n} - \mathbf{s} - \mathbf{h}] - 1}{2 \exp \left[\frac{\mathbf{n} - \mathbf{s} - \mathbf{h}}{2} \right]} \quad (3.8)$$

Let:

$$f(\mathbf{s}, \mathbf{h}, \mathbf{n}, \boldsymbol{\alpha}) = \mathbf{h} + \mathbf{C} \log \left\{ 1 + \exp [C^{-1}(\mathbf{n} - \mathbf{s} - \mathbf{h})] + 2\boldsymbol{\alpha} \exp \left[\frac{\mathbf{C}^{-1}(\mathbf{n} - \mathbf{s} - \mathbf{h})}{2} \right] \right\}, \quad (3.9)$$

be the mismatch function. We can write Eq. (3.7) as follows:

$$\mathbf{y} = \mathbf{s} + f(\mathbf{s}, \mathbf{h}, \mathbf{n}, \boldsymbol{\alpha}). \quad (3.10)$$

Based on Eq.(3.10), we can see that the speech features are altered by $f(\mathbf{s}, \mathbf{h}, \mathbf{n}, \boldsymbol{\alpha})$. It is a nonlinear function which has four unknown variables including the speech signal itself and we only know $\mathbf{y}$. For that reason, robust speech recognition is difficult to solve.

3.2 Noise robust speech recognition methods

Most robust speech recognition methods assume that speech and noise are uncorrelated. This assumption is made since the expected value of the cross-term is zero. Under this assumption, the cross-term can be ignored and Eq. (1.7) can be simplified as:

$$|Y|^2 = |S|^2|H|^2 + |N|^2. \quad (3.11)$$

In the Mel frequency domain, Eq. (3.11) can be written as follows:

$$|\check{Y}|^2 = |\check{S}|^2|\check{H}|^2 + |\check{N}|^2. \quad (3.12)$$

Similarly, by taking the log and applying DCT matrix $\mathbf{C}$ to Eq. (3.12), the MFCC representation of Eq. (3.12) is:

$$\mathbf{y} = \mathbf{s} + \mathbf{h} + \mathbf{C} \log \left\{ 1 + \exp [C^{-1}(\mathbf{n} - \mathbf{s} - \mathbf{h})] \right\}. \quad (3.13)$$

Therefore, by assuming speech and noise are uncorrelated, the mismatch function becomes

$$f(\mathbf{s}, \mathbf{h}, \mathbf{n}) = \mathbf{h} + \mathbf{C} \log \left\{ 1 + \exp [\mathbf{C}^{-1}(\mathbf{n} - \mathbf{s} - \mathbf{h})] \right\}. \quad (3.14)$$

Eq. (3.14) can be simplified as follows:

$$\mathbf{y} = \mathbf{s} + f(\mathbf{s}, \mathbf{h}, \mathbf{n}). \quad (3.15)$$

The mismatch function in Eq. (3.14) is simpler than that of Eq. (3.9). By ignoring the cross-term, the phase difference between noise and speech, which is difficult to estimate [Alsteris and Paliwal, 2007], is not ignored.

3.2.1 Model compensation approach

The simplest way to adapt the clean speech models to the noisy environments is by re-training the speech recognizer using noisy speech. However, it requires a large amount of training data. Besides, it is also difficult to obtain noisy speech for all possible noisy environments. Parallel model combination (PMC) [Gales and Young, 1992] and VTS are two popular methods for model compensation methods.

3.2.1.1 Parallel model combination (PMC)

Figure 3.1 illustrates the PMC process. In PMC, a noisy speech HMM is created by combining a clean speech HMM and a noise HMM. The speech and noise model parameters are first transformed into the log spectral [Gales and Young, 1996] or linear spectral domains [Gales and Young, 1993]. The parameters of speech and noise are then combined to obtain the parameters for noisy speech. After the combining process, the estimate noisy speech parameters are then transformed back into the cepstral domain to build their HMMs. In PMC, clean speech is modeled using a mixture of Gaussians and noise is modeled with single Gaussian. The resulting distribution of noisy speech is also assumed to be a mixture of Gaussians.

The mean $\mu_{\mathbf{s}}^l$ and the covariance $\Sigma_{\mathbf{s}}^l$ of clean speech in the log spectral domain respectively are obtained using the following formulas:

$$\mu_{\mathbf{s}}^l = \mathbf{C}^{-1} \mu_{\mathbf{s}}^c, \quad (3.16)$$

$$\Sigma_{\mathbf{s}}^l = \mathbf{C}^{-1} \Sigma_{\mathbf{s}}^c (\mathbf{C}^{-1})^T, \quad (3.17)$$

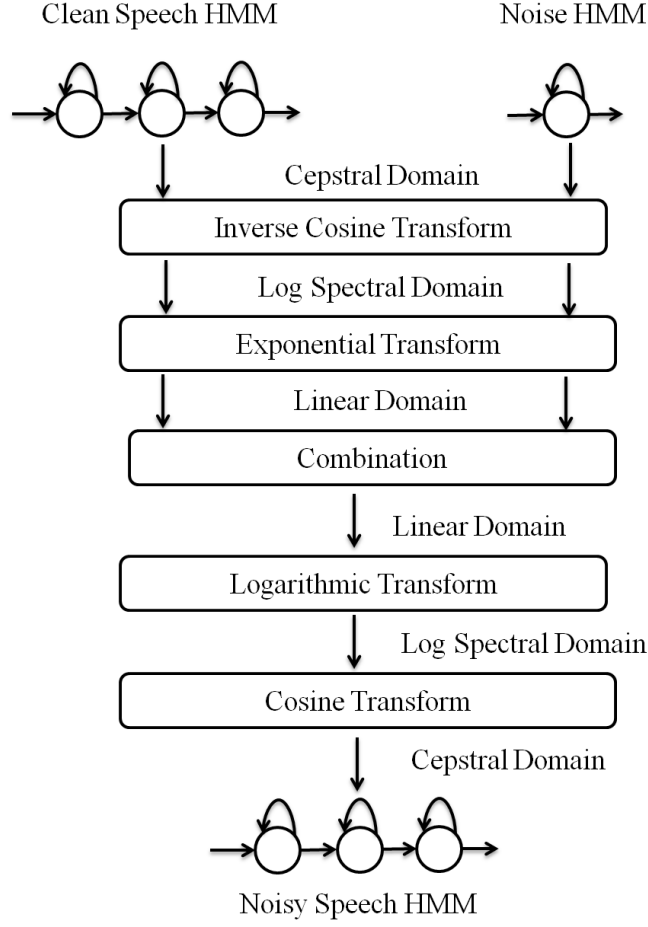


Figure 3.1: Parallel Model Combination

where μ_s^c and Σ_s^c are the mean and the covariance of a Gaussian probability density function (pdf) of clean speech in the cepstral domain respectively. Since clean speech is assumed to follow Gaussian distribution in the cepstral domain, its distribution in the linear spectral domain follows a lognormal distribution. The mean μ^s and the covariance Σ^s in linear spectral domain are formulated as follows:

$$\mu_s^d[i] = \exp \left(\mu_s^l[i] + 0.5 \Sigma_s^l[i, i] \right), \quad (3.18)$$

$$\Sigma_s^d[i, j] = \mu_s^d[i] \mu_s^d[j] \left[\exp \left(\Sigma_s^l[i, j] \right) - 1 \right], \quad (3.19)$$

where i and j are the indices of the components of the distribution in the linear domain. For noise, $\mu_{\mathbf{n}}^l$ and $\Sigma_{\mathbf{n}}^l$ can be calculated in the same way as Eq. (3.16) and (3.17) respectively.

The parameters of noisy speech are obtained by adding the parameters of clean speech with those of noise in either the linear spectral domain or the log spectral domain. Since noise and speech are assumed to be uncorrelated, the parameters of clean speech and noise are combined in the linear spectral domain using the following formulas:

$$\mu_{\mathbf{y}}^d = g\mu_{\mathbf{s}}^d + \mu_{\mathbf{n}}^d, \quad (3.20)$$

$$\mu_{\mathbf{s}}^d = g^2\Sigma_{\mathbf{s}}^d + \Sigma_{\mathbf{n}}^d, \quad (3.21)$$

where g is the gain matching term, which is estimated using the following formula:

$$g = \frac{E_y - E_n}{E_s}, \quad (3.22)$$

where E_y , E_n and E_s are the average energy of the noisy speech, noise, and clean speech respectively.

3.2.1.2 Vector Taylor series (VTS) for model compensation

In this method, vector Taylor expansion is used to approximate the mismatch function. VTS can be implemented either for feature enhancement [Moreno et al., 1996] or model compensation [Acero et al., 2000; Kim et al., 1998]. In this section, we explain VTS-based method for model compensation. We explain VTS-based feature enhancement method in Section 3.2.2.4.

For model compensation, the purpose is the same as for PMC, i.e, to adapt the clean speech model into noisy environments. In this method, both clean and noisy speech are also modeled using GMMs with one-to-one corresponds to each other. It is also assumed that noise and speech are uncorrelated. It is called phase insensitive VTS since the phase information is excluded since the cross-term is ignored.

From Eq. (3.14), we can re-write the mismatch function in cepstral domain as:

$$f(\mathbf{s}, \mathbf{h}, \mathbf{n}) = \mathbf{h} + g(\mathbf{s}, \mathbf{h}, \mathbf{n}), \quad (3.23)$$

where:

$$g(\mathbf{s}, \mathbf{h}, \mathbf{n}) = \mathbf{C} \log [1 + \exp(\mathbf{C}^{-1}(\mathbf{n} - \mathbf{h} - \mathbf{s}))]. \quad (3.24)$$

Eq. (3.10) can be re-written, by replacing $g(\mathbf{s}, \mathbf{h}, \mathbf{n})$ with its first-order vector Taylor series approximation, as follows:

$$\mathbf{y} = \mu_{\mathbf{s}} + \mu_{\mathbf{h}} + g(\mu_{\mathbf{s}}, \mu_{\mathbf{h}}, \mu_{\mathbf{n}}) + \frac{\partial \mathbf{y}}{\partial \mathbf{s}}(\mathbf{s} - \mu_{\mathbf{s}}) + \frac{\partial \mathbf{y}}{\partial \mathbf{h}}(\mathbf{h} - \mu_{\mathbf{h}}) + \frac{\partial \mathbf{y}}{\partial \mathbf{n}}(\mathbf{n} - \mu_{\mathbf{n}}). \quad (3.25)$$

The partial differential of $\mathbf{y}$ in Eq. (3.25) with respect to $\mathbf{s}$, $\mathbf{h}$, and $\mathbf{n}$ evaluated at $\mu_{\mathbf{z}} = \mu_{\mathbf{n}} - \mu_{\mathbf{s}} - \mu_{\mathbf{h}}$ can be written:

$$\left. \frac{\partial \mathbf{y}}{\partial \mathbf{s}} \right|_{(\mu_{\mathbf{n}}, \mu_{\mathbf{s}}, \mu_{\mathbf{h}})} = \left. \frac{\partial \mathbf{y}}{\partial \mathbf{h}} \right|_{(\mu_{\mathbf{n}}, \mu_{\mathbf{s}}, \mu_{\mathbf{h}})} = \mathbf{A}, \quad (3.26)$$

$$\left. \frac{\partial \mathbf{y}}{\partial \mathbf{n}} \right|_{(\mu_{\mathbf{n}}, \mu_{\mathbf{s}}, \mu_{\mathbf{h}})} = \mathbf{I} - \mathbf{A}, \quad (3.27)$$

where:

$$\mathbf{A} = \mathbf{C}\mathbf{F}\mathbf{C}^{-1}, \quad (3.28)$$

Where $\mathbf{F}$ is a diagonal matrix of which its elements are given by:

$$f(\mu_{\mathbf{z}}) = \frac{1}{1 + \exp(\mathbf{C}^{-1}\mu_{\mathbf{z}})}. \quad (3.29)$$

Then, the mean and the covariance of the pdf of noisy speech $\mathbf{y}$ are estimated using the following formulas:

$$\mu_{\mathbf{y}} = E[\mathbf{y}] \approx \mu_{\mathbf{s}} + \mu_{\mathbf{h}} + g(\mu_{\mathbf{s}}, \mu_{\mathbf{h}}, \mu_{\mathbf{n}}). \quad (3.30)$$

$$\Sigma_{\mathbf{y}} = \mathbf{A}\Sigma_{\mathbf{s}}\mathbf{A}^T + \mathbf{A}\Sigma_{\mathbf{h}}\mathbf{A}^T + (\mathbf{I} - \mathbf{A})\Sigma_{\mathbf{n}}(\mathbf{I} - \mathbf{A})^T. \quad (3.31)$$

3.2.2 Feature enhancement

3.2.2.1 Spectral subtraction

Spectral subtraction is a popular method to remove additive noise from noisy speech in the spectral domain. It was first proposed by Boll [1979]. In spectral subtraction, it

is assumed that the noise spectra are known. The estimate of the clean speech spectra are obtained by simply subtracting the noisy speech spectra with the noise spectra. It is formulated as follows:

$$|\hat{X}|^2 = |Y|^2 - |N|^2. \quad (3.32)$$

Spectral subtraction maintains a linear relation between noisy speech, noise and clean speech. Therefore, it is also called linear spectral subtraction (LSS).

In practice, the noise spectra are not known and must be estimated. For that reason, the performance of spectral subtraction heavily relies on the accuracy of the estimation of the noise spectra, and hence, it may not be suitable when noise is highly non-stationary.

3.2.2.2 MMSE-based feature enhancement

Spectral subtraction is an optimal maximum likelihood variance estimator under the Gaussian assumption. It is not an optimal spectral estimator. Ephraim and Malah [1984] proposed an MMSE-based feature enhancement to derive an optimal spectral estimator. Their method is derived on the magnitude spectral domain. Similar approaches have also been carried out in other domains such as the log magnitude spectral [Ephraim and Malah, 1985], filterbank [Stark and Paliwal, 2011] and cepstral domains [Yu et al., 2008]. In this section, we explain the derivation of this approach in the magnitude spectral domain.

The objective of this approach is to find the estimate of the clean speech magnitude spectrum $\hat{X}$, given the noisy spectrum Y based on MMSE criterion:

$$\begin{aligned} \hat{X} &= \hat{f}(Y) \\ &= \underset{f}{\operatorname{argmin}} E \left[(f(Y) - X)^2 \right] \\ &= \underset{f}{\operatorname{argmin}} \int (f(Y) - X)^2 p(X) dX. \end{aligned} \quad (3.33)$$

The solution for Eq. (3.33), assuming speech and noise are independent, is:

$$\hat{X} = E[X | \mathbf{Y}], \quad (3.34)$$

where $\mathbf{Y} = [Y(k=1)(Yk=2) \cdots Y(k=K)]$ denotes the observed noisy spectra for all spectra. Eq. (3.34) is basically the posterior probability of X . In [Ephraim and Malah, 1984], it is assumed that the real and imaginary parts of the Fourier transform coefficients follow zero mean Gaussian distributions and are independent to each other. It is also assumed that there is no correlation between each spectral component. Using these assumptions, Eq. (3.34) is simplified to:

$$\begin{aligned}\hat{X} &= E[X | Y], \\ &= \int_0^\infty xp(x | Y)dx, \\ &= \frac{\int_0^\infty xp(Y | x)p(x)}{\int_0^\infty p(Y|x)p(x)dx},\end{aligned}\tag{3.35}$$

where x is a realization of the random variable X . Since:

$$p(Y | X)p(X) = \int_0^{2\pi} p(Y | x, \theta_x)p(x, \theta_x)d\theta_x,\tag{3.36}$$

where θ_x is the realization of the phase random variable of X . Then we obtain:

$$\hat{X} = \frac{\int_0^\infty \int_0^{2\pi} xp(Y | x, \theta_x)p(x, \theta_x)d\theta_x dx}{\int_0^\infty \int_0^{2\pi} p(Y | x, \theta_x)p(x, \theta_x)d\theta_x dx}.\tag{3.37}$$

To obtain $\hat{X}$, we need to estimate $p(Y | x, \theta_x)$ and $p(x, \theta_x)$. Since speech and noise are assumed to be Gaussians, then Y is sum of two zero mean complex Gaussian random variables. Therefore, $p(Y | x, \theta_x)$ is also a Gaussian:

$$p(Y | x, \theta_x) = \frac{1}{\pi\lambda_n} \exp\left(-\frac{1}{\lambda_n}|Y - X|^2\right),\tag{3.38}$$

where $\lambda_n = E[|N|^2]$ is the noise variance. For complex Gaussian random variables, the magnitude and phase are independent. Therefore, $p(x, \theta_x)$ is the product of $p(x)$ and $p(\theta_x)$. The pdf of x is a Rayleigh distribution since the real and imaginary parts are Gaussian distributions. Meanwhile, the pdf of θ_x is uniform in $(-\pi, \pi)$. So, we obtain:

$$p(x, \theta_x) = \frac{x}{\pi\lambda_x} \exp\left(-\frac{x^2}{\lambda_x}\right),\tag{3.39}$$

where $\lambda_x = E[|X|^2]$ is the clean speech variance. Substituting (3.38) and (3.39) into (3.37), we obtain:

$$\hat{X} = \sqrt{\lambda} \Gamma(1.5) \Phi(-0.5, 1; -\nu), \quad (3.40)$$

where $\Gamma(\cdot)$ denotes Gamma function:

$$\Gamma(z) = \int_0^\infty \exp(-t) t^{z-1} dt, \quad (3.41)$$

$\Phi(a, b; c)$ denotes the confluent hypergeometric function:

$$\Phi(a, b; c) = 1 + \frac{a}{b} \frac{c}{1!} + \frac{a(a+1)}{b(b+1)} \frac{c}{2!} + \frac{a(a+1)(a+2)}{b(b+1)(b+2)} \frac{c}{3!} + \dots, \quad (3.42)$$

λ is defined by:

$$\lambda = \frac{\lambda_x \lambda_n}{\lambda_x + \lambda_n}, \quad (3.43)$$

and ν is formulated as follows:

$$\nu = \frac{\xi}{1 + \xi} \gamma, \quad (3.44)$$

where γ and ξ are defined by:

$$\gamma = \frac{|Y|^2}{\lambda_n}, \quad (3.45)$$

$$\xi = \frac{\lambda_x}{\lambda_n}. \quad (3.46)$$

Eq. (3.40) can also be written in term of Bessel function:

$$\hat{X} = \frac{\sqrt{\pi\nu}}{2\gamma} \exp\left(\frac{-\nu}{2}\right) \left[(1 + \nu) I_0\left(\frac{\nu}{2}\right) + \nu I_1\left(\frac{\nu}{2}\right) \right] Y, \quad (3.47)$$

where $I_0(\cdot)$ and $I_1(\cdot)$ are the modified Bessel functions of zero and first order:

$$I_v(a) = \sum_{m=0}^{\infty} \frac{a^{v+2m}}{2^{v+2m} m! \Gamma(v+m+1)}. \quad (3.48)$$

where j denotes the complex number.

3.2.2.3 Feature normalization

Feature normalization is a well established method to remove convolutive noise. CMN [Furui, 1981] is a popular feature normalization method. In CMN, we subtract the long term average from a feature, on the assumption that convolutive noise is stationary. LSMN [Avendano and Hermansky, 1997] is also based on the same principle as CMN. Their difference is the operational domain; CMN operates in the cepstral domain, whereas LSMN in the log spectral domain. In CMN, convolutive noise is assumed to be constant within each filter-bank, whereas in LSMN, it is constant within each frequency bin. Therefore, depend on the channel characteristics, the performance of LSMN and CMN can be different.

Let us consider the case when only convolutive noise exists. Therefore, Eq. (3.1) becomes:

$$|\check{Y}(m, l)|^2 = |\check{S}(m, l)|^2 |\check{H}(m, l)|^2, \quad (3.49)$$

In the cepstral domain, Eq. 3.49 becomes:

$$\mathbf{y}(m, l) = \mathbf{s}(m, l) + \mathbf{h}(m, l). \quad (3.50)$$

The long term average of $\mathbf{y}(m, l)$ is given by:

$$\bar{\mathbf{y}}(l) = \bar{\mathbf{s}}(l) + \bar{\mathbf{h}}(l), \quad (3.51)$$

where:

$$\bar{\mathbf{y}}(l) = \frac{1}{M} \sum_{m=1}^M \mathbf{y}(m, l), \quad (3.52)$$

$$\bar{\mathbf{s}}(l) = \frac{1}{M} \sum_{m=1}^M \mathbf{s}(m, l), \quad (3.53)$$

$$\bar{\mathbf{h}}(l) = \frac{1}{M} \sum_{m=1}^M \mathbf{h}(m, l). \quad (3.54)$$

Let $\tilde{\mathbf{s}}(m, l)$ be the normalized feature which is obtained by subtracting $\bar{\mathbf{s}}(l)$ from $\mathbf{s}(m, l)$.

It is mathematically written:

$$\tilde{\mathbf{s}}(m, l) = \mathbf{s}(m, l) - \bar{\mathbf{s}}(l). \quad (3.55)$$

By performing the same operation on noisy speech features, we obtain:

$$\begin{aligned} \tilde{\mathbf{y}}(m, l) &= \mathbf{y}(m, l) - \bar{\mathbf{y}}(l) \\ &= (\mathbf{s}(m, l) - \bar{\mathbf{s}}(l)) + (\mathbf{h}(m, l) - \bar{\mathbf{h}}(l)). \end{aligned} \quad (3.56)$$

Since convolutive noise is assumed to be stationary, then $\mathbf{h}(m, l) = \bar{\mathbf{h}}(l)$. Eq. (3.57) becomes:

$$\tilde{\mathbf{y}}(m, l) = \mathbf{s}(m, l) - \bar{\mathbf{s}}(l) = \tilde{\mathbf{s}}(m, l). \quad (3.57)$$

From Eq. (3.57), it is clear that CMN removes convolutive noise.

Mean variance normalization (MVN) is proposed in [Viikki and Laurila, 1998], which is an extension of CMN by including variance normalization. MVN is formulated as:

$$\tilde{\mathbf{y}}(m, l) = \frac{\mathbf{y}(m, l) - \bar{\mathbf{y}}(l)}{\sigma(l)}. \quad (3.58)$$

MVN is shown to improve the robustness against additive noise.

3.2.2.4 Vector Taylor series (VTS) for feature enhancement

While VTS can be used to adapt the clean speech model into the noisy environments, it can also be used in feature enhancement method. In this method, the noisy speech parameters are utilized to obtain the clean speech estimate using MMSE criterion [Moreno et al., 1996]:

$$\begin{aligned} \hat{\mathbf{s}}_{\text{MMSE}} &= E(\mathbf{s} \mid \mathbf{y}) \\ &= \int \mathbf{s} p(\mathbf{s} \mid \mathbf{y}) d\mathbf{s} \end{aligned} \quad (3.59)$$

Expressing $\mathbf{s}$ as a function of $\mathbf{y}$ and the mismatch function, $f(\mathbf{s}, \mathbf{h}, \mathbf{n})$, we can write Eq. (3.59):

$$\hat{\mathbf{s}}_{\text{MMSE}} = \int \mathbf{y} - f(\mathbf{s}, \mathbf{h}, \mathbf{n}) p(\mathbf{s} \mid \mathbf{y}) d\mathbf{s}. \quad (3.60)$$

GMMs are used to model speech and noisy speech with one-to-one correspond between speech and noisy speech models and noise is modeled as a Gaussian random variable. Therefore, Eq. (3.60) can be written:

$$\hat{\mathbf{s}}_{\text{MMSE}} = \mathbf{y} - \int \sum_{r=1}^R P[r | \mathbf{y}] f(\mathbf{s}, \mathbf{h}, \mathbf{n}) p[\mathbf{s} | r, \mathbf{y}] d\mathbf{s}, \quad (3.61)$$

where r is the index of mixture components and R denotes the total number of mixtures. We replace the mismatch function with its first order Taylor expansion, with the prior speech mean $\mu_{\mathbf{s}}$ and $\mu_{\mathbf{n}}$ are used as the operating points of the Taylor series. Therefore, we approximate Eq. (3.61) as follows:

$$\begin{aligned} \hat{\mathbf{s}}_{\text{MMSE}} &\approx \mathbf{y} - \sum_{r=1}^R P[r | \mathbf{y}] \int f(\mu_{\mathbf{s}}, \mathbf{h}, \mu_{\mathbf{n}}) p[\mathbf{s} | r, \mathbf{y}] d\mathbf{s}, \\ &= \mathbf{y} - \sum_{r=1}^R P[r | \mathbf{y}] f(\mu_{\mathbf{s}}, \mathbf{h}, \mu_{\mathbf{n}}). \end{aligned} \quad (3.62)$$

We need to estimate the posterior probability of each mixture component $P(r | \mathbf{y})$. It can be calculated as follows:

$$P(r | \mathbf{y}) = \frac{p(\mathbf{y} | r) P(r)}{p(\mathbf{y})} = \frac{p(\mathbf{y} | r) P(r)}{\sum_{r=1}^R p(\mathbf{y} | r) P(r)} \quad (3.63)$$

The distribution of noisy speech $p(\mathbf{y} | r)$ can be calculated as described in Section 3.2.1.2.

3.2.2.5 ETSI advanced front end (AFE)

The European Telecommunications Standards Institute (ETSI) released a standard front end for distributed speech recognition [ETSI standard doc., 2002]. This front-end is shown to perform well on the Aurora-2 database, a popular database used to evaluate robust speech recognition methods [Macho et al., 2002]. The noise reduction process of ETSI AFE is shown in Fig. 3.2. In this standard, the two stage mel-warped Wiener filters [Agarwal and Cheng, 1999] are implemented for removing additive noise, and blind equalization (BE) [Mauuary, 1996] is employed for removing convolutive noise.

The transfer functions of the Wiener filters are calculated in the frequency domain, but the filtering is performed in the time domain. To obtain the estimation of the

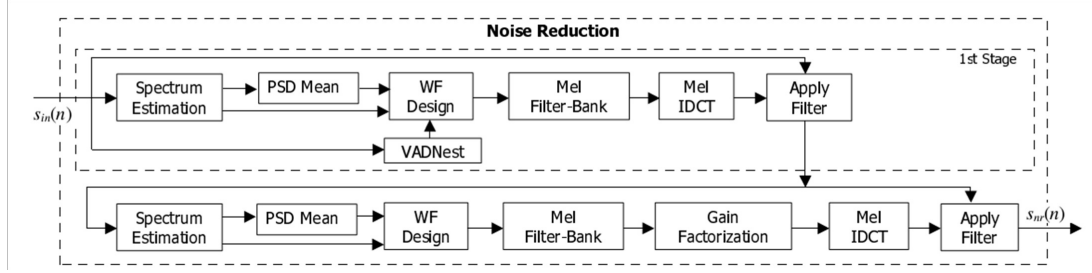


Figure 3.2: Noise reduction scheme in ETSI AFE [Macho et al., 2002]

filter, $S_{\text{in}}(m, k)$ and $S_{\text{N}}(m, k)$, the power spectrum density (PSD) of the noisy signal and noise respectively, is first calculated using the following formula:

$$S_{\text{in}}(m, k) = \frac{1}{T_{\text{PSD}}} \sum_{i=0}^{T_{\text{PSD}}-1} |Y(m-i, k)|^2, \quad (3.64)$$

$$S_{\text{N}}(m, k) = \frac{1}{T_{\text{PSD}}} \sum_{i=0}^{T_{\text{PSD}}-1} |\hat{N}(m-i, k)|^2, \quad (3.65)$$

where $T_{\text{PSD}} = 2$, and $|\hat{N}(m-i, k)|^2$ is the estimate of noise power spectrum of the previous frame. The first stage Wiener filter is formulated as:

$$H_1(m, k) = \frac{\eta_1(m, k)}{1 + \eta_1(m, k)}, \quad (3.66)$$

where:

$$\eta_1(m, k) = \frac{S_{\text{den}}(m, k)}{S_{\text{N}}(m, k)}, \quad (3.67)$$

and $S_{\text{den}}(m, k)$ is the smoothed noisy spectrum, which is obtained using the following formula:

$$S_{\text{den}}(m, k) = \delta S_{\text{den3}}(m-1, k) + (1 - \delta) \max(S_{\text{in}}(m, k) - S_{\text{N}}(m, k), 0), \quad (3.68)$$

where $\delta = 0.98$ and $S_{\text{den3}}(m-1, k)$ is the de-noised spectrum from the previous frame:

$$S_{\text{den3}}(m-1, k) = H_2(m-1, k) S_{\text{in}}(m-1, k). \quad (3.69)$$

The second stage of the Wiener filter is calculated as follows:

$$H_2(m, k) = \frac{\eta_2(m, k)}{1 + \eta_2(m, k)}, \quad (3.70)$$

where:

$$\eta_2(m, k) = \max \left(\frac{S_{\text{den2}}(m, k)}{S_N(m, k)}, \eta_{th} \right), \quad (3.71)$$

$\eta_{th} = 0.079432823$, which corresponds to the maximum filter attenuation of 11.33 dB, and $S_{\text{den2}}(m, k)$ is computed by applying the first stage Wiener filter to the input signal PSD:

$$S_{\text{den2}}(m, k) = H_1(m, k)S_{\text{in}}(m, k). \quad (3.72)$$

After estimating the filters' transfer function in the frequency domain, they are transformed into a Mel-frequency scale by using 23 triangular Mel-warped frequency windows. Then, the results are transformed back to the time domain to obtain the impulse response of the filters by taking a Mel-warped inverse cosine transform. The estimated clean speech is obtained by convolving the noisy input signal with the impulse response of the Wiener filters.

In BE, the speech cepstrum is normalized by subtracting the bias of each frame from the speech cepstrum. The bias is calculated from a reference cepstrum, which corresponds to the cepstrum of a flat spectrum. By performing BE, the effect of convolutive noise could be reduced.

3.2.3 Dealing with the cross-term

The methods explained in Section 3.2.1 and 3.2.2 are derived based on the assumption that speech and noise are uncorrelated. Under this assumption, the cross-term is ignored and Gaussian distributions are used to model clean and noisy speech features. However, these assumptions are generally not true. Even though speech and noise are statistically independent, the use of short-time window introduce a cross-term between speech and noise in practice. Hence, the additivity does not generally hold. The expected value of the cross-term is zero. However, its magnitude at a particular frame could be significantly large. It has been shown to significantly degrade the performance

of speech recognition [Deng et al., 2004; Evans et al., 2006; Faubel et al., 2008]. Furthermore, the Gaussian distributions are not the best model for speech either [Chen and Loizou, 2007; Martin, 2005]. The importance of the cross-term in speech recognition attracts many studies to include it in deriving robust speech recognition methods. Other types of distributions have also been proposed to replace the Gaussian model. In this section, we explain some variants of robust speech recognition methods that consider correlation between noise and speech.

3.2.3.1 Non-linear spectral subtraction

A simplest way to reduce the effect of correlation between speech and noise is introducing a weighting factor as employed in nonlinear spectral subtraction (NSS) [Berouti et al., 1979]. In this method, an oversubtraction factor, α , is introduced. The original intention is to reduce the musical noise which often exists after implementing spectral subtraction. This parameter is usually set to depend on the signal-to-noise ratio (SNR). In NSS, the spectral subtraction formula is modified:

$$|\hat{X}|^2 = |Y|^2 - \alpha(\text{SNR})|N|^2. \quad (3.73)$$

Several variants of NSS have been develop [Flores and Young, 1994; Lockwood and Boudy, 1992; Poruba, 2002; Udrea et al., 2008]. They are basically modification of Eq. (3.73).

Zhu and Alwan [2002] reported that α was related to the missing cross-term. This can be justified as follows. Recalling Eq. (1.7), which is the relation between noisy speech, clean speech, and additive noise in the power spectral domain:

$$|Y(m, k)|^2 = |X|^2 + |N|^2 + 2|X||N| \cos(\theta_X - \theta_N), \quad (3.74)$$

By denoting $\xi = \frac{|N|}{|X|}$, Eq. (3.74) can be written:

$$|Y|^2 = [1 + 2\xi \cos(\theta_X - \theta_N)] |X|^2 + |N|^2 \quad (3.75)$$

From Eq. (3.75), it can be seen that the cross-term can be treated as a weight factor, which depends on the ratio between the magnitude spectra of speech and noise. In other words, inverse of the square root of the SNR. This weight is close to one when

the SNR is high, but it becomes significant when the SNR is low. As the consequences, $\alpha > 1$ is implemented when the SNR is low. NSS is shown to be better than LSS in improving the robustness of speech recognition [Zhu and Alwan, 2002].

3.2.3.2 MMSE-based methods using non Gaussian distributions

In Section 3.2.2.2, MMSE spectral estimator based on the assumption that speech spectrum follows a complex Gaussian distribution is derived. However, several studies show that short time speech spectrums do not follow Gaussian distributions [Gazor and Zhang, 2003; Wilbur B. Davenport, 1952]. Laplace [Chen and Loizou, 2007] and Gamma [Martin, 2002] distributions are often used to model speech instead since they are found to fit speech spectra better than Gaussian distributions. Accordingly, their MMSE spectral estimators have also been derived. It is found that the quality of the clean speech estimate derived under these models is better than that of the Gaussian distribution when it is evaluated using objective quality measures. However, to our knowledge, the evaluation these methods for speech recognition tasks has not been performed yet.

3.2.3.3 Phase sensitive VTS

VTS-based robust speech recognition methods in Sections 3.2.2.4 and 3.2.1.2 are derived under the assumption that the cross-term is zero. There have been several attempts to derive VTS-based methods without this assumption [Deng et al., 2004; Faubel et al., 2008; Li et al., 2009; Stouten et al., 2005]. Since this approach includes the phase information, it is called phase sensitive approach. This approach can also be implemented for either feature enhancement [Deng et al., 2004; Stouten et al., 2005] or model compensation techniques [Li et al., 2009]. In this section, we only explain the VTS-based method for model compensation.

In this method, the relation in Eq. (3.7) is used instead of Eq. (3.15). Assuming that α is independent of $\mathbf{n}$, $\mathbf{s}$ and $\mathbf{h}$, we can approximate Eq. (3.7) using first order VTS as follows:

$$\mathbf{y} = \mu_{\mathbf{s}} + \mu_{\mathbf{h}} + g(\mu_{\mathbf{s}}, \mu_{\mathbf{h}}, \mu_{\mathbf{n}}) + \mathbf{G}(\mathbf{s} - \mu_{\mathbf{s}}) + \mathbf{G}(\mathbf{h} - \mu_{\mathbf{h}}) + (1 - \mathbf{G})(\mathbf{n} - \mu_{\mathbf{n}}) \quad (3.76)$$

where:

$$\mathbf{G} = \mathbf{I} - \mathbf{C} \text{diag} \left[\frac{\exp(\mathbf{C}^{-1}(\mu_{\mathbf{n}} - \mu_{\mathbf{s}} - \mu_{\mathbf{h}})) + \alpha \exp(0.5\mathbf{C}^{-1}(\mu_{\mathbf{n}} - \mu_{\mathbf{s}} - \mu_{\mathbf{h}}))}{1 + \exp(\mathbf{C}^{-1}(\mu_{\mathbf{n}} - \mu_{\mathbf{s}} - \mu_{\mathbf{h}})) + 2\alpha \exp(0.5\mathbf{C}^{-1}(\mu_{\mathbf{n}} - \mu_{\mathbf{s}} - \mu_{\mathbf{h}}))} \right] \mathbf{C}^{-1} \quad (3.77)$$

By taking the expectation of both sides of Eq. (3.76), we obtain the mean of noisy speech distribution as follows:

$$\mu_{\mathbf{y}} \approx \mu_{\mathbf{s}} + \mu_{\mathbf{h}} + g(\mu_{\mathbf{s}}, \mu_{\mathbf{h}}, \mu_{\mathbf{n}}) \quad (3.78)$$

And the variance:

$$\Sigma_{\mathbf{y}} = \mathbf{G}\Sigma_{\mathbf{s}}\mathbf{G}^T + (\mathbf{I} - \mathbf{G})\Sigma_{\mathbf{s}}(\mathbf{I} - \mathbf{G})^T \quad (3.79)$$

Eqs. (3.78) and (3.79) are similar with Eqs. (3.30) and (3.31) except there is an additional variable α needs to be estimated as well. To simplify the solution, in [Li et al., 2009], the convolutive noise and α are considered as deterministic quantities, which are obtained empirically.

The distribution obtained from this approach can also be used to obtain the clean speech estimate using MMSE criterion [Deng et al., 2004]. In this method, the phase factor α is assumed to follow a Gaussian distribution. The phase sensitive approach is shown to be better than phase insensitive VTS-based method for improving the robustness of speech recognition.

3.3 Summary

In this chapter, we explained several popular methods for improving the robustness of speech recognition against environmental noise. We can classify those methods into two category: feature enhancement and model compensation. Feature enhancement methods work on the feature domain. The goal is to remove noise from speech features. The examples of this category are spectral subtraction, CMN, and ETSI AFE. On the other hand, model compensation operates in the model domain. The objective is to adapt the clean speech models into noisy environments and then performs the recognition using the new model. PMC and VTS are popular examples of this category.

The assumption used in deriving these methods is that noise and speech are uncorrelated. Since this assumption is generally not true, the performance of these methods are not optimal. When these methods are extended, by considering the correlation term between speech and noise, a significantly better results are obtained. These shows the importance of considering the correlation between noise and speech. However, finding the correlation between speech and noise are not easy and in many cases are impossible to perform.

Chapter 4

Non-extensive statistics: a review

4.1 Introduction

Many systems in physics, economics, sociology and biology are considered *complex*. In such systems, we do not fully know about the sub-systems and their relations to each other. Therefore, the additive property does not hold. Boltzmann-Gibbs statistics and Shannon entropy, which are extensive frameworks, are unable to explain some phenomena in the complex systems. Hence, their implementations may not be optimal for such systems. A theory of non-extensive statistics is introduced [Tsallis, 1988] to explain the non-additive phenomena in the complex systems in statistical mechanics. This theory generalizes Boltzmann-Gibbs statistics. It has been proved that this framework represents the complex systems better than Boltzmann-Gibbs statistics do.

In this chapter, we briefly review the theory of non-extensive statistics. In Section 4.2, we explain the non-extensive functions, which are the fundamental functions in this theory, and their properties. In Section 4.3, we briefly explain the Tsallis entropy, an entropy defined in this framework. In Section 4.4, we explain two distributions derived from Tsallis entropy: q -exponential and q -Gaussian distributions. In Section 4.5, we give several implementations of the non-extensive statistics. Section 4.6 summarizes this chapter and gives an outlook about the main challenge in implementing this framework.

4.2 Non-extensive functions

4.2.1 Definitions

The theory of non-extensive statistics generalizes Boltzmann-Gibbs statistics by utilizing the q -exponential (q -exp) function and its inverse, the q -logarithmic (q -log) functions. The q -exp function is defined as the following:

$$\exp_q(x) = (1 + (1 - q)x)^{\frac{1}{1-q}}, \quad (4.1)$$

and the q -log function is defined as:

$$\log_q(x) = \frac{x^{1-q} - 1}{1 - q}. \quad (4.2)$$

These functions asymptotically approach the exponential and natural logarithmic functions as q approaches 1. Therefore, the q -exp and the q -log functions are one type of generalizations of the exponential and the natural log functions respectively. Fig. 4.1 and 4.2 illustrate the q -exp and the q -log functions for real number x respectively. It can be seen that these functions vary with parameter q . They approach $\exp(x)$ and $\log(x)$ respectively when q is close to 1. They become linear when $q = 0$.

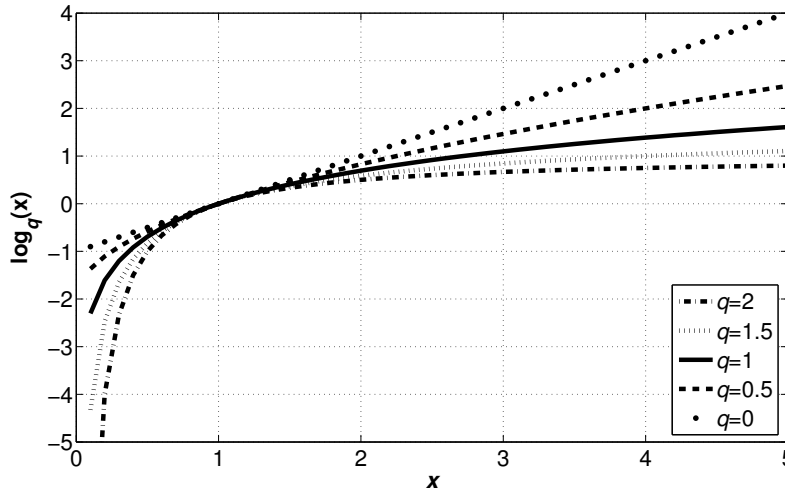


Figure 4.1: The q -logarithmic function of real variable x for different q -values.

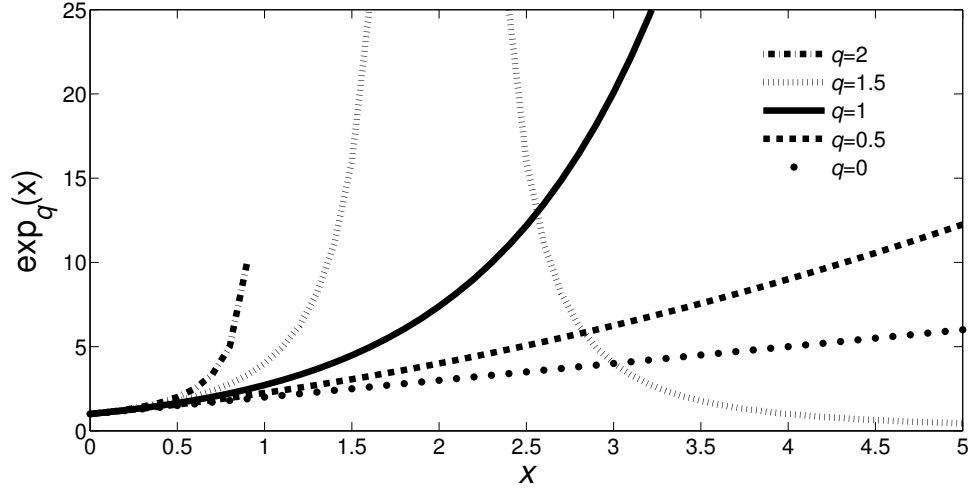


Figure 4.2: The q -exponential function of real variable x for different q -values.

4.2.2 Properties

The exponential and natural logarithmic functions are morphism functions. These functions transform a multiplication into an addition and vice versa: $\exp(x)\exp(y) = \exp(x+y)$ and $\log(xy) = \log(x) + \log(y)$. The q -exp and q -log functions however do not share the same properties when $q \neq 1$. They introduce a pseudo-additivity when $q \neq 1$ [Borges, 2004; Nivanen et al., 2003]:

$$\exp_q(x)\exp_q(y) = \exp_q(x+y+(1-q)xy), \quad (4.3)$$

$$\log_q(xy) = \log_q(x) + \log_q(y) + (1-q)\log_q(x)\log_q(y). \quad (4.4)$$

Therefore, the standard addition and multiplication operators cannot be applied to them.

The properties of the q -exp and the q -log functions inspired a new type of operators. Let $+_q$ and $\times_q$ be the generalized addition and generalized multiplication operators respectively. These operators depend on parameter q and are defined as:

$$x+_q y = x+y+(1-q)xy, \quad (4.5)$$

$$x \times_q y = (x^{1-q} + y^{1-q} - 1)^{\frac{1}{1-q}}. \quad (4.6)$$

Based on these operators, we can re-write the properties of the q -exp and the q -log functions as the following:

$$\exp_q(x) \exp_q(y) = \exp_q(x +_q y), \quad (4.7)$$

$$\exp_q(x) \times_q \exp_q(y) = \exp_q(x + y), \quad (4.8)$$

$$\log_q(xy) = \log_q(x) +_q \log_q(y), \quad (4.9)$$

$$\log_q(x \times_q y) = \log_q(x) + \log_q(y). \quad (4.10)$$

In the same way, the generalized subtraction, $-_q$, and division, $/_q$ operators can be defined as follows:

$$x -_q y = \frac{x + y}{1 + (1 - q)xy}, \quad (4.11)$$

$$x /_q y = (x^{1-q} + y^{1-q} + 1)^{\frac{1}{1-q}}. \quad (4.12)$$

Based on these operators, the rest of the properties of the q -exp and the q -log functions can be written:

$$\exp_q(x) / \exp_q(y) = \exp_q(x -_q y), \quad (4.13)$$

$$\exp_q(x) /_q \exp_q(y) = \exp_q(x - y), \quad (4.14)$$

$$\log_q(x/y) = \log_q(x) -_q \log_q(y), \quad (4.15)$$

$$\log_q(x /_q y) = \log_q(x) - \log_q(y). \quad (4.16)$$

The properties of the q -exp and the q -log functions are the same as those of the ex-

ponential and logarithmic functions respectively when $q = 1$. Further details about the q -exp and the q -log functions and their properties can be found in [Borges, 2004; Nivanen et al., 2003].

4.3 Tsallis entropy

In this framework, entropy is redefined as the following:

$$\begin{aligned} S_q &= -k \sum_{i=1}^W p_i^q \log_q p_i \\ &= k \frac{1 - \sum_{i=1}^W p_i^q}{q - 1}, \end{aligned} \quad (4.17)$$

where W is the number of states, k is the Boltzmann constant, p_i is the probability for each state, and $\sum_{i=1}^W p_i = 1$. This entropy, Tsallis entropy, is a generalization of Shannon entropy.

In Shannon theory, the total entropy of a system is the sum of its subsystems' entropies. In other words, Shannon entropy is an extensive entropy. It can be applied to those systems which have a known structure, e.g, the number of subsystems and the relation between them. However, it cannot be applied to *complex* systems where we do not know well about the subsystems and their relation to each other.

Tsallis entropy is, on the other hand, a non-extensive entropy. The pseudo-additivity properties of the q -log function are used to explain the non-additive phenomenon in complex systems. Let A and B be two sub-systems of a complex system X . Then:

$$S_q(X) = S_q(A) + S_q(B) + (1 - q)S_q(A)S_q(B). \quad (4.18)$$

The third term of Eq. (4.18) represents the correlation between A and B , the degree of which depends on the choice of q . Tsallis entropy is a non-additive property when $q \neq 1$. It will be additive and the same as Shannon entropy when $q = 1$. The more detail about Tsallis entropy can be found in [Tsallis, 1988, 2002, 2009].

4.4 Tsallis distributions

Several distributions have been derived from maximizing Tsallis entropy under appropriate constraints. In this section, two of them are explained, the q -Gaussian and

q -exponential distributions.

4.4.1 Q -Gaussian distribution

A q -Gaussian distribution can be obtained by maximizing the Tsallis entropy in a similar way as a Gaussian distribution can be derived from Shannon entropy. The density function for a q -Gaussian distribution with mean μ_q and variance λ_q is defined by:

$$p_q(x) = \frac{A_q B_q}{\sqrt{\lambda_q}} \exp_q \left(-\frac{B_q^2 (x - \mu_q)^2}{\lambda_q} \right), \quad (4.19)$$

where A_q is a normalization term and is defined as:

$$A_q = \begin{cases} \frac{\Gamma\left(\frac{5-3q}{2-2q}\right)}{\Gamma\left(\frac{2-q}{1-q}\right)} \sqrt{\frac{1-q}{\pi}} & -\infty < q < 1 \\ \frac{1}{\sqrt{\pi}} & q = 1 \\ \frac{\Gamma\left(\frac{1}{q-1}\right)}{\Gamma\left(\frac{3-q}{2q-2}\right)} \sqrt{\frac{q-1}{\pi}} & 1 < q < 3, \end{cases} \quad (4.20)$$

and B_q is a scaling factor and in a normalized distribution $B_q = \frac{1}{\sqrt{3-q}}$. Fig. 4.3 shows the probability distributions of q -Gaussian for several values of q . The q -Gaussian distribution is a compact support distribution when $q < 1$ and a heavy-tailed distribution when $1 < q < 3$. The q -Gaussian distribution is identical with the Gaussian distribution when $q = 1$. In the heavy tail regions, the q -Gaussian is equivalent to the Student's t -distribution, which probability density function is formulated as:

$$p(x) = \frac{\Gamma\left(\frac{w+1}{2}\right)}{\sqrt{w\pi}\Gamma\left(\frac{w}{2}\right)} \left(1 + \frac{x^2}{w}\right)^{-\frac{w+1}{2}}, \quad (4.21)$$

where w is the number of degrees of freedom. The relation between q and w is formulated as follows:

$$q = \frac{w+2}{w+1}. \quad (4.22)$$

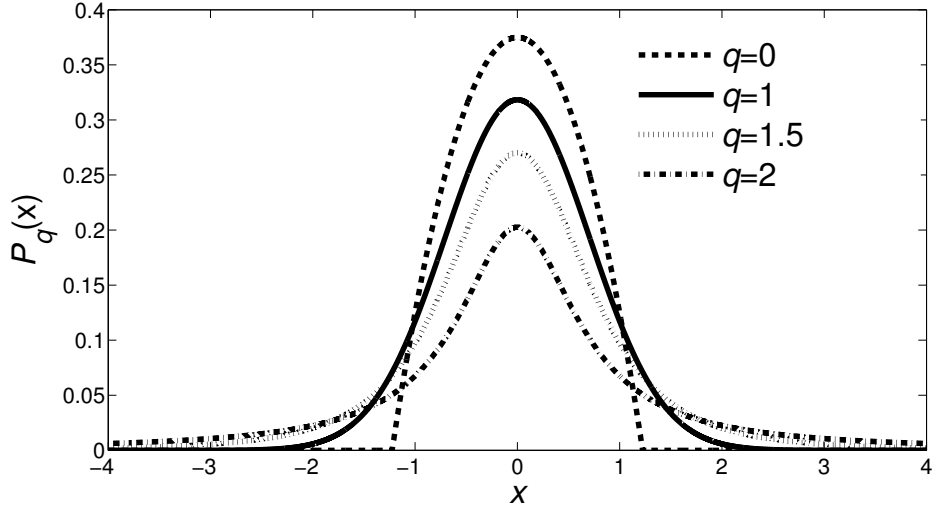


Figure 4.3: q -Gaussian distribution for several q

4.4.2 Q -Exponential distribution

The q -exponential distribution is derived by maximizing the Tsallis entropy when the variable domain is constrained to be positive. The density function for a q -exponential distribution is defined as follows:

$$p_q(y) = (2 - q)\lambda \exp_q(-\lambda y). \quad (4.23)$$

The mean μ_q is calculated using the following formula:

$$\mu_q = \begin{cases} \frac{1}{\lambda(3-2q)} & q < \frac{3}{2}, \\ \text{otherwise undefined,} & \end{cases} \quad (4.24)$$

and the variance σ_q :

$$\sigma_q = \frac{q - 2}{(2q - 3)^2(3q - 4)\lambda^2} \quad (4.25)$$

Figure 4.4 shows the probability distributions of q -exponential for several values of q and λ . Similar to the q -Gaussian distribution, the q -exponential distribution is also a heavy-tailed distribution when $q > 1$.

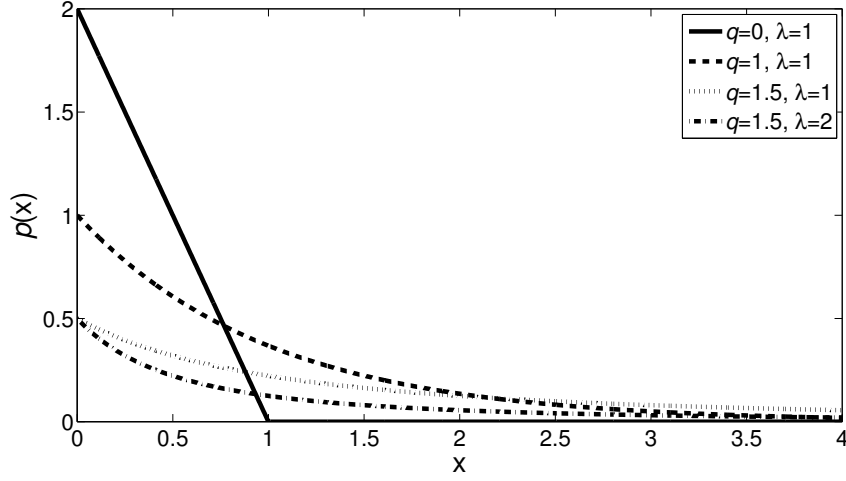


Figure 4.4: q -exponential distribution for several q and λ

The q -exponential distribution is a special case of the Lomax distribution (Pareto Type II):

$$p(y) = \frac{\alpha}{\lambda_l} \left[1 + \frac{y}{\lambda} \right]^{-(\alpha-1)}, \quad y \geq 0. \quad (4.26)$$

The relations between the parameters of the Lomax distribution and the q -exponential distribution are given by:

$$\alpha = \frac{2-q}{q-1}, \quad (4.27)$$

$$\lambda_l = \frac{1}{\lambda(q-1)}, \quad (4.28)$$

4.5 Implementations

The non-extensive framework has been employed in the study of complex systems in many areas in physics such as cosmology [Plastino et al., 2004], fractals [Olemskoi et al., 2010], nuclear physics [Wilk and Włodarczyk, 2002], and self-gravitating systems [Jiulin, 2007]. This theory has also been successfully applied to the other fields. For instance in finance and economics, Tsallis entropy and the q -Gaussian distribution have been implemented to model the financial risk in the stock market [Gradojevic

and Gencay, 2011]. In biological system, the q -Gaussian distribution is found to fit the distribution of the protein volume [Moret, 2011]. In medicine, Tsallis entropy is used to measure the complexity of the brain rhythms [Bezerianos et al., 2003; Weili et al., 2009]. The q -exponential distribution has been found to fit the empirical data of the cities' population [Malacarne et al., 2001], the citations of scientific journal [Tsallis and de Albuquerque, 2000], the DNA sequences [Oikonomou et al., 2008], and the earthquakes [Hasumi, 2009].

4.6 Summary

In this chapter, we briefly explain the theory of non-extensive statistics. This theory utilizes the pseudo-additivity properties of the q -exp and the q -log functions to generalize Boltzmann-Gibbs statistisc. This framework is able to explain some non-additive phenomena in the complex systems. On this framework, the entropy, which is called Tsallis entropy, is redefined and set of distributions are derived, e.g. the q -Gaussian distribution and the q -exponential distribution. This framework has been successfully implemented in many complex systems in physics, economics and finance, medicine, biology and geology.

In this framework, the parameter q is used to represent the degree of complexity of a system [Tsallis, 2002]. However, the physical meaning of q which is still not clear and highly debated. This becomes one of the challenges in non-extensive statistics. To our knowledge, an automatic method for optimizing q has not been found. Usually, it is empirically determined to fit the non-additive phenomena in the complex systems.

Chapter 5

Feature normalization based on non-extensive statistics

5.1 Introduction

The performance of the robust speech recognition techniques is improved when we consider speech and noise to be correlated. In the current extensive framework, we need to estimate the cross-term to calculate the correlation between speech and noise. To estimate the cross-term, we need to estimate both the noise spectra and the phase difference between speech and noise [Deng et al., 2004; Faubel et al., 2008; Lu and Loizou, 2008]. However, it is difficult to estimate the speech phase in a short window [Alsteris and Paliwal, 2007]. To estimate the noise spectra is also difficult since noise is usually non-stationary. The cross-term introduces non-linearity between speech components and alters the distribution of the speech features in the log spectral or cepstral domain. It cannot be removed by a simple linear normalization method such as CMN.

In this chapter, we derive a feature normalization method based on non-extensive statistics to solve the problem of cross-term. We introduce the q -log spectral domain in non-extensive statistics [Tsallis, 1988], which can naturally represent noisy speech including the cross-term, and propose q -log spectral mean normalization (q -LSMN) which is an extension of LSMN to the q -log spectral domain. Our proposed method, q -LSMN, is a nonlinear normalization method which only normalizes speech features but also removes the cross-term.

Few studies have already employed the non-extensive statistics for speech recognition as well. Rufiner et al. [Rufiner et al., 2004] proposed to add Tsallis entropy and its relative change to the standard MFCC features for capturing the dynamics of the speech signals. The q -log function was employed in [Kobayashi and Imai, 1984] as the spectral smoother for the speech features so that they were more robust against noise. The purpose of this approach was to reduce the dynamic range of the spectrum in the low frequency region. The q -log function was also implemented in [Ito et al., 2000] to provide a forward masking scale for the dynamic cepstra. In contrast to these studies, we use the q -log function to model non-additivity in the noisy speech features when noise and speech are correlated.

This chapter is organized as follows: In Section 5.2, we explain how the cross-term causes non-additivity in the speech features. In Section 5.3, we explain our framework in more detail. In Section 5.4, we derive and explain our proposed method, q -LSMN. In Section 5.5, we describe the details of spectral subtraction which we implement to remove additive noise. In Section 5.6, we describe the experimental setup to evaluate our proposed method. In Section 5.7, we present and discuss our experimental results. Section 5.8 concludes this chapter.

5.2 Non-additivity in noisy speech features

In this section, we examine the effect of the cross-term on the speech features. We show how clean speech cannot be recovered even though the noise spectrum can be estimated correctly. We recall Eq. (1.7), that is the relation between noisy speech, clean speech, additive noise, and convolutive noise in the power spectral domain (we omit the subscripts m and k for simplicity):

$$|Y|^2 = |X|^2 + |N|^2 + 2|X||N|\cos(\theta_X - \theta_N), \quad (5.1)$$

Let $|\hat{N}|^2$ be the estimate of the noise power spectrum. We can obtain $|\hat{X}|^2$, which is the estimation of $|X|^2$, as the following:

$$|\hat{X}|^2 = |Y|^2 - |\hat{N}|^2. \quad (5.2)$$

Substituting (1.7) into (5.2) and denoting $\theta = \theta_X - \theta_N$, we obtain:

$$|\hat{X}|^2 = |X|^2 + \left(|N|^2 - |\hat{N}|^2\right) + 2|X||N|\cos\theta. \quad (5.3)$$

If we assume the noise spectrum is estimated correctly, i.e. $|N|^2 - |\hat{N}|^2$ is zero, Eq. (5.3) becomes:

$$|\hat{X}|^2 = |X|^2 + 2|X||N|\cos\theta. \quad (5.4)$$

Meanwhile, in the power spectral domain:

$$|X|^2 = |S|^2|H|^2, \quad (5.5)$$

where $|S|^2$ and $|H|^2$ are the power spectra of clean speech and convolutive noise respectively. By substituting (5.5) into (5.4), we can rewrite Eq. (5.4) as follows:

$$|\hat{X}|^2 = |S|^2|H|^2\chi, \quad (5.6)$$

where:

$$\chi = 1 + 2\frac{|N|}{|S||H|}\cos\theta. \quad (5.7)$$

By taking the log of Eq. (5.6), we obtain:

$$\hat{\mathbf{x}} = \mathbf{s} + \mathbf{h} + \boldsymbol{\chi}, \quad (5.8)$$

where $\hat{\mathbf{x}}$, $\mathbf{s}$, and $\mathbf{h}$ are the log of $|X|^2$, $|S|^2$, and $|H|^2$ respectively, and:

$$\boldsymbol{\chi} = \log\left(1 + 2\frac{|N|}{|S||H|}\cos\theta\right). \quad (5.9)$$

The cross-term in Eq. (5.9) is zero when the distance between the speech and noise components in frequency space is greater than the frequency resolution in STFT [Kadambé and Boudreaux-Bartels, 1992]. However, the noise and speech spectra are usually close to each other, causing them to overlap, in which case the resulting cross-term could be substantially large in magnitude.

By performing a linear normalization, such as CMN or LSMN, we obtain:

$$\tilde{\mathbf{x}} = (\mathbf{s} + \mathbf{h} + \boldsymbol{\chi}) - (\bar{\mathbf{s}} + \bar{\mathbf{h}} + \bar{\boldsymbol{\chi}}), \quad (5.10)$$

where:

$$\bar{\boldsymbol{\chi}} = \frac{1}{M} \sum_{m=1}^M \boldsymbol{\chi}^{(m,k)}. \quad (5.11)$$

Since $\mathbb{E}\{\cos \theta\} = 0$, then we can safely assume $\bar{\boldsymbol{\chi}} = 0$. By assuming convolutive noise to be stationary, Eq. (5.10) becomes:

$$\begin{aligned} \tilde{\mathbf{x}} &= \mathbf{s} - \bar{\mathbf{s}} + \boldsymbol{\chi} \\ &= \tilde{\mathbf{s}} + \boldsymbol{\chi}. \end{aligned} \quad (5.12)$$

From Eq. (5.12), it is obvious that the clean speech can not be recovered even though additive noise and convolutive noise can be removed completely.

Several studies have reported that the cross-term deteriorates the performance of ASR. Several reports [Deng et al., 2004; Zhu and Alwan, 2002] showed that the clean speech estimate was not able to achieve the same accuracy as the clean speech when the cross-term was ignored, even when the noise spectra were perfectly estimated. Furthermore, Evans et al. [Evans et al., 2006] reported that the effect of the cross-term is more significant on the accuracy of speech recognition in the low SNR conditions.

5.3 Framework

In Section 5.2, it is shown that the cross-term, $\boldsymbol{\chi}$, is introduced in the log spectral domain when a short time window is used in signal processing. It significantly degrades the performance of speech recognition. Inspired by the theory of non-extensive statistics, we implement the q -log function to provide a new domain, q -log spectral domain, which can effectively represent phenomena with non-additive nature.

It is difficult to replace the whole extensive framework for the present speech recognition systems with the non-extensive one, since the additivity does not hold in this domain and algorithms for model parameter estimation cannot always be provided. Here, we employ a *plug-in* approach with the q -log spectral domain as an intermediate

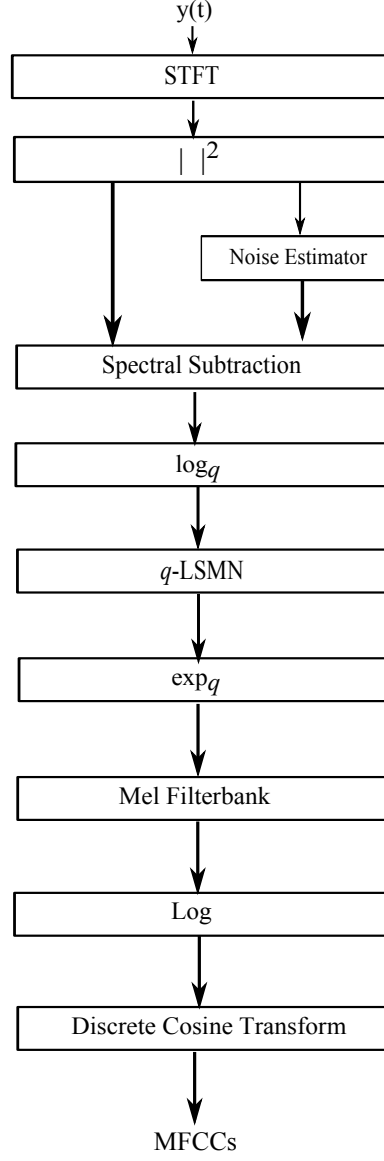


Figure 5.1: Block diagram of the proposed front-end.

domain. In this approach, first, the speech features are transformed into those in the q -log spectral domain. Next, the features are normalized in this domain using a new technique, q -log spectral mean normalization (q -LSMN). Then, the normalized features are re-transformed back to the original spectral domain. The rest of the speech recognition process remains the same as the conventional process. In this way, we can utilize a method based on non-extensive statistics within the present framework of the extensive

statistics.

Figure 5.1 shows the block diagram of our front-end. The speech signal is first framed using a 25 ms Hamming window with a 10 ms frame shift. Then we perform STFT and take the square of each spectral component to obtain the power spectra of the signal. We implement spectral subtraction to remove the additive noise spectra. After q -LSMN is carried out, a standard MFCC feature extraction as explained in Section 2.2 is performed.

It should be noted that non-additivity in the speech features not only comes from the cross-term but also from other factors such as the nature of speech pattern itself, which is also *complex*. Various long-term correlations also exist among its different spectral components in complex ways in various time-scales [Cohen, 2005; McAuley et al., 2005; Ming et al., 1996]. Such correlations may also be compensated for using our approach.

5.4 Q -Log spectral mean normalization (Q -LSMN)

In Eq. (5.12), we show that the conventional feature normalization methods cannot remove the cross-term in noisy speech spectra. It still exists even when the spectral subtraction and LSMN are ideally conducted. In this section, we derive our method, q -LSMN, and show the effect of q -LSMN on reducing the cross-term.

In spectral domain, LSMN is equal to normalizing the spectrum with its arithmetical mean of the logarithm of $|S(m, k)|^2$:

$$|\tilde{S}(m, k)|^2 = \frac{|S(m, k)|^2}{\exp \left[\frac{1}{M} \sum_{m=1}^M \mathbf{s}(m, k) \right]}. \quad (5.13)$$

Meanwhile, the arithmetical mean of the q -logarithm of $|S(m, k)|^2$ is calculated as follows:

$$\bar{\mathbf{s}}_q(k) = \frac{1}{M} \sum_{m=1}^M \mathbf{s}_q(m, k). \quad (5.14)$$

By normalizing the spectrum with its arithmetical mean of the q -logarithm of $|S(m, k)|^2$,

we obtain:

$$|\tilde{S}(m, k)|^2 = \frac{|S(m, k)|^2}{\exp_q \left[\frac{1}{M} \sum_{m=1}^M \mathbf{s}_q(m, k) \right]}. \quad (5.15)$$

For $q = 1$, Eq. (5.15) is the same as Eq. (5.13). By taking the q -log of Eq.(5.15) we obtain:

$$\tilde{\mathbf{s}}_q(m, k) = \frac{\mathbf{s}_q(m, k) - \bar{\mathbf{s}}_q(k)}{1 + (1 - q)\bar{\mathbf{s}}_q(k)}. \quad (5.16)$$

Eq. (5.16) is q -LSMN formula. It is identical with LSMN when $q = 1$.

We investigate the effect of q -LSMN on the cross-term. For readability, we drop the subscripts m and k from now on. We denote:

$$Z = |S|^2 \chi. \quad (5.17)$$

We can rewrite Eq. (5.6) as follows:

$$|\hat{X}|^2 = Z|H|^2. \quad (5.18)$$

By taking the q -log of Eq. (5.18), we obtain:

$$\hat{\mathbf{x}}_q = \mathbf{z}_q + \mathbf{h}_q + (1 - q)\mathbf{z}_q\mathbf{h}_q. \quad (5.19)$$

where $\hat{\mathbf{x}}_q$, $\mathbf{z}_q$, and $\mathbf{h}_q$ are the q -log of $|\hat{X}|^2$, Z , and $|H|^2$ respectively. $\hat{\mathbf{x}}_q$ is the estimate of $\mathbf{x}_q$ after spectral subtraction. By normalizing the spectra with its long term average according to Eq. (5.16), and assuming convolutive noise is stationary, we obtain:

$$\begin{aligned} \tilde{\mathbf{x}}_q &= \frac{(\mathbf{z}_q - \bar{\mathbf{z}}_q)(1 + (1 - q)\mathbf{h}_q)}{(1 + (1 - q)\bar{\mathbf{z}}_q)(1 + (1 - q)\mathbf{h}_q)} \\ &= \frac{\mathbf{z}_q - \bar{\mathbf{z}}_q}{1 + (1 - q)\bar{\mathbf{z}}_q} \\ &= \tilde{\mathbf{z}}_q, \end{aligned} \quad (5.20)$$

where:

$$\bar{\mathbf{z}}_q = \frac{1}{M} \sum_{m=1}^M \mathbf{z}_{q,m}. \quad (5.21)$$

From Eq. (5.20) it is clear that q -LSMN is robust against convolutive noise. By taking the q -log of Eq. (5.17) we obtain:

$$\mathbf{z}_q = \mathbf{s}_q + \boldsymbol{\chi}_q + (1 - q)\mathbf{s}_q\boldsymbol{\chi}_q, \quad (5.22)$$

where $\boldsymbol{\chi}_q$ is the q -log of χ . Since $\mathbb{E}\{\cos \theta\} = 0$, then $\mathbb{E}\{\chi\} = 0$. While $\boldsymbol{\chi}_q$ is obtained by a nonlinear transformation from χ , we find that $\mathbb{E}\{\boldsymbol{\chi}_q\}$ is very close to zero for $0 < q < 1$. Therefore, the long term average of $\mathbf{z}_q$ equals:

$$\bar{\mathbf{z}}_q = \bar{\mathbf{s}}_q. \quad (5.23)$$

By substituting (5.22) and (5.23) to (5.20), we obtain:

$$\begin{aligned} \tilde{\mathbf{x}}_q &= \frac{\mathbf{s}_q + \boldsymbol{\chi}_q + (1 - q)\mathbf{s}_q\boldsymbol{\chi}_q - \bar{\mathbf{s}}_q}{1 + (1 - q)\bar{\mathbf{s}}_q} \\ &= \frac{(\mathbf{s}_q - \bar{\mathbf{s}}_q) + \boldsymbol{\chi}_q(1 + (1 - q)\mathbf{s}_q)}{1 + (1 - q)\bar{\mathbf{s}}_q} \\ &= \tilde{\mathbf{s}}_q + \frac{1 + (1 - q)\mathbf{s}_q}{1 + (1 - q)\bar{\mathbf{s}}_q} \boldsymbol{\chi}_q. \end{aligned} \quad (5.24)$$

From (5.24), we can see that normalizing in the q -log spectral domain introduces a weight to $\boldsymbol{\chi}_q$. For the spectral valleys, i.e. $\mathbf{s}_q < \bar{\mathbf{s}}_q$, this weight is smaller than 1, hence the cross-term is smaller. On the other hand, for the spectral peaks, i.e. $\mathbf{s}_q > \bar{\mathbf{s}}_q$, the weight is larger than 1. Removing the cross-term in the spectral valleys may be more beneficial for speech recognition than that in the spectral peaks. In the spectral valleys, noise and the cross-term are more dominant. Hence, the SNR is lower. Meanwhile, speech is more dominant in the spectral peaks. It could mask noise and the cross-term [Schroeder et al., 1979] and hence reduce their effect.

The use of cube root [Hermansky, 1990] and power-law functions [Kim and Stern, 2009] to replace the log function has been investigated previously. These can be seen as special cases of our framework. The cube-root method corresponds to q -LSMN at $q = 0.66$ and the power-law method corresponds to q -LSMN at $q = 0.9$. Our method

can be seen as an extension of these methods to various q -values.

5.5 Spectral Subtraction

Since q -LSMN only removes convolutive noise and the cross-term, we also implement spectral subtraction (SS) to remove additive noise. In this chapter, we implement nonlinear spectral subtraction which is formulated as follows:

$$|\hat{X}(m, k)|^2 = \max(|Y(m, k)|^2 - \alpha_m |\hat{N}(m, k)|^2, \beta |Y(m, k)|^2), \quad (5.25)$$

where $|\hat{X}(m, k)|^2$ is the estimate of filtered speech spectrum, $|\hat{N}(m, k)|^2$ is the estimated noise spectrum and α_m and β are the control parameters. The parameter α_m is set to be a function of the noisy signal-to-noise ratio (NSNR) for frame m with the following relation [Berouti et al., 1979]:

$$\alpha_m = \begin{cases} 1 & \text{if NSNR}_m \geq 20\text{dB}, \\ 4 - \frac{3}{20}\text{NSNR}_m & \text{if } -5\text{dB} \leq \text{NSNR}_m < 20\text{dB}, \\ 4.75 & \text{if NSNR}_m < -5\text{dB}, \end{cases} \quad (5.26)$$

where NSNR_m is formulated as:

$$\text{NSNR}_m = 10 \log \frac{\sum_{k=1}^K |Y(m, k)|^2}{\sum_{k=1}^K |\hat{N}(m, k)|^2}. \quad (5.27)$$

We set the spectral floor parameter, β , to 0.1.

For estimation of the noise spectrum, we implement the minima tracking algorithm [Doblinger, 1995]. First, we find $|\tilde{N}|^2$, the pre-estimation of noise power spectrum:

$$\begin{aligned} |\tilde{N}(m, k)|^2 = & \gamma |\hat{N}(m-1, k)|^2 \\ & + \frac{1-\gamma}{1-\lambda} \left(|\ddot{Y}(m, k)|^2 - |\ddot{Y}(m-1, k)|^2 \right), \end{aligned} \quad (5.28)$$

where $|\ddot{Y}(m, k)|^2$ is the smoothed noisy power spectrum that is calculated as follows:

$$|\ddot{Y}(m, k)|^2 = \delta |\ddot{Y}(m-1, k)|^2 + (1 - \delta) |Y(m, k)|^2. \quad (5.29)$$

We use the values $\gamma = 0.998$, $\lambda = 0.96$ and $\delta = 0.9$ in this paper.

Since noise is often nonstationary, it is important to keep updating the noise spectrum. The decision when to update the noise spectrum is based on a simple voice activity detector (VAD) algorithm proposed by [Hirsch and Ehrlicher, 1995]. It uses the ratio of the noisy spectrum and the noise spectrum:

$$\xi_{\text{rel}}(m, k) = \frac{\xi(m, k) - \xi_{\min}(m, k)}{\xi_{\max}(m, k) - \xi_{\min}(m, k)}, \quad (5.30)$$

where $\xi(m, k) = \frac{|\hat{N}(m, k)|^2}{|Y(m, k)|^2}$. The value of $\xi_{\min}(m, k)$ and $\xi_{\max}(m, k)$ are determined from 20 previous successive frames. The updating rules are:

$$|\hat{N}(m, k)|^2 = \begin{cases} |\hat{N}(m-1, k)|^2 & \text{if } \xi_{\text{rel}}(m, k) < T, \\ |\tilde{N}(m, k)|^2 & \text{else,} \end{cases} \quad (5.31)$$

where T is a threshold. We set T to 0.15.

5.6 Experimental Setup

5.6.1 Databases

Our proposed method was evaluated using two databases, Aurora-2 [Hirsch and Pearce, 2000] and CENSREC-2 [Nakamura et al., 2006]. Both databases are designed for concatenated digit recognition evaluation. These databases have different environmental settings. While noise was added artificially in Aurora-2, CENSREC-2 was recorded in real car driving environments.

5.6.1.1 The Aurora-2 database

In this database, eight types of noise: subway, babble, car, exhibition hall, restaurant, street, airport and train station, were added to clean speech artificially. There are two training conditions: the clean-condition training and the multi-condition training. For the clean condition training, only clean speech is used for training. For the multi-condition training data corrupted with four types of noise: subway, babble, car and

exhibition hall, at SNRs of 20 dB, 15 dB, 10 dB and 5 dB are used. We used only the clean-condition training in this study. For testing, this database provides three test sets: A, B and C. In Test Set A, the same noise as in the multi-condition training were added to clean speech data. Test Set B uses the same utterances as Test Set A but with different added noise (restaurant, street, airport and train station). Both Test Sets A and B use G.712 channel characteristics. In Test Set C, MIRS channel characteristic is used with subway and street noise. For all test sets, noise was added at SNRs of 20 dB, 15 dB, 5 dB, 0 dB and -5 dB.

5.6.1.2 The CENSREC-2 database

The Aurora-2 database is a popular database to evaluate robust speech recognition methods, it however has a major shortcoming. This database does not reflect the real situation conditions even though noises used in this database are real noise. In the real situations, the speakers tend to modify the way they speak in noisy environment. This phenomenon is called the Lombard effect. For this reason, we also used the CENSREC-2 database.

CENSREC-2 is a Japanese spoken database. It has 11 environmental conditions: the combinations of three vehicle speeds (idling, low-speed driving on city streets and high speed driving on expressways) and four kinds of in-car environments (normal, with air conditioner on, with audio CD player on, and with windows open). There were two types of microphones used: HF (hands free) and CT (close talking).

For the evaluation, the CENSREC-2 database provides four evaluation conditions. For Condition 1, the speech data for training and testing were recorded in the same environment and using the same microphone (HF). For Condition 2, the training data and test data were recorded with different environments: idling condition for training and low speed and high speed conditions for testing. Both testing and training data were recorded using HF microphones. In this condition, the main cause of mismatch is additive noise. For Condition 3, the training and testing data were recorded in the same environments, but using different microphones. CT microphones were used for recording the training data and HF microphones for the testing data. In this condition, channel mismatch is the main cause of the performance degradation. Lastly, for Condition 4, the training and test data were recorded using different environments and different microphones. The idling environment and CT microphones were used for

training data, whereas for testing, the data was recorded in the low and high speed conditions and the HF microphones were used. In this condition, both channel and additive noise are the sources of mismatch.

5.6.2 ASR configuration

For extracting the features, we applied 23 triangle mel-filterbanks and extracted 12 MFCC's coefficients and log energy as the static features. For recognition, 38 dimensional MFCC features were used. They consist of 12 static MFCC features, their first and second-order derivatives, the first and second order derivatives of log energy. We excluded the log energy from the features.

We implemented an HMM-based speech recognition system. Each digit was modeled by a left-to-right HMM with 16 states. Each state has three Gaussian mixtures. Two pause models were used: "sil" and "sp". The "sil" model consisted of three states. Each state in the "sil" model has six mixtures. The "sp" model consisted of a single state which was tied to the middle state of the "sil" model.

We measured the recognition performance with the word accuracy rate (%). For the Aurora-2 database, the average accuracy denotes the average over SNR 0dB to 20dB, while for CENSREC-2, the average accuracy denotes the average over all four evaluation conditions.

5.7 Experimental results and discussions

We first investigate the effect of q -LSMN on noisy speech. We examine whether q -LSMN can reduce the non-additive term in noisy speech. Then, we present our evaluation results of q -LSMN on the speech recognition performance.

5.7.1 The effect of q -LSMN on noisy speech

We investigated the nature of our proposed q -LSMN under the condition that the additive noise was known (artificially added). We filtered 1001 clean speech utterances of the Aurora-2 database (from first set of the test set A) through three types of communication channel, G.712, MIRS, and IRS. Then, q -LSMN was performed on both clean speech and filtered speech and the log spectral distance (LSD) between

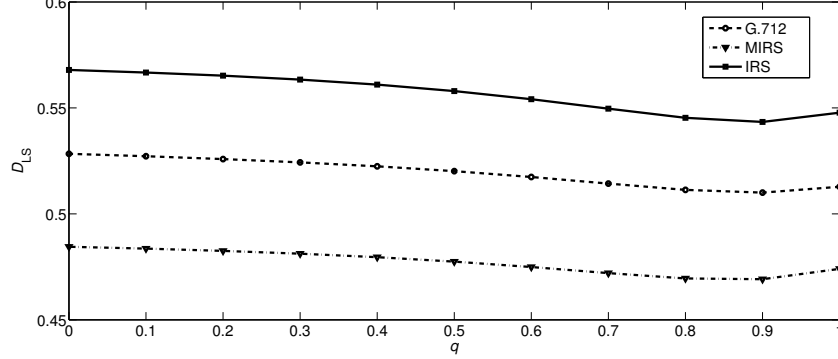


Figure 5.2: The log spectral distance between “filtered speech”, speech affected by convolutive noise, and the clean speech after applying q -LSMN.

them was calculated:

$$D_{LS} = \frac{1}{KM} \sum_{m=1}^M \sum_{k=1}^K \sqrt{\left(\log \frac{|S(m,k)|^2}{|\hat{S}(m,k)|^2} \right)^2}. \quad (5.32)$$

Fig 5.2 shows the log spectral distance between clean speech and filtered speech after q -LSMN. The LSD was smaller for some $q \neq 1$. These results indicated that speech and convolutive noise might be non-additive in the log spectral domain even without additive noise.

We added to the filtered speech (G.712 and MIRS) two types of additive noise from JEIDA noise database [Itahashi, 1990], babble noise and white noise. Noise was added from -5dB to 20dB with 5 dB increments. To obtain the clean speech estimate, we first carried out spectral subtraction as in Eq. 5.25 with $\alpha_m = 1$ and $\beta = 0.01$. We used “the true spectrum” of noise, artificially added to the clean speech. We next normalized the spectra obtained from spectral subtraction with their long term average. The resulting spectra can be assumed as the clean speech and the cross-term as shown in Eq. (5.12). We applied q -LSMN on the normalized spectra. We then calculate the LSD between the estimated clean speech spectra after q -LSMN and the normalized clean spectra.

We conducted a recognition experiment using the filtered speech, i.e. speech contaminated with convolutive noise only. We applied q -LSMN using $0 \leq q \leq 1$. The results are shown in Fig 5.5. We found that the performance of speech recognition im-

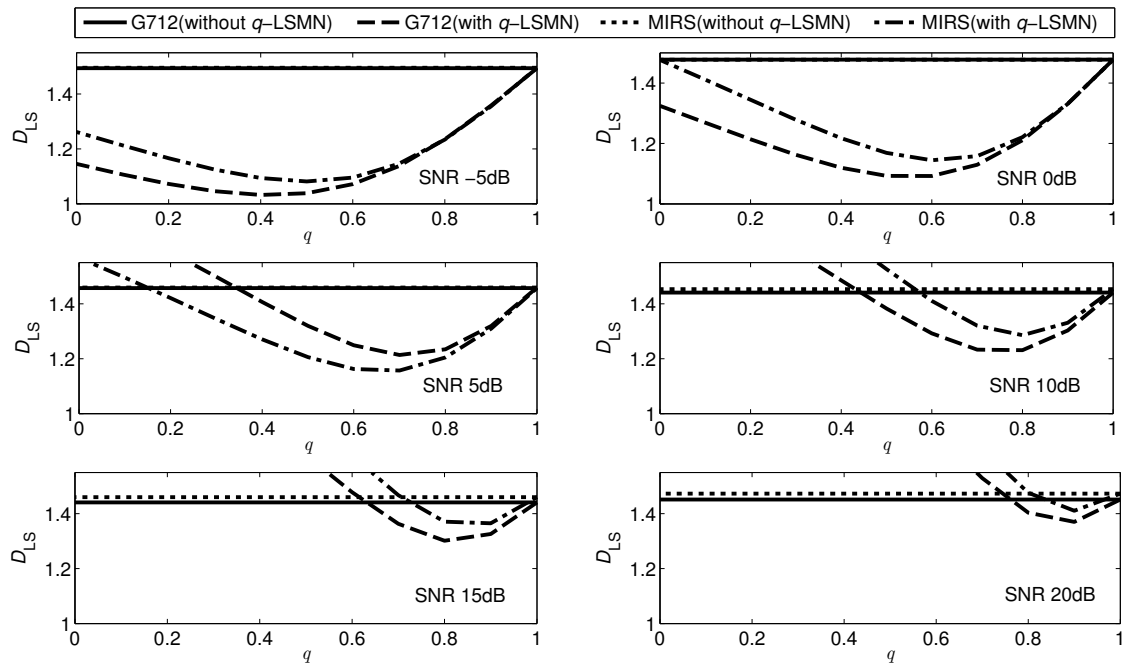


Figure 5.3: The log spectral distance between the clean speech estimate and the clean speech with and without applying q -LSMN for the spectral valleys. The clean speech estimate is obtained after employing spectral subtraction and normalization on the noisy speech spectra. In this figure, babble noise is used as additive noise.

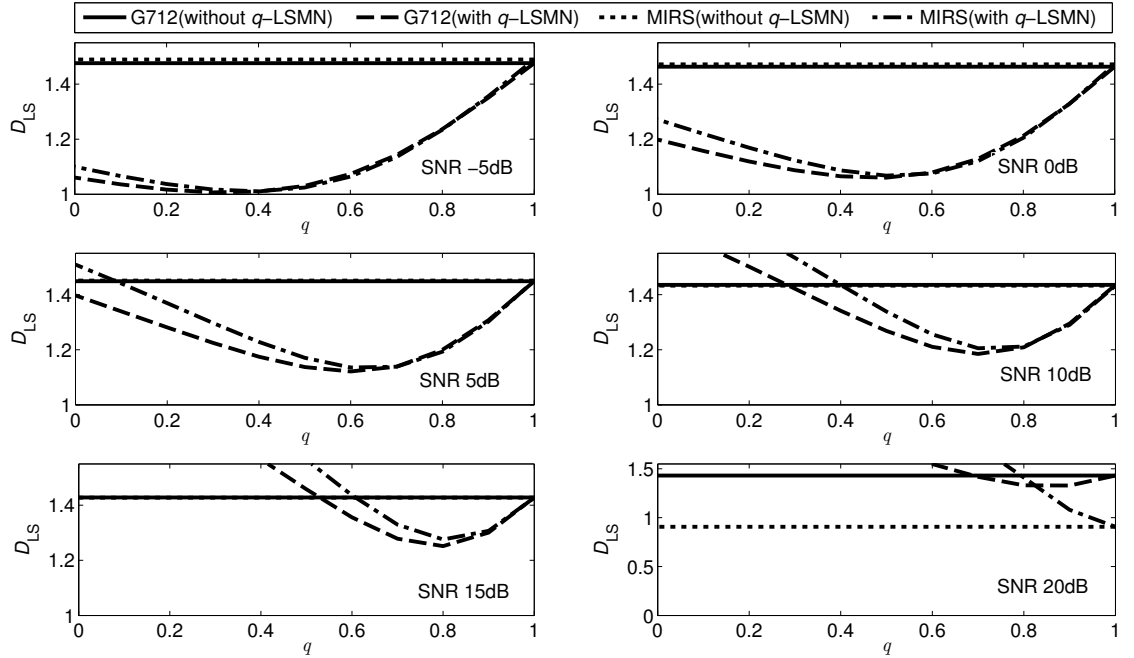


Figure 5.4: The log spectral distance between the clean speech estimate and the clean speech with and without applying q -LSMN for the spectral valleys. The clean speech estimate is obtained after employing spectral subtraction and normalization on the noisy speech spectra. In this figure, white noise is used as additive noise.

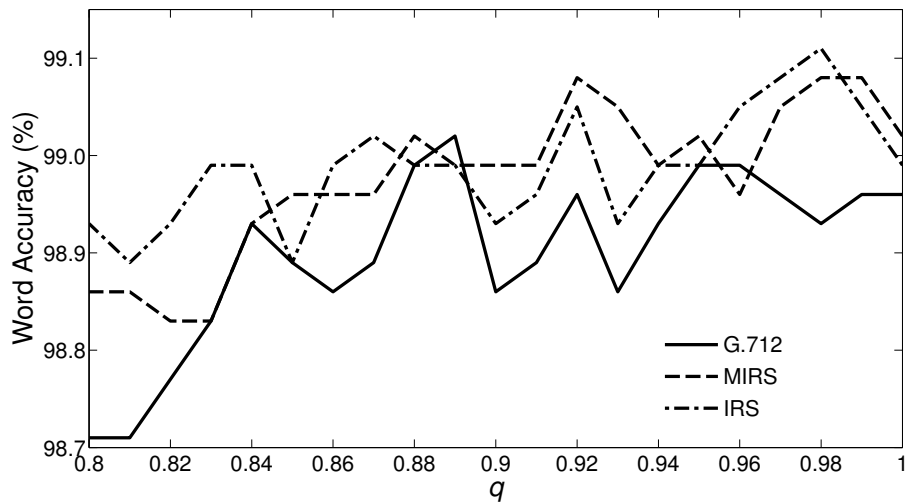


Figure 5.5: The word accuracies (%) of q -LSMN for filtered speech, i.e speech contaminated with convolutive noise only.

Table 5.1: Word accuracy (%) of the Aurora-2 task when the cross-term is assumed to be zero.

SNR (dB)	G.712		MIRS	
	Babble	White	Babble	White
Clean	98.6	98.6	98.6	98.6
20	97.8	97.6	97.8	97.7
15	97.5	97.7	97.6	97.6
10	97.6	97.1	97.1	97.8
5	97.1	97.4	97.2	97.9
0	96.9	96.9	96.4	97.5
-5	95.2	96.1	94.3	96.3

proved for some $q \neq 1$. These results suggest that speech and convolutive noise are not additive in the log spectral domain as they are usually assumed, while the differences in accuracy are marginal.

Figures 5.3 and 5.4 show the LSD between the estimated clean speech spectra and the clean speech spectra with and without q -LSMN for the spectral valleys. We select the spectral valleys by choosing the spectra that satisfy $\mathbf{s}_q < \bar{\mathbf{s}}_q$ for each utterance. The LSD between clean speech and its estimate is smaller with q -LSMN than without it for some $q \neq 1$. We obtained 0.2 of the LSD improvement in average over all SNR conditions. These results indicate the effect of q -LSMN on the cross-term.

Table 5.1 shows the recognition results using the clean speech estimate. The accuracies were lower than that for the clean speech. This result indicated that there was still a mismatch even though the additive noise and convolutive noise spectra were removed. From Eq. (5.4), it is obvious that the cross-term is the source of this mismatch. This result agrees with the previous studies [Deng et al., 2004; Faubel et al., 2008; Zhu and Alwan, 2002].

5.7.2 Evaluation of q -LSMN without spectral subtraction

Figure 5.6 shows the average word accuracies over test set A, B, and C of the Aurora 2 database when q is varied from 0.0 to 1.0. For certain values of $q \neq 1$, q -LSMN was better than $q = 1$, the case when q -LSMN became identical with LSMN. The best accuracy was achieved at $q = 0.7$, with 21.9% error reduction rate over LSMN.

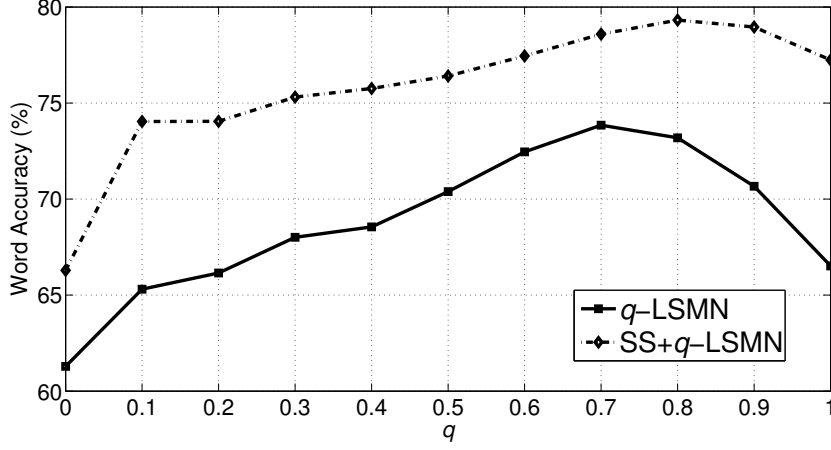


Figure 5.6: The word accuracies (%) of q -LSMN with and without spectral subtraction for the Aurora-2 task for different q values.

Figure 5.7 shows that the optimum q is different for each SNR condition of the Aurora 2 database. The optimum q value was closer to 1 for the high SNR and became smaller as SNR became lower except for -5 dB SNR. This result is consistent with the fact that the cross-term became larger when SNR was lower. These results coincided with the improvement in the LSD on the spectral valleys in Section 5.7.1. Fig. 5.8 shows the optimum q for four types of noise in test set A of the Aurora-2 database for each SNR condition. This result suggests the optimum q is influenced by the noise types. Noises in real environments affect each spectral band differently. As a result, the cross-terms are different among different kinds of noise, and different q is required to compensate for it.

Figure 5.9 shows the average word accuracies over Conditions 1, 2, 3, and 4 of the CENSREC-2 database. The best performance was achieved at $q = 0.4$. This value was different from that in the Aurora-2 database. In real noisy environment, additive noise may have various different sources and their transmission channels may be largely different. Hence, a lower q is required to compensate for the cross-term.

Figure 5.10 shows the performance of q -LSMN for each evaluation condition of the CENSREC-2 database. The word accuracies were largely improved except in Condition 1. In Condition 1, training and testing conditions are the same. Hence, there should be only a small mismatch, and the performance improves only slightly from LSMN

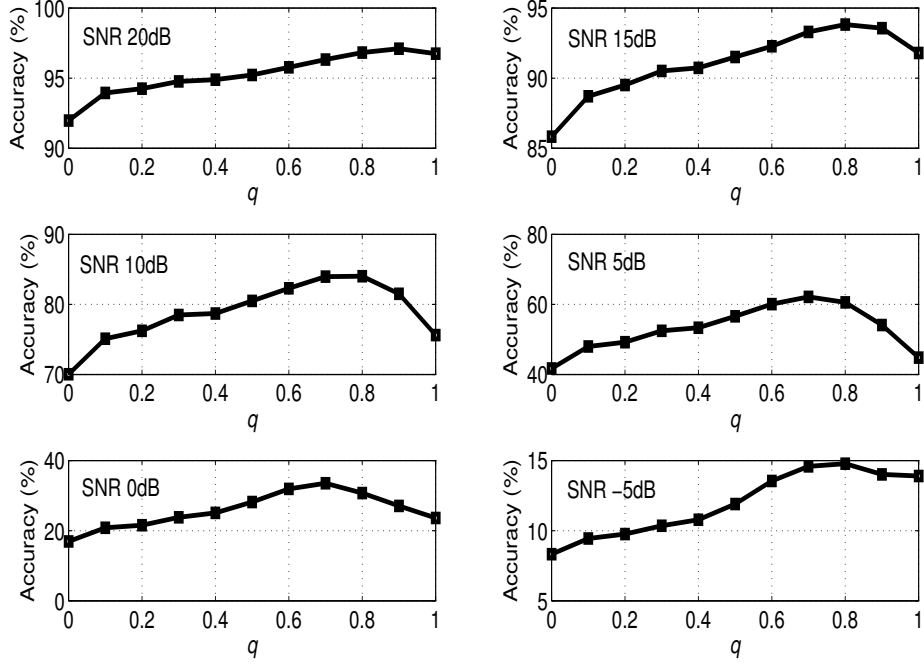


Figure 5.7: The word accuracies (%) of q -LSMN for various SNR conditions in the Aurora-2 task.

($q = 1$). Larger improvements were obtained in Condition 3 and 4 where there are channel differences between training and testing.

Tables 5.2 and 5.3 summarize the recognition results of q -LSMN and the other conventional normalization methods. For both databases, the performance of q -LSMN was consistently better than that of CMN and MVN. It reduced errors by 20.1% and 18.2% respectively when $q = 0.7$ for Aurora-2, and by 38.5% and 27.2% respectively when $q = 0.4$ for CENSREC-2. These results confirm that q -LSMN improves the robustness of ASR systems.

5.7.3 Evaluation of q -LSMN with spectral subtraction

The combination of q -LSMN with spectral subtraction further improves the performance of the front-end as expected (Fig. 5.6 and 5.9). For both databases, the optimum q -value was different from that without SS. When SS was implemented, $q = 0.8$ was

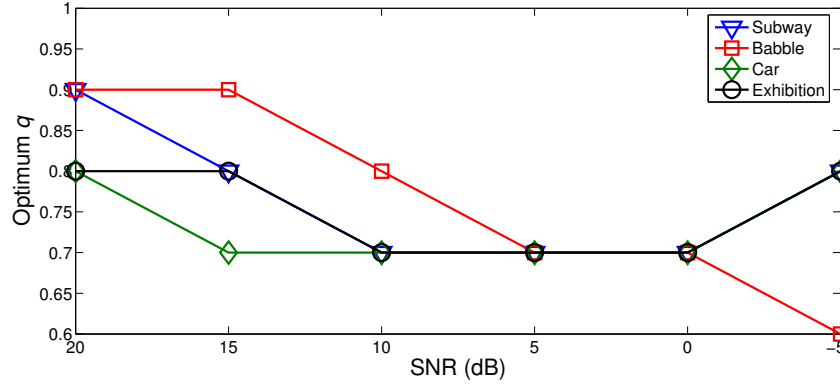


Figure 5.8: The optimum q -values for four types of noise in Test Set A of the Aurora-2 database for various SNR conditions.

Table 5.2: The word accuracy (%) of the Aurora-2 task.

Methods	Set A	Set B	Set C	Ave.
No compensation	65.8	68.6	60.9	65.9
q -LSMN ($q = 0.7$)	72.2	76.9	71.1	73.8
LSMN	64.7	70.2	62.8	66.5
CMN	65.7	70.5	63.9	67.3
MVN	68.3	69.3	64.9	68.0

Table 5.3: The word accuracy (%) of the CENSREC-2 task.

Methods	Cond 1	Cond 2	Cond 3	Cond 4	Ave.
No compensation	85.5	78.7	46.2	40.4	62.7
q -LSMN ($q = 0.4$)	88.6	85.2	73.5	63.3	77.7
LSMN	88.3	83.3	64.4	48.7	71.1
CMN	85.9	77.7	52.9	38.9	63.6
MVN	85.5	82.5	63.2	46.0	69.3

the optimum for Aurora-2, whereas $q = 0.5$ was the optimum for CENSREC-2. The cross-term becomes smaller as SNR becomes higher. This may be the reason that the optimum q is smaller after SS.

Tables 5.4 and 5.5 summarize the recognition results of q -LSMN and the other conventional normalization methods after SS. For the Aurora-2 database, q -LSMN reduced errors by 10.4% and 21.1% from CMN and MVN respectively at $q = 0.8$. For

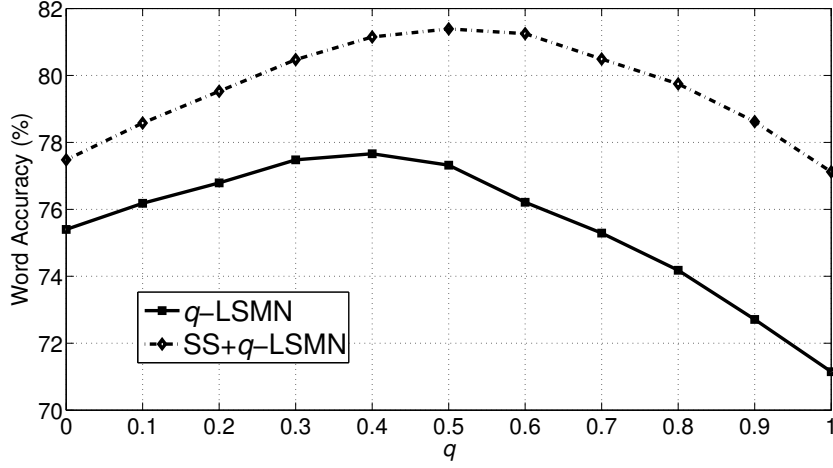


Figure 5.9: The word accuracies (%) of q -LSMN with and without spectral subtraction for CENSREC-2 task with different q values.

Table 5.4: The word accuracy (%) of q -LSMN and the other methods after spectral subtraction for the Aurora-2 task.

Methods	Set A	Set B	Set C	Ave.
SS	70.0	76.7	75.9	77.5
SS + q -LSMN ($q = 0.8$)	78.5	80.7	78.1	79.3
SS + LSMN	76.2	79.3	75.2	77.2
SS + CMN	76.2	78.4	75.2	76.9
SS + MVN	74.3	74.3	71.8	73.8
ETSI AFE	80.1	82.1	79.5	80.8

CENSREC-2, q -LSMN reduced errors by 41.1% and 35.3% from CMN and MVN respectively at $q = 0.5$. We also compared our method with ETSI AFE. ETSI AFE performed better than our method on the Aurora-2 database. However, q -LSMN achieved better word accuracies on the CENSREC-2 database. q -LSMN alone reduced errors by 26.1% from ETSI AFE. Combination with SS reduced errors by 38.4% from ETSI AFE.

In ETSI AFE, blind equalization (BE) is used to remove convolutive noise. In BE, the bias of each frame, which will be subtracted from the features, is calculated using pre-determined reference means and weighting factors which depend on the log of the energy of the frame. The reference means used in ETSI AFE correspond to flat spectra. In CENSREC-2, channel mismatch is caused not only by the difference of microphones

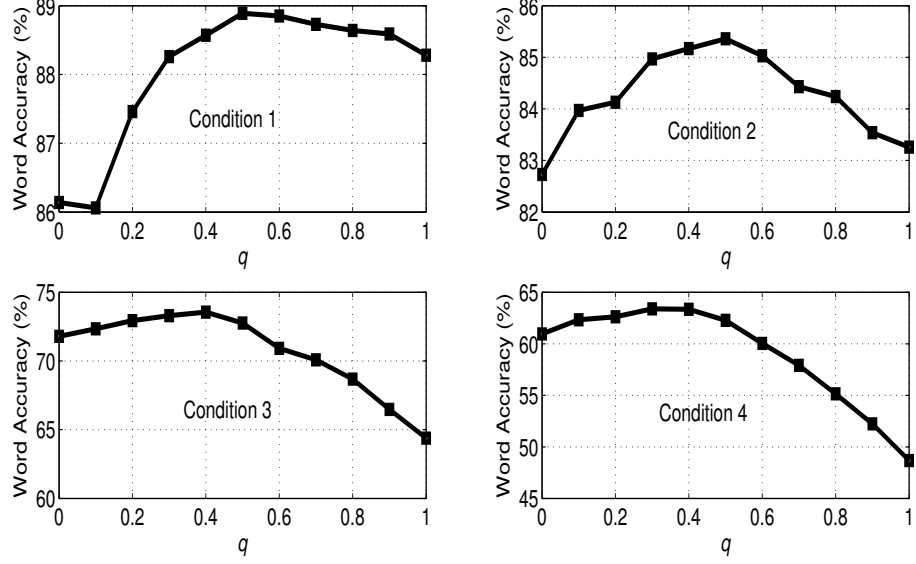


Figure 5.10: The word accuracies (%) of q -LSMN for CENSREC-2 task for each condition evaluation with different q values.

Table 5.5: The word accuracy (%) of q -LSMN and the other methods after spectral subtraction for the CENSREC-2 task.

Methods	Cond 1	Cond 2	Cond 3	Cond 4	Ave.
SS	86.7	79.9	53.2	47.0	66.7
SS + q -LSMN ($q = 0.5$)	89.2	86.1	78.5	71.6	81.4
SS + LSMN	88.9	84.7	72.7	62.2	77.1
SS + CMN	86.7	79.7	59.5	47.6	68.4
SS + MVN	86.7	83.0	65.9	51.8	71.8
ETSI AFE	85.6	80.3	59.9	53.3	69.8

types used in training and testing, but also by the change of distance between the speakers and the microphones, which would cause the reference and actual means to be different. Hence, BE may not suitable for the recording conditions in CENSREC-2.

5.8 Conclusion

We propose q -LSMN, a feature normalization method in the q -log spectral domain. The use of the q -log function enables us to represent nonadditivity of speech features,

which exists when noise and speech are correlated. Our evaluation using two types of databases, Aurora-2 and CENSREC-2 databases, showed the effectiveness of our approach. q -LSMN was better than CMN and MVN in both databases. Our method outperformed ETSI AFE on the CENSREC-2 database, where we gained up to 26.1% relative improvement. This result may indicate that our method is more effective in real environments. The combination of spectral subtraction with q -LSMN was also better than the combination of spectral subtraction with CMN or MVN. We believe that our method can be used complementary to any noise removal methods other than spectral subtraction.

We should say that our non-extensive approach can not provide a clear solution to the problems of non-extensivity of the noisy speech features, but does provide an alternative approach to them. In some $q \neq 1$, it provides a better solution than the original extensive ones such as CMN and LSMN from the viewpoint of efficiency and/or performance.

Chapter 6

Q -Gaussian based spectral subtraction

6.1 Introduction

Spectral subtraction (SS) is perhaps one of the most popular methods to remove additive background noise. In spectral subtraction, the clean speech spectrum estimate is obtained by simply subtracting the noisy spectrum with the noise spectrum estimate. From the statistical sense, spectral subtraction is a variance estimator which is derived in the extensive framework. In this framework, it is assumed that the speech and noise spectra follow Gaussian distributions and are uncorrelated. Hence, the noisy speech spectra also follow Gaussian distributions. By maximizing the likelihood of the noisy speech distribution, the spectral subtraction formula can be derived [McAulay and Malpass, 1980]. Spectral subtraction is usually combined with feature normalization methods to remove both additive and convolutive noise [Kitaoka et al., 2001; Yoma et al., 1998].

The Gaussian distribution however is not the best distribution to represent the speech spectrum due to the nature of a speech pattern which is *complex*. The length of the window used in speech processing is often smaller than various correlations between the spectral components of a speech pattern. For this reason, while speech may approximate the Gaussian distribution when a very long window is employed, the short-time speech spectra do not follow the Gaussian distributions [Wilbur B. Davenport, 1952] but show heavy-tailed distributions instead. Unfortunately, a long window cannot be

used in speech processing because we may lose the time resolution of the signal. Several other distributions have been used instead of Gaussian distributions, for instance: Laplace [Chen and Loizou, 2007] and Gamma [Martin, 2002, 2005] distributions. They have shown to fit speech spectrum better than the Gaussian distribution does. When speech is contaminated with noise, the short-time spectra of noisy speech are likely to follow heavy-tailed distributions as well and not Gaussian distributions. Therefore, it is not surprising that spectral subtraction has some limitation [Evans et al., 2006] and may not give a sufficiently high performance. A weighting factor is often introduced to improve its performance as in NSS. However, this factor is decided heuristically.

Our focus in this chapter is how to deal with the fact that noisy speech does not follow a Gaussian distribution. A new family of distributions has been derived from the theory of non-extensive statistics. These distributions, as explained in Chapter 4, maximize Tsallis entropy under appropriate constraints. The q -Gaussian distribution is one of them. It is a heavy-tailed distribution when $q > 1$. For this reason, it has been employed to model many complex systems. Their examples are stock market [Gradojevic and Gencay, 2008, 2011; Kozaki and Sato, 2008], DNA molecules [Moreira et al., 2008], earthquakes [Caruso et al., 2007], and dynamical systems [Moyano and Anteneodo, 2006; Tirnakli et al., 2007].

In this chapter, we propose a new spectral subtraction by assuming noisy speech follows a q -Gaussian distribution. By doing so, we can employ a heavy-tailed distribution to model noisy speech distribution. We then derive a spectral subtraction method in a similar way as spectral subtraction does. We call our method q -spectral subtraction (q -SS).

The remaining of this chapter is organized as follows. In Section 6.2, we derive the spectral subtraction formula and explain some of its variants. In Section 6.3, our proposed method, q -ss, is explained. We also derive a way to optimize the parameter in q -SS in this section. In Section 6.4, we explain our experimental setup for evaluating q -SS. In Section 6.5, we evaluate our method, q -SS, on the speech recognition tasks and discuss the results. In Section 6.6, we conclude this chapter.

6.2 Spectral subtraction

6.2.1 Derivation

In this section, we explain how to derive the spectral subtraction formula in more detail. Spectral subtraction is derived as follows. Let $y(t)$ denote noisy speech consisting of clean speech $x(t)$ and additive noise $n(t)$. By taking the short-time fourier transform of the signals, we obtain their spectral representation. Consider a spectral component at frequency f . We assume a spectral component, X_f , of clean speech is a complex random variable that follows a Gaussian distribution with zero mean and variance $\sigma(f)$. Similarly, a spectral component of noise signal, N_f , is also a complex random variable that has a Gaussian distribution with zero mean and variance $\tau(f)$. We also assume that X_f and N_f are statistically independent, and hence, noisy speech, Y_f , also follows a zero mean Gaussian distribution with the probability density of Y_f is given by:

$$P(Y_f) = \frac{1}{\pi\nu(f)} \exp\left(-\frac{|Y_f|^2}{\nu(f)}\right), \quad (6.1)$$

where $\nu(f)$ is the variance of noisy speech. Since speech and noise are independent, $\nu(f) = \sigma(f) + \tau(f)$. We would like to find the estimate of the clean speech variance from an observation of $|Y_f|^2$ assuming $\tau(f)$ is known. By differentiating $P(Y_f)$ with respect to $\sigma(f)$ and equating it to zero, we obtain $\hat{\sigma}(f)$, the maximum likelihood estimation of $\sigma(f)$ as the following:

$$\hat{\sigma}(f) = |Y_f|^2 - \tau(f). \quad (6.2)$$

Let $|X_f|^2$ and $|N_f|^2$ be the observed power spectra of clean speech and noise respectively. For zero mean distributions, the variance of a distribution is the average of the squared of the spectrum. Therefore $|X_f|^2 = \sigma(f)$ and $|N_f|^2 = \tau(f)$. Eq. (6.2) becomes:

$$|\hat{X}(f)|^2 = |Y_f|^2 - |N_f|^2. \quad (6.3)$$

Eq. (6.3) is the power spectral subtraction formula. It maintains a linear relation between noisy speech, noise and clean speech. Therefore, it is also called linear spectral subtraction (LSS).

6.2.2 Nonlinear spectral subtraction (NSS)

The simplicity of spectral subtraction comes with a price. The inaccuracy of noise estimation causes distortions and information losses in speech. There have been many variants of spectral subtraction proposed to improve its performance. One popular variant of spectral subtraction is the one proposed by Berouti et al. [1979]. They introduce an oversubtraction factor, α , and the spectral subtraction formula becomes:

$$|\hat{X}_f|^2 = |Y_f|^2 - \alpha |N_f|^2. \quad (6.4)$$

Since the introduction of α makes the subtraction nonlinear, it is called nonlinear spectral subtraction (NSS). Zhu and Alwan [2002] reported that the use of α also compensates for nonlinear relation between noise and speech. Even though NSS has shown to improve the robustness of ASR better than LSS, the parameter α is determined heuristically. There exists no consistent ways to optimize α .

Many variants of NSS have been proposed [Kamath and Loizou, 2002b; Lockwood and Boudy, 1992; Udrea et al., 2008]. They are basically modifications of the Eq.(6.4). In this chapter, we determine α for each frequency bin, denoting α_f , using the following relation [Berouti et al., 1979]:

$$\alpha_f = \begin{cases} 1 & \text{if } \Phi_f \geq 20\text{dB}, \\ \alpha_0 - \frac{3}{20}\Phi_f & \text{if } -5\text{dB} \leq \Phi_f < 20\text{dB}, \\ 4.75 & \text{if } \Phi_f < -5\text{dB}. \end{cases} \quad (6.5)$$

Parameter α_0 is the desired value of α at 0 dB SNR. It is usually set between 4 and 6. In this chapter we use $\alpha_0 = 4$. Φ_f is the noisy signal to noise ratio (in dB), i.e. the *a posteriori* SNR. In this chapter, we calculated it for each frequency bin using the following formula:

$$\Phi_f = 10 \log \gamma_f, \quad (6.6)$$

where:

$$\gamma_f = \frac{|Y_f|^2}{|\hat{N}_f|^2}. \quad (6.7)$$

In this chapter, we denote NSS of which α_f is determined using Eq. (6.5) as NSS_n. When a single and constant α_f is used for all spectra, we denote it as NSS_c.

6.3 Q -Spectral subtraction (q -SS)

6.3.1 Derivation

When speech and noise are Gaussian random variables, noisy speech can still be a Gaussian even when speech and noise are correlated Chen and Shao [2004]. However, the short-time speech spectra are not likely to follow Gaussian distributions Gazor and Zhang [2003]; Wilbur B. Davenport [1952]. Therefore, the distribution of noisy speech is likely not a Gaussian, even when speech and noise are independent. For this reason, we assume that noisy speech follows the q -Gaussian distribution, which has heavy-tailed. Theoretically, the q -Gaussian distributions can emerge from either the sum of correlated random variables Thistleton et al. [2009]; Umarov et al. [2008]; Vignat and Plastino [2007] or the sum of independent N q -Gaussian random variables for small number of N Tsallis and Duarte Queirós [2007]. It has also been shown that the long term behavior of a locally stationary system that follows a Gamma distribution exhibits a q -Gaussian distribution Beck and Cohen [2003]; Touchette and Beck [2005].

The q -spectral subtraction (q -SS) formula is derived as follows. Consider a spectral component at frequency f . We assume speech and noise to be q -Gaussian and independent. Therefore, the spectral component of noisy speech follows the q -Gaussian distribution with variance $\nu_q(f) = \sigma_q(f) + \tau_q(f)$, where $\sigma_q(f)$ and $\tau_q(f)$ is the variance of speech and noise respectively. Let $Y_R = \text{Re}(Y_f)$ and $Y_I = \text{Im}(Y_f)$ be the real and imaginary parts of the speech spectrum respectively. Assuming both Y_R and Y_I follow q -Gaussian and are identically distributed with variance $\nu_q(f)/2$. Then, the probability density functions for Y_R and Y_I are as follow:

$$P_q(Y_R) = \frac{\sqrt{2}A_qB_q}{\sqrt{\nu_q(f)}} \exp_q \left(-\frac{2B_q^2|Y_R|^2}{\nu_q(f)} \right), \quad (6.8)$$

$$P_q(Y_I) = \frac{\sqrt{2}A_qB_q}{\sqrt{\nu_q(f)}} \exp_q \left(-\frac{2B_q^2|Y_I|^2}{\nu_q(f)} \right). \quad (6.9)$$

We assume that the real and imaginary part of each Y_f are independent since it was reported that their dependency was small in average Martin [2005]. Then, the distribution for noisy speech is formulated as follows:

$$P_q(Y_f) = \frac{2A_q^2 B_q^2}{\nu_q(f)} \exp_q \left(-\frac{2B_q^2 |Y_f|^2}{\nu_q(f)} \right). \quad (6.10)$$

Equation (6.10) is identical with Eq. (6.1) when $q = 1$. By differentiating $P_q(Y_f)$ with respect to $\sigma_q(f)$, and equating to zero, we obtain the maximum likelihood estimate, $\hat{\sigma}_q(f)$, as the following:

$$\hat{\sigma}_q(f) = \frac{2(2-q)}{3-q} |Y_f|^2 - \tau_q(f). \quad (6.11)$$

Since $\sigma_q(f) = |X_f|^2$ and $\tau_q(f) = |N_f|^2$, Eq. (6.11) becomes:

$$|\hat{X}_f|^2 = \frac{2(2-q)}{3-q} |Y_f|^2 - |N_f|^2. \quad (6.12)$$

Eq. (6.12) is the q -spectral subtraction (q -SS) formula. It is the same as LSS when $q = 1$.

6.3.2 Relation to nonlinear spectral subtraction

In this section, we relate q -SS with NSS in Eq. (6.4). Denoting $v(q) = \frac{2(2-q)}{3-q}$, we can rewrite Eq. (6.12) as follows:

$$|\hat{X}_f|^2 = v(q) |Y_f|^2 - |N_f|^2. \quad (6.13)$$

By dividing Eq. (6.13) with $v(q)$, we obtain:

$$\frac{1}{v(q)} |\hat{X}_f|^2 = |Y_f|^2 - \frac{1}{v(q)} |N_f|^2. \quad (6.14)$$

Since scaling does not affect the performance of speech recognition, we can relate α in Eq. (6.4) with Eq. (6.14) as follows:

$$\begin{aligned} \alpha &= \frac{1}{v(q)}, \\ &= \frac{3-q}{2(2-q)}. \end{aligned} \quad (6.15)$$

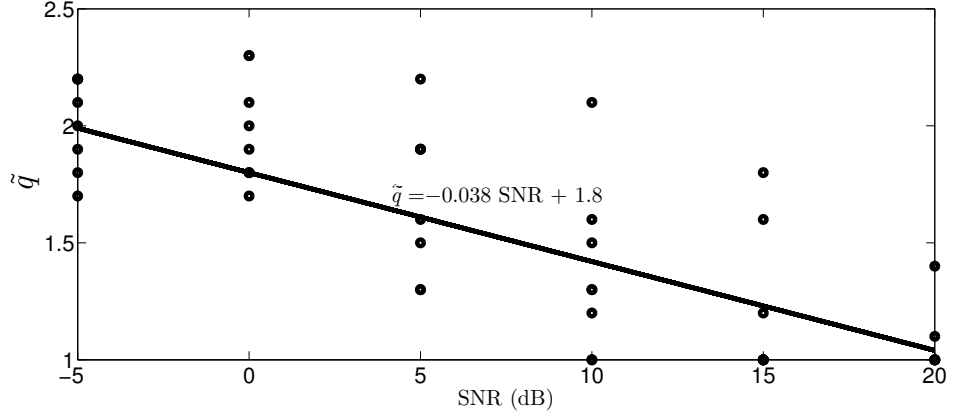


Figure 6.1: The scatter plot between $\tilde{q}$, the q -value that fits the empirical distribution of noisy speech based on the MMSE criterion, and the SNR condition. The solid line is the relation between $\tilde{q}$ and the SNR based on linear regression method.

Based on Eq. (6.15), α is positive when $q < 2$. It is infinity when $q = 2$. By this way, our method becomes identical with NSS. Our q -SS formulation also gives a consistent way to estimate the control parameter α in NSS.

6.3.3 Q -SS based on the optimum q -Gaussian distribution

In this section, we derive another way to determine q . The optimum q is estimated by finding the value of q of the q -Gaussian distribution that fits best the distribution of noisy speech based on the minimum mean square error (MMSE) criterion.

To estimate q , we first find the q -Gaussian distribution that fits the empirical distribution $S(Y_R)$ of noisy speech, where Y_R is the real part of DFT coefficients of Y and $S(Y_R)$ is the normalized histogram of Y_R whose total area is 1. We use the first 10 utterances from the test set A of the Aurora-2 database to obtain $S(Y_R)$. We estimate $S(Y_R)$ for each SNR conditions. It should be noted that the Aurora-2 database provides the *a priori* SNR, i.e. the ratio between clean speech and noise. This is different from Φ_f in Eq. (6.7) which is the *a posteriori* SNR. Based on the data, we obtain its variance and $S_i(Y_R)$ where $i = 1, 2, \dots, n$ are the center point of each histogram bin. The q -value that minimizes the sum of the mean squared error between the normalized histogram, $S_i(Y_R)$ and the corresponding q -Gaussian distribution, $P_{q_i}(Y_R)$, which we denote as $\tilde{q}$,

is selected using the following formula:

$$\tilde{q} = \underset{q}{\operatorname{argmin}} \frac{1}{n} \sum_{i=1}^n (S(Y_{R_i}) - P_q(Y_{R_i}))^2. \quad (6.16)$$

After obtaining $\tilde{q}$ for every SNR, we find the relation between SNR and $\tilde{q}$ using linear regression.

Figure 6.1 shows the scatter plot between the estimated q and the SNR conditions and the result of linear regression. We limit $q = 1.88$ when the SNR is low (lower than -5dB) and $q = 1$ when the SNR is high (higher than 20dB). These are the same limits used in NSS. The relation between the SNR and $\tilde{q}$ is formulated as:

$$\tilde{q} = \begin{cases} 1 & \text{if } \Psi_f(m) \geq 20\text{dB}, \\ -0.038\Psi_f(m) + 1.8 & \text{if } -5\text{dB} \leq \Psi_f(m) < 20\text{dB}, \\ 1.88 & \text{if } \Psi_f < -5\text{dB}, \end{cases} \quad (6.17)$$

where $\Psi_f(m)$ is the clean speech signal-to-noise ratio (in dB), i.e. the *a priori* SNR. It is different from Φ_f of Eq. (6.7), which is the ratio between noisy speech and noise. It is calculated using the following formula:

$$\Psi_f(m) = 10 \log \xi_f(m), \quad (6.18)$$

where:

$$\xi_f(m) = \frac{|X_f|^2}{|N_f|^2}. \quad (6.19)$$

In practice, $\xi_f(m)$ is unknown and needs to be estimated. We obtain $\hat{\xi}_f(m)$, the estimate of ξ_f , using the maximum likelihood method [Ephraim and Malah, 1984]. In the maximum likelihood method, $\xi_f(m)$ is estimated based on L past observations of the noisy speech spectra $\mathbf{Y}_f(m) = \{Y_f(m), Y_f(m-1), \dots, Y_f(m-L+1)\}$, assuming the noise variance, $\tau(f)$ is known. By assuming the statistical independence of the L

observations and using Gaussian model, we obtain the likelihood function:

$$p(\mathbf{Y}_f(m)|\sigma(f), \tau(f)) = \prod_{i=0}^{L-1} \frac{1}{\phi(\sigma(f) + \tau(f))} \exp\left(-\frac{Y_f^2(m-i)}{\sigma(f) + \tau(f)}\right), \quad (6.20)$$

where $\sigma(f)$ is the clean speech variance. By maximizing Eq. (6.20) with respect to $\sigma(f)$, we obtain:

$$\hat{\sigma}(f) = \max\left(\frac{1}{L} \sum_{i=0}^{L-1} Y_f^2(m-i) - \tau_f(m), 0\right). \quad (6.21)$$

By dividing both side of Eq. (6.21) by $\tau_f(m)$, we obtain:

$$\hat{\xi}_f(m) = \max\left[\frac{1}{L} \sum_{i=0}^{L-1} \gamma_f(m-i) - 1, 0\right], \quad (6.22)$$

where $\gamma_f(m)$ is the *a posteriori* SNR given in Eq. (6.7). In practice, Eq. (6.22) is replaced by recursive operation:

$$\hat{\xi}_f(m) = \max[\bar{\gamma}_f(m) - 1, 0], \quad (6.23)$$

where:

$$\bar{\gamma}_f(m) = a\bar{\gamma}_f(m-1) + (1-a)\frac{\gamma_f(m)}{b}, \quad (6.24)$$

In this paper, we apply $a = 0.725$ and $b = 2$. We denote q -SS that use the relation in Eq. (6.17) as q -SS_m. When a single and constant q is used, we denote it as q -SS_c.

6.3.4 Noise estimation

In spectral subtraction, it is assumed that the noise power spectra known. However, they are unknown in practice and need to be estimated. In this paper, we applied the minima tracking algorithm [Doblinger, 1995] to estimate noise. In this method, the

noise power spectrum is pre-estimated using the following formula:

$$|\tilde{N}_f(m)|^2 = \gamma |\hat{N}_f(m-1)|^2 + \frac{1-\gamma}{1-\lambda} \left(|\ddot{Y}_f(m)|^2 - |\ddot{Y}_f(m-1)|^2 \right), \quad (6.25)$$

where $|\ddot{Y}(m, k)|^2$ is the smoothed noisy power spectrum which is obtain using the following formula:

$$|\ddot{Y}_f(m)|^2 = \delta |\ddot{Y}_f(m-1)|^2 + (1-\delta) |Y_f(m)|^2. \quad (6.26)$$

We use the values $\gamma = 0.998$, $\lambda = 0.96$ and $\delta = 0.9$ in this paper.

Since noise is usually nonstationary, it is important to keep updating the noise spectrum. We implement a voice activity detector (VAD) algorithm proposed by Hirsch and Ehrlicher [1995]. In this method, the ratio of the noisy spectrum and the noise spectrum is used to determine when to update the noise spectra. It is calculated as follows:

$$\zeta_f^{\text{rel}}(m) = \frac{\zeta_f(m) - \zeta_f^{\min}(m)}{\zeta_f^{\max}(m) - \zeta_f^{\min}(m)}, \quad (6.27)$$

where $\zeta_f(m) = \frac{|\hat{N}_f(m)|^2}{|Y_f(m)|^2}$. The value of $\zeta_f^{\min}(m)$ and $\zeta_f^{\max}(m)$ are determined from 20 previous successive frames. The updating rules are:

$$|\hat{N}(m, k)|^2 = \begin{cases} |\hat{N}_f(m-1)|^2 & \text{if } \zeta_f^{\text{rel}}(m) < T, \\ |\tilde{N}_f(m)|^2 & \text{else,} \end{cases} \quad (6.28)$$

where T is a threshold. We set T to 0.15.

6.3.5 Flooring

Due to inaccuracies in the estimation of noise spectrum, the power spectrum estimate of clean speech, $|\hat{X}_f|^2$, could be negative. To avoid this, a flooring rule is usually applied.

$$|\hat{X}_f|^2 = \beta |Y_f|^2 \quad \text{if } |\hat{X}_f|^2 < \beta |Y_f|^2. \quad (6.29)$$

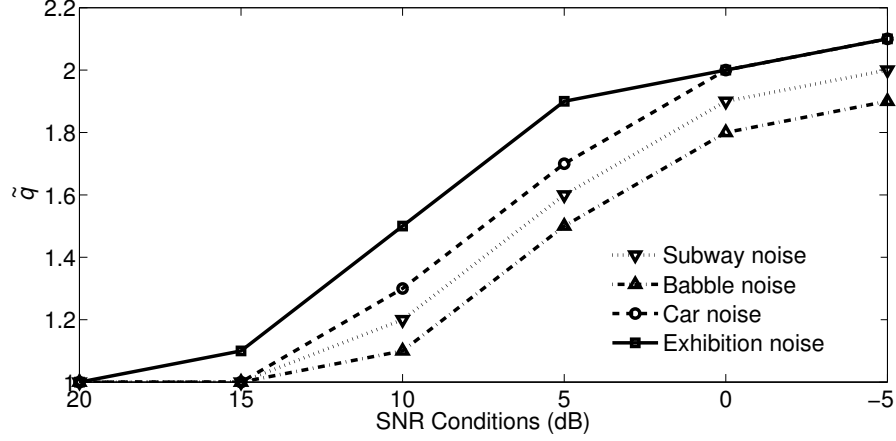


Figure 6.2: The optimal q -values for different SNR conditions.

We set $\beta = 0.01$ in this chapter. It is applied for the three spectral subtraction methods, LSS, NSS and q -SS.

6.4 Experimental setup

We evaluated q -SS on the Aurora-2 and CENSREC-2 tasks. We implemented the same features, and speech recognition system as explained in Section 5.6.

6.5 Experimental results and discussions

6.5.1 Q -Gaussian representation of noisy speech

We investigated the q -Gaussianity of noisy speech in a similar way explained in Section 6.3.3. The difference was the amount of utterances we used to build the histograms. In this section, we used 200 utterances of female speakers for each SNR condition from Test Set A of the Aurora-2 database. We only considered a single DFT coefficient (50-th coefficient) from a total of 256 coefficients.

Figure 6.2 shows the estimated q -value for each noise conditions and for each SNR condition. As we can see, the optimum q -value is higher when the SNR is lower.

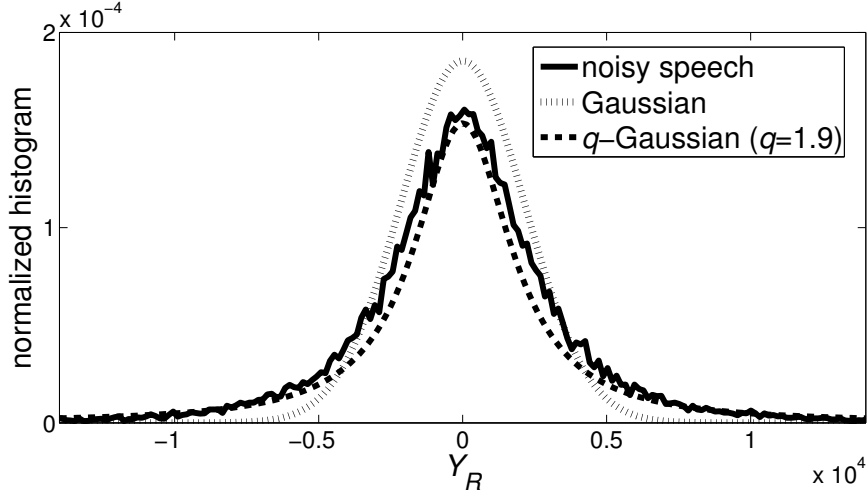


Figure 6.3: Gaussian and q -Gaussian distributions fitted to the histogram of the speech data corrupted with subway noise at 0 dB SNR.

Figure 6.3 shows that the q -Gaussian distribution with $q = 1.9$ better fits the noisy speech than a Gaussian distribution ($q = 1$) does.

6.5.2 Performance of q -SS

We first conducted several experiments to evaluate the performance of q -SS_c. We varied q from 1 to 2. We varied q with increments 0.1 from $q = 1.0$ to $q = 1.8$. Since there was an abrupt changes from $q = 1.8$ to $q = 2$, we varied q with increments 0.01 in this region.

Figures 6.4 and 6.5 show the average accuracy of q -SS_c when the q -value is varied from 1 to 2 on the Aurora-2 and CENSREC-2 databases respectively. We found that the performance of q -SS was improved when $q > 1$, the case when q -Gaussian is a heavy tailed distribution. The improvements were obtained for both databases. For Aurora-2, the best accuracy was obtained when $q = 1.88$, with 18.1% error reduction rate from the case when $q = 1$, i.e the case when q -SS_c was the same as LSS. For CENREC2, the best accuracy was obtained when $q = 1.87$, where we achieved 25.6% error reduction rate. We noticed that spectral subtraction was not so effective on the CENSREC-2 database. This is probably because of inaccuracies of the noise estimation due to the non-stationary characteristics of noise in the CENSREC-2 database.

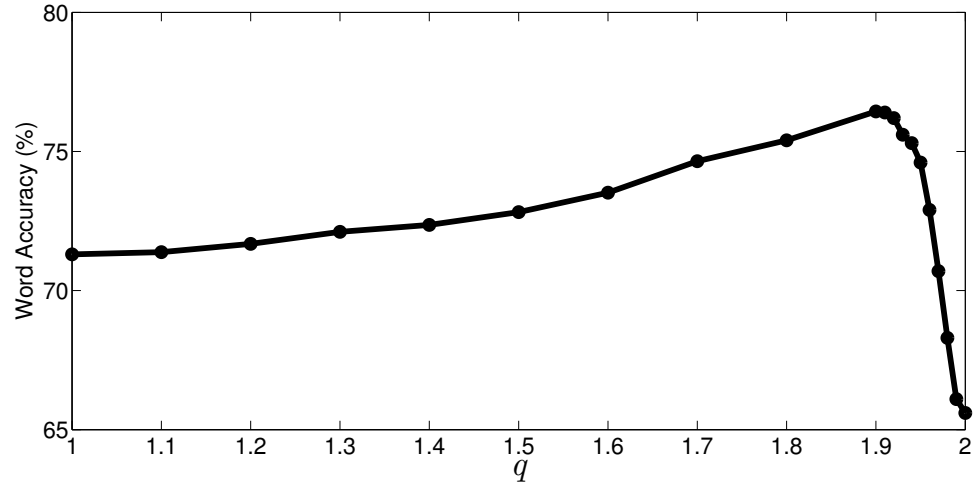


Figure 6.4: Performance of q -SS-c for $1 \leq q \leq 2$ for the Aurora-2 database. The average values over 0-20 dB of the SNR conditions of all the noise types are shown

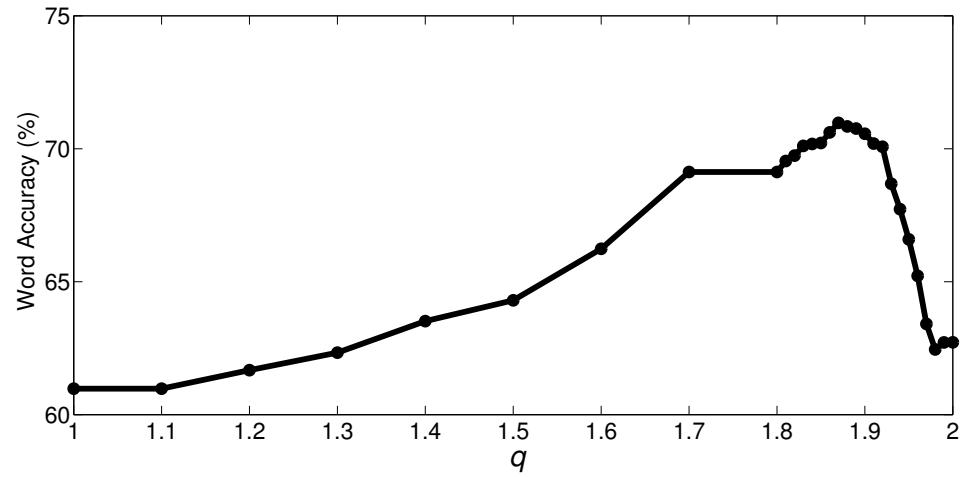


Figure 6.5: Performance of q -SS-c for $1 \leq q \leq 2$ for the CENSREC-2 database. The average values over all the evaluation conditions are shown

Figure 6.6 shows the performance of q -SS_c for different SNR conditions on the Aurora-2 database. We noticed that the word accuracies drastically degraded when q was from 1.9 to 2.0. This occurred for all SNR conditions except for 20dB SNR. This is similar with the results for the CENSREC-2 database as shown in Fig. 6.7. The performance was dropped drastically for $1.9 \leq q \leq 2$. When q approaches 2, the weight factor in q -SS, i.e. $v(q)$, approaches zero. Therefore, it is very likely that $v(q)|Y_f|^2 < |N_f|^2$. In this condition, the clean speech estimate, $|\hat{X}|^2$, equals to the flooring, $\beta|Y_f|^2$, and hence the recognition accuracy drops to the condition where no spectral subtraction method is applied.

For Aurora 2, we found that q -SS was significantly effective when the SNR conditions are between 0 to 10 dB. It was not significantly effective under the higher SNR conditions. Similar results were found for CENSREC-2. Our method, q -SS, was much more effective for Conditions 3 and 4. These results were as expected since the SNR in Conditions 3 and 4 were lower than that in Conditions 1 and 2.

We noticed some inconsistencies of our experimental results with the results in Section 6.3.3 and 6.5.1. When the SNR > 5 dB, the performance was better when a higher q was applied. We did not expect these results. These results suggested that the SNR might not be the only factors that could affect q or the relation between the SNR and q might not be linear. In the previous sections, we used the variance from the data to find the optimum distribution. It should be noted that the variance of a q -Gaussian distribution depends on q . The variance of the q -Gaussian distribution is larger when the q is higher. In other words, we could fit the distribution of the clean spectra into the q -Gaussian distributions at a higher q if the variance is optimized. In high SNR conditions, where speech is more dominant than noise, the distribution of the noisy speech spectra was heavily influenced by the distribution of the clean speech spectra. This could affect the optimum q -value for recognition task. Further study is needed to investigate whether we could relate the distribution of noisy speech and clean speech.

We also conducted the experiments using an adaptive q -value, i.e. q -SS_m. Tables 6.1 and 6.2 summarize the comparison of q -SS_m and q -SS_c ($q = 1.88$) for eight types of noise in Aurora-2. The best performances for each noise type and SNR were printed in bold. We found that q -SS_m was better in average than that of q -SS_c. q -SS_m was better especially in the lower SNR conditions. These results suggest that the SNR could be used as a parameter to control q . Table 6.3 compares q -SS_m and q -SS_c

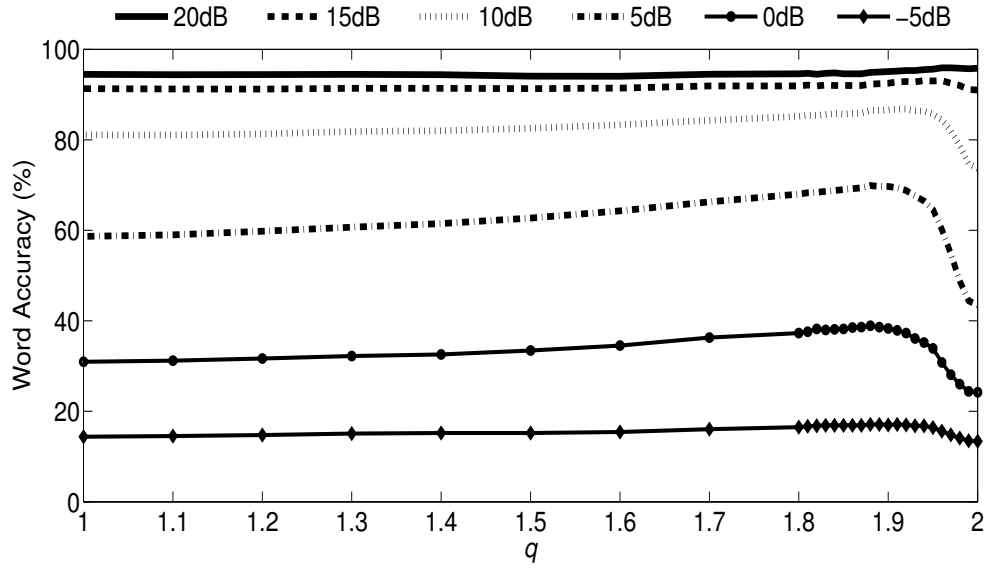


Figure 6.6: Word Accuracy of q -SS_c for different SNR conditions of the Aurora-2 database. The averaged values over all the noise types are shown.

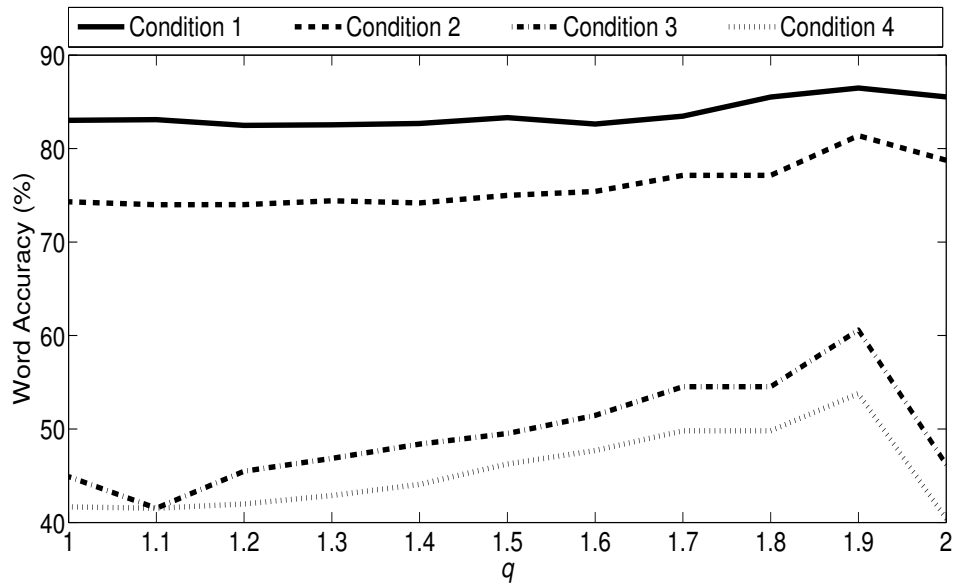


Figure 6.7: Word Accuracy of q -SS_c for each evaluation condition of the CENSREC-2 database.

Table 6.1: Performance comparison (Word Accuracy (%)) of q -SS_c ($q = 1.88$) and q -SS_m for different types of noise and SNR conditions of the of Aurora-2 database

Methods	Noise Types	SNR (dB)					
		20	15	10	5	0	-5
q -SS _c ($q=1.88$)	Subway	97.5	95.0	90.4	75.4	44.2	20.3
q -SS _m		96.6	94.4	89.7	76.1	45.0	20.3
q -SS _c ($q=1.88$)	Babble	91.0	87.5	81.6	65.6	38.1	17.0
q -SS _m		91.3	89.2	82.4	67.8	39.3	16.1
q -SS _c ($q=1.88$)	Car	97.9	97.0	92.3	75.8	38.0	15.7
q -SS _m		97.3	96.5	92.0	75.7	37.6	15.3
q -SS _c ($q=1.88$)	Exhibition	95.7	93.2	87.7	69.6	34.9	14.4
q -SS _m		95.7	92.9	87.7	72.6	39.1	15.7
q -SS _c ($q=1.88$)	Restaurant	88.9	84.4	77.3	62.2	38.5	15.2
q -SS _m		89.1	85.6	79.1	64.0	39.6	16.1
q -SS _c ($q=1.88$)	Street	97.2	95.2	89.3	72.2	40.2	17.6
q -SS _m		96.7	94.7	89.0	73.2	40.8	17.5
q -SS _c ($q=1.88$)	Airport	91.4	89.0	84.9	71.6	45.7	20.1
q -SS _m		92.2	90.4	85.2	72.1	46.3	20.3
q -SS _c ($q=1.88$)	Station	95.4	93.4	88.8	74.0	43.1	18.3
q -SS _m		94.5	92.9	88.7	73.6	42.2	18.1

Table 6.2: Performance comparison (Word Accuracy (%)) of q -SS_c ($q = 1.88$) and q -SS_m for each test set in Aurora-2.

Algoritm	Test A	Test B	Test C	Ave.
q -SS _c ($q = 1.88$)	77.4	76.1	75.3	76.5
q -SS _m	77.9	76.5	76.6	77.1

Table 6.3: Performance comparison (Word Accuracy (%)) of q -SS_c ($q = 1.87$) and q -SS_m for each evaluation condition in CENSREC-2.

Methods	Evaluation Condition				Average
	Condition 1	Condition 2	Condition 3	Condition 4	
q-ss _c	86.6	81.8	60.45	55.0	71.0
q-ss _m	85.1	79.7	62.0	56.1	70.7

($q = 1.87$) for each evaluation condition of the CENSREC-2 database. We found that q -SS_m was better for Condition 3 and 4, where the SNR conditions were lower.

6.5.3 Comparison with NSS

In this section, we compared the performance of q -SS and NSS. First, we compared NSS_c and q -SS_c, for $1 \leq q \leq 1.9$, i.e. $1 \leq \alpha \leq 5.5$. The results for Aurora-2 and CENSREC-2 are shown in Figs. 6.8 and 6.9 respectively. We found that the performance of q -SS was better than that of NSS for each pair of q and α .

To analyze these results, we compared the attenuation curves of q -SS and NSS. The attenuation curve tells us about how much a signal is suppressed (in dB) when we apply NSS and q -SS for each SNR condition. To obtain the attenuation curve, we first find the transfer function of q -SS filter. It is formulated as follows:

$$|\hat{X}_f| = H_{q\text{-SS}}(f)|Y_f|, \quad (6.30)$$

where $H_{q\text{-SS}}(f)$ can be seen as a time-varying filter given by:

$$H_{q\text{-SS}} = \left(\frac{v(q)|Y_f|^2 - |N_f|^2}{|Y_f|^2} \right)^{0.5}. \quad (6.31)$$

Eq. (6.31) can be expressed in term of the *a posteriori* SNR, γ_f , as follows:

$$H_{q\text{-SS}} = \left(\frac{v(q) \cdot \gamma_f - 1}{\gamma} \right)^{0.5}. \quad (6.32)$$

The attenuation (dB) is then calculated using the following formula:

$$\text{Attenuation} = 20 \log_{10} H_{q\text{-SS}}. \quad (6.33)$$

Meanwhile, the transfer function of NSS filter is formulated as:

$$H_{\text{NSS}} = \left(\frac{|Y_f|^2 - \alpha|N_f|^2}{|Y_f|^2} \right)^{0.5}. \quad (6.34)$$

Eq. (6.34) can be expressed in term of the *a posteriori* SNR, γ_f , as follows:

$$H_{\text{NSS}} = \left(\frac{\gamma_f - \alpha}{\gamma} \right)^{0.5}. \quad (6.35)$$

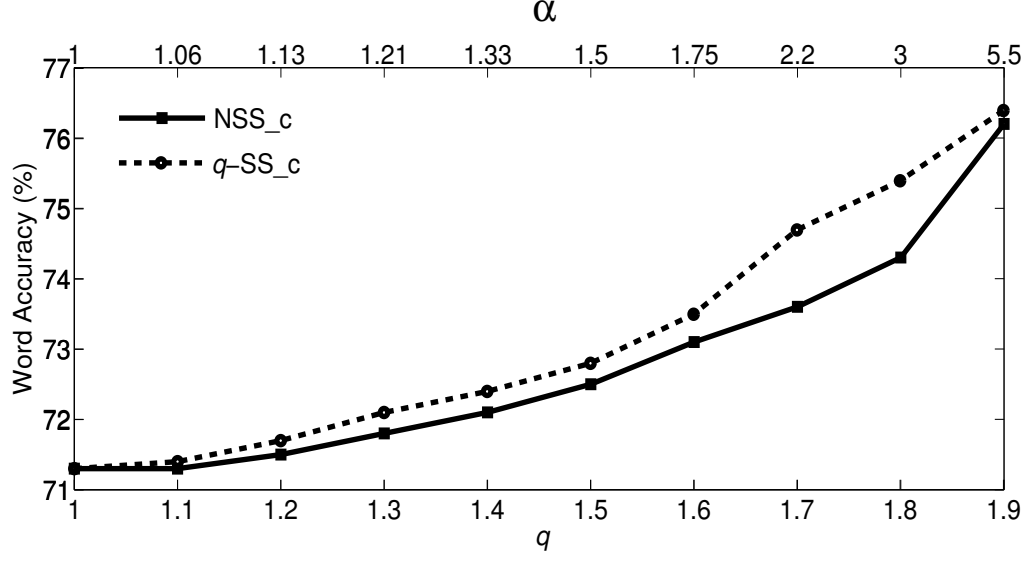


Figure 6.8: Performance comparison of q -SS_c and NSS_c for $1 \leq q \leq 2$ and its equivalent α value for Aurora-2. The average values over 0-20 dB of the SNR conditions of all the noise types are shown.

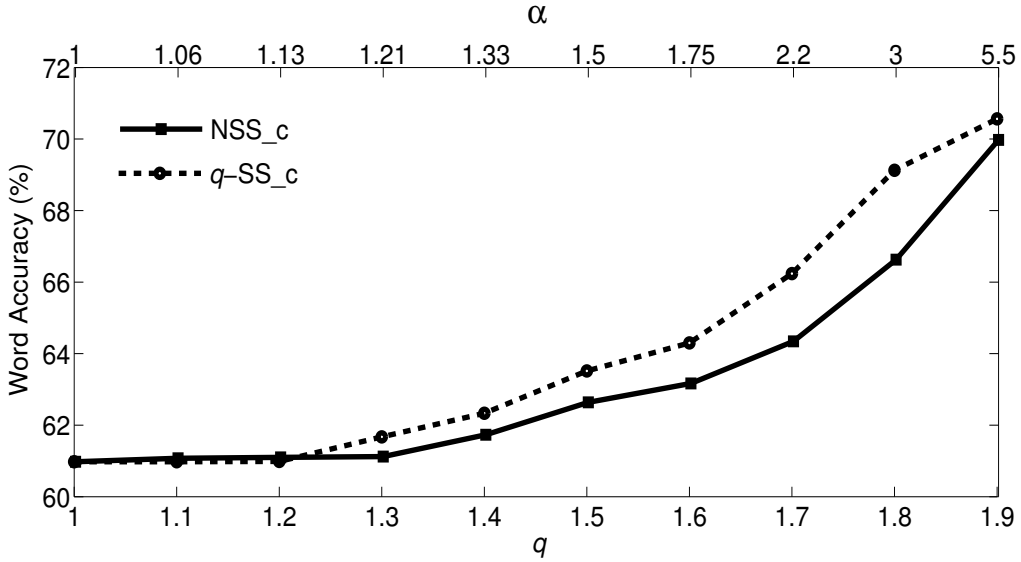


Figure 6.9: Performance comparison of q -SS_c and NSS_c for $1 \leq q \leq 2$ and its equivalent α value for CENSREC-2. The average values over all the evaluation conditions are shown.

We can rewrite Eq. (6.35) as follows:

$$H_{\text{NSS}} = \left(\frac{\alpha \left(\frac{\gamma_f}{\alpha} - 1 \right)}{\gamma_f} \right)^{(0.5)} \quad (6.36)$$

Since $\alpha = \frac{1}{v(q)}$, we can write Eq. (6.36) as follows:

$$H_{\text{NSS}} = \left(\frac{1}{v(q)} \right)^{0.5} H_{q\text{-SS}} \quad (6.37)$$

We can see that, q -SS is basically NSS with more attenuation.

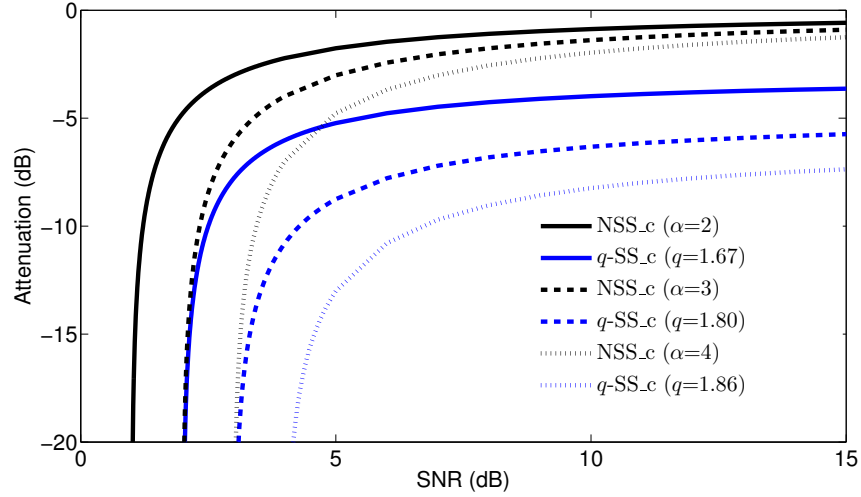


Figure 6.10: The comparison of the attenuation curve of q -SS_c and NSS_c for several values of q and their respective α value.

Figure 6.10 compares the attenuation curve of NSS and q -SS as functions of the SNR. We noticed that q -SS applied more attenuation than that of NSS at high SNR condition. For instance, for $\alpha = 2$, i.e. $q = 1.67$, if the SNR was 5dB, the attenuation was -2.22dB with NSS and -5.23dB with q -SS. As we can see from Eq. (6.37), the transfer function of q -SS was the transfer function of NSS attenuated with a factor $v(q)^{0.5}$. In speech recognition, scaling down (attenuation) of the signals did not affect its performance. Therefore, the performance of speech recognition would not be affected much by the attenuation at the high SNR conditions. However, flooring occurred at a higher SNR for q -SS than for NSS. For instance, for $\alpha = 3$, i.e. $q = 1.8$, flooring

Table 6.4: The performance comparison (Word accuracy (%)) of several spectral subtraction methods for each SNR conditions of the Aurora-2 database. The average values over all noise types are shown.

Methods	SNR						Average (0-20 dB)
	20	15	10	5	0	-5	
No Compensation	95.9	91.0	73.7	43.5	24.2	13.4	65.7
LSS	94.5	91.3	81.1	58.7	31.0	14.4	71.3
NSS _c ($\alpha = 5.5$)	94.4	91.9	86.1	70.1	38.8	16.7	76.2
NSS _n	94.1	91.4	83.7	65.5	35.1	15.4	74.0
q -SS _c ($q = 1.88$)	95.1	92.5	86.6	69.7	38.3	17.0	76.4
q -SS _m	94.6	92.5	86.8	71.3	40.1	17.3	77.1

Table 6.5: The performance comparison (Word accuracy (%)) of several spectral subtraction methods for each evaluation condition of the CENSREC-2 database.

Methods	Evaluation Condition				Average
	Condition 1	Condition 2	Condition 3	Condition 4	
No Compensation	85.5	78.8	46.2	40.4	62.7
LSS	83.0	74.3	44.9	41.7	61.0
NSS _c	85.2	79.1	60.6	55.0	70.0
NSS _n	83.5	76.3	56.9	50.4	66.8
q -SS _c ($q=1.87$)	86.7	81.8	60.1	55.0	71.0
q -SS _m	85.1	79.7	62.0	56.1	70.7

occurred at 2dB SNR for NSS and 3dB for q -SS. Since it was more difficult to estimate noise when the SNR was low, it is very likely that the clean speech estimate had more distortions in this region. Thus, flooring these regions may minimized the distortions. In other words, the q -SS formulation has better masking than NSS.

Table 6.4 summarizes the performance of several spectral subtraction methods for each SNR conditions of the Aurora-2 database. The results for CENSREC-2 is summarized in Table 6.5. For Aurora-2, the performance of q -SS_m was better than NSS for both constant α (NSS_c) and adaptive α (NSS_n). Using adaptive q -value, i.e. q -SS_m achieved the best performance in average compared to the other spectral subtraction methods. Meanwhile, for CENSREC-2, q -SS_c achieved the best performance compared to other methods in average. However, q -SS_m performed better in Condition 3

and 4, where the SNR is lower. These results also confirmed that the parameter q can be controlled using the SNR information.

6.6 Conclusions

We derive a nonlinear spectral subtraction method based on the q -Gaussian distribution assumption for noisy speech. We call it q -SS. The q -Gaussian distribution is a heavy tailed distribution which can arise from the sum of correlated random variables. In our analysis, the q -Gaussian distributions fit noisy speech better than Gaussian distributions do.

Our approach gives a consistent way to estimate the control parameter α in NSS from the spectra of observed noisy speech. Our speech recognition results on the Aurora-2 database showed that q -SS was better than the conventional spectral subtraction and nonlinear spectral subtraction. Our experiments also confirmed that the SNR can be used to control the parameter q .

Chapter 7

Conclusions and future works

7.1 Conclusions

In this thesis, we implement the theory non-extensive statistics to improve the robustness of speech recognition in noisy environments. The purpose is to address the correlations between the speech spectral components and between speech and noise. These conditions are often ignored in most robust speech recognition techniques. Several efforts have been put to solve these problems. However, most of them solve them on the extensive framework. In the extensive framework, the correlations between speech components and between speech and noise need to be estimated explicitly. However, they have been proven difficult and often impossible to obtain.

By implementing a non-extensive framework, we consider that the correlations exist, which degree is affected by a parameter q . When q equals to 1, it will be the same with the extensive framework. Our preliminary findings show that the correlations and the parameter q are deeply related in statistical sense.

We proposed two methods: q -log spectral mean normalization (q -LSMN) and q -spectral subtraction (q -SS). Q -LSMN is a feature normalization method in the q -log spectral domain. In q -LSMN, we implement the q -log function from non-extensive statistics to provide a domain with non-additivity. The purpose is the same as the implementation of this function in Tsallis entropy: to represent nonadditivity of the speech features, which exists when noise and speech are correlated.

Our second method, q -SS, is a spectral subtraction method which is derived based on the assumption that noisy speech follows the q -Gaussian distribution. The q -Gaussian

distribution a heavy-tailed typed distribution when $q > 1$ which is found to be useful to model complex systems. Our experiments showed that the q -Gaussian distribution fits noisy speech better than a Gaussian distribution does. Our derivation shows a surprising result. We can naturally express a weight factor to the spectral subtraction formula, which is similar to the control parameter α in NSS. This control parameter has been shown to compensate the correlation between speech and noise. In other words, q -SS gives a consistent way to estimate α .

We evaluated the proposed methods on two types of database: Aurora 2 and CENSREC-2. These databases have different environmental conditions. Aurora-2 is a synthesized noisy database, where noises are added artificially to clean speech. This database is a popular database used to evaluate robust speech recognition techniques. However, this database has a major shortcoming. While noises used are real noise, this database does not reflect the real situation where speakers tend to modify the way they speak in noisy environment (Lombard effect). For this reason, we also used the CENSREC-2 database. This database is recorded in real driving environments.

Our evaluation results showed that our proposed methods are better than the methods that are derived in the extensive framework. Improvements were obtained when $q \neq 1$ were used, a condition of the extensive statistics. For q -LSMN, an improvement in performance was obtained when $q < 1$. It was better than that of CMN and MVN for both databases. We found different optimum q for both databases, 0.7 for Aurora-2 and 0.4 for CENSREC-2. The difference might be caused by the difference in environmental conditions between these databases. In real word, the sources of noise are more likely more than one source. This increases the cross-term between noise and speech, hence a lower q is required. We also combined q -LSMN with NSS and found that our method can be used as complementary to additive noise removal. The combination of NSS and q -LSMN was better than the same combination with CMN and MVN on both databases. For q -SS, an improvement in performance was obtained when $q > 1$, the condition when the q -Gaussian distribution is a heavy-tailed distribution. q -SS is better than LSS when $1 < q < 1.9$. We also derive a way to optimize q . We estimate q from the distribution that best fit the q -Gaussian distribution to noisy speech. Our experiments show that this estimation proves to be effective especially in the low SNR conditions.

7.2 Future works

In this thesis, we have shown that non-extensive statistics was useful to improve the robustness of speech recognition against environmental noise. Our experiments showed that the use of $q \neq 1$, where the non-extensivity exists, was effective in improving the performance of the robust speech recognition techniques.

Our q -LSMN is based on non-extensive statistics. While it has been shown to be successful in interpreting some physical phenomena that cannot be fully explained by extensive statistics, the physical meaning of the q -value has not yet been explored much and should be investigated in future. Another interesting topic would be the implementation of other compensation methods in the q -log domain.

Noises in real world affect each frequency band of speech differently [Kamath and Loizou, 2002a]. For that reason, the optimal q -value may be different in different frequency bands. While we used the same values of q for all the bands in this research, the optimization of q -values to each band might be promising.

It is also interesting to apply other compensation methods in the non-extensive framework. Several researches have shown that the use of root function produced more robust features against noise than log function. However, their performance is strongly affected by the normalization method [Hunt, 1999; Lockwood and Alexandre, 1994]. In this research, we implement the q -log function in the spectral domain. It is interesting to investigate the performance of the q -log function when it is implemented in the cepstral domain, as a normalization method using this function can easily be derived.

Our q -SS is an optimum variance estimator of the speech spectrum in the non-extensive framework. While, we found that the parameter q in q -SS can be controlled using the SNR, their relation may not be linear. Therefore, the meaning of q should be investigated further in the future. It would also be interesting to extend the q -Gaussian assumption to derive the true spectral estimator of signal. While in this research no assumption was made on the clean speech distribution, it would be interesting to model the clean speech spectra using the q -Gaussian distributions.

Finally, our work has shown it is possible to implement the non-extensive framework without changing much the current framework of speech recognition system. However, considering the effectiveness of this framework in robust speech recognition problems,

it may be interesting to extend the current speech recognition systems into the non-extensive framework.

List of publications

Publications related to Chapter 5

Journal

1. H. Pardede, K. Iwano and K. Shinoda, “ Feature normalization based on non-extensive statistics for speech recognition”, Speech Communication, Vol 55, Issue 5, pp. 587599, 2013

International Conferences and Workshops (*Peer reviewed*)

1. H. Pardede and K. Shinoda, “ Non-extensive statistics for feature normalization in speech recognition”, IWSML 2012, Mar 2012.
2. H. Pardede and K. Shinoda, “ Generalized-Log Spectral Mean Normalization for Speech Recognition ”, Proc Interspeech 2011, pp.1645-1648, Aug.27, 2011.

National Conferences and Meetings (*Not reviewed*)

1. H. Pardede, K. Shinoda, and K. Iwano “A Compensation Technique Using q -Logarithm for Noisy Speech Recognition ”, ASJ Spring Meeting 2012, pp. 19-20, Mar 2012.
2. H. Pardede, K. Shinoda, and K. Iwano “Nonlinear Normalization Using q -Logarithm for Robust Speech Recognition ”, IEICE Technical Report, Vol. 111, No. 153, pp. 45-50, Jul 2011.

Publications related to Chapter 6

Journal

1. H. Pardede, K. Iwano and K. Shinoda, “Spectral subtraction based on non-extensive statistics for speech recognition”, IEICE Transaction on Information and Systems, accepted for publication

International Conferences and Workshops (*Peer reviewed*)

1. H. Pardede, K. Shinoda, and K. Iwano “ Q -Gaussian based spectral subtraction for robust speech recognition ”, Proc Interspeech 2012, Sept., 2012.

National Conferences and Meetings (*Not reviewed*)

1. H. Pardede, K. Shinoda, and K. Iwano “Spectral Subtraction Based on q -Gaussian Assumption for Noise Robust Speech Recognition ”, ASJ Spring Meeting 2012, pp. 21-22 , Mar 2012.

References

- Acero, A., Deng, L., Kristjansson, T., and Zhang, J. (2000). HMM adaptation using vector Taylor series for noisy speech recognition. *Proc. IEEE Internat. Conf. Spoken Language Processing*, 3:869–872.
- Agarwal, A. and Cheng, Y. M. (1999). Two-stage Mel-warped Wiener filter for robust speech recognition. *Proc. IEEE Workshop on Automatic Speech Recognition and Understanding*, pages 12–15.
- Alexandre, P., Boudy, J., and Lockwood, P. (1993). Root homomorphic deconvolution schemes for speech processing in car noise environments. *IEEE Internat. Conf. Acoustics, Speech and Signal Processing*, 2:99–102 vol.2.
- Alsteris, L. D. and Paliwal, K. K. (2007). Short-time phase spectrum in speech processing: A review and some experimental results. *Digit. Signal Process.*, 17(3):578 – 616.
- Atal, B. S. (1974). Effectiveness of linear prediction characteristics of the speech wave for automatic speaker identification and verification. *J. Acoust. Soc. Amer.*, 55(6):1304–1312.
- Avendano, C. and Hermansky, H. (1997). On the effects of short-term spectrum smoothing in channel normalization. *IEEE Trans. Speech Audio Process.*, 5(4):372–374.
- Baum, L. E. and Eagon, J. A. (1967). An inequality with applications to statistical estimation for probabilistic functions of Markov processes and to a model for ecology. *Bull. Amer. Math. Soc.*, 73:360–363.
- Beck, C. and Cohen, E. (2003). Superstatistics. *Physica A*, 322:267 – 275.

- Berouti, M., Schwartz, R., and Makhoul, J. (1979). Enhancement of speech corrupted by acoustic noise. *Proc. IEEE Internat. Conf. on Acoustics, Speech and Signal Processing*, 4:208 – 211.
- Bezerianos, A., Tong, S., and Thakor, N. (2003). Time-dependent entropy estimation of EEG rhythm changes following brain ischemia. *Ann. Biomed. Eng.*, 31:221–232.
- Boll, S. F. (1979). Suppression of acoustic noise in speech using spectral subtraction. *IEEE Trans. Acoust. Speech Signal Process.*, 27(2):113–120.
- Borges, E. P. (2004). A possible deformed algebra and calculus inspired in nonextensive thermostatics. *Physica A*, 340:95–101.
- Caruso, F., Pluchino, A., Latora, V., Vinciguerra, S., and Rapisarda, A. (2007). Analysis of self-organized criticality in the olami-feder-christensen model and in real earthquakes. *Phys. Rev. E*, 75:055101.
- Chen, B. and Loizou, P. C. (2007). A Laplacian-based MMSE estimator for speech enhancement. *Speech Commun.*, 49(2):134–143.
- Chen, L. H. Y. and Shao, Q.-M. (2004). Normal approximation under local dependence. *The Annals of Probability*, 32(3):pp. 1985–2028.
- Cohen, I. (2005). Relaxed statistical model for speech enhancement and a priori SNR estimation. *IEEE Trans. Speech Audio Process.*, 13(5):870 – 881.
- Dahl, G., Yu, D., Deng, L., and Acero, A. (2012). Context-dependent pre-trained deep neural networks for large-vocabulary speech recognition. *IEEE Trans. Audio Speech Lang. Processing*, 20(1):30 –42.
- Davis, S. and Mermelstein, P. (1980). Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences. *IEEE Trans. Acoust. Speech Signal Process.*, 28(4):357 – 366.
- Dempster, A. P., Laird, N. M., and Rubin, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm. *J. R. Stat. Soc. Series B. Stat. Methodol.*, 39(1):1–38.

- Deng, L., Droppo, J., and Acero, A. (2004). Enhancement of log Mel power spectra of speech using a phase-sensitive model of the acoustic environment and sequential estimation of the corrupting noise. *IEEE Trans. Speech Audio Process.*, 12(2):133 – 143.
- Doblinger, G. (1995). Computationally efficient speech enhancement by spectral minima tracking in subbands. *Proc. Eurospeech*, pages 1513–1516.
- Ephraim, Y. and Malah, D. (1984). Speech enhancement using a minimum-mean square error short-time spectral amplitude estimator. *IEEE Trans. Acoust. Speech Signal Process.*, 32(6):1109 – 1121.
- Ephraim, Y. and Malah, D. (1985). Speech enhancement using a minimum mean-square error log-spectral amplitude estimator. *IEEE Trans. Acoust. Speech Signal Process.*, 33(2):443 – 445.
- ETSI standard doc. (2002). *Speech processing, transmission and quality aspects (STQ); Distributed speech recognition; Advanced feature extraction algorithm; Compression algorithm*. ETSI ES 202 050 Ver.1.1.1 (2002-10).
- Evans, N., Mason, J., Liu, W., and Fauve, B. (2006). An assessment on the fundamental limitations of spectral subtraction. *Proc. IEEE Internat. Conf. on Acoustics, Speech and Signal Processing*, 1:1520–6149.
- Faubel, F., McDonough, J., and Klakow, D. (2008). A phase-averaged model for the relationship between noisy speech, clean speech and noise in the log-mel domain. *Proc. Interspeech*, pages 553–556.
- Flores, J. and Young, S. (1994). Continuous speech recognition in noise using spectral subtraction and HMM adaptation. *Proc. IEEE Internat. Conf. on Acoustics, Speech and Signal Processing*, i:I/409 –I/412 vol.1.
- Forney, G.D., J. (1973). The Viterbi algorithm. *Proceedings of the IEEE*, 61(3):268 – 278.
- Furui, S. (1981). Cepstral analysis technique for automatic speaker verification. *IEEE Trans. Acoust. Speech Signal Process.*, 29(2):254 – 272.

- Gales, M. and Young, S. (1992). An improved approach to the hidden Markov model decomposition of speech and noise. In *Acoustics, Speech, and Signal Processing, 1992. ICASSP-92., 1992 IEEE International Conference on*, volume 1, pages 233–236 vol.1.
- Gales, M. and Young, S. (1993). Cepstral parameter compensation for HMM recognition in noise. *Speech Commun.*, 12(3):231 – 239.
- Gales, M. and Young, S. (1996). Robust continuous speech recognition using parallel model combination. *IEEE Trans. Speech Audio Process.*, 4(5):352–359.
- Gazor, S. and Zhang, W. (2003). Speech probability distribution. *IEEE Signal Process. Lett.*, 10(7):204–207.
- Gradojevic, N. and Gencay, R. (2008). Overnight interest rates and aggregate market expectations. *Economics Letters*, 100(1):27 – 30.
- Gradojevic, N. and Gencay, R. (2011). Financial applications of nonextensive entropy [applications corner]. *IEEE Signal Process. Mag.*, 28(5):116–141.
- Hasumi, T. (2009). Hypocenter interval statistics between successive earthquakes in the two-dimensional burridgeknopoff model. *Physica A*, 388(4):477 – 482.
- Hermansky, H. (1990). Perceptual linear predictive (PLP) analysis of speech. *J. Acoust. Soc. Am.*, 87(4):1738–1752.
- Hirsch, H. and Ehrlicher, C. (1995). Noise estimation techniques for robust speech recognition. *Proc. IEEE Internat. Conf. on Acoustics, Speech and Signal Processing*, 1:153–156.
- Hirsch, H. and Pearce, D. (2000). The aurora experimental framework for the performance evaluation of speech recognition systems under noisy conditions. *Proc. ISCA ITRW ASR2000*, pages 181–188.
- Hunt, M. J. (1999). Spectral signal processing for ASR. *Proc IEEE Workshop Autom. Speech Recognit. Underst.*, pages 17–25.

- Iso, K. and Watanabe, T. (1990). Speaker-independent word recognition using a neural prediction model. In *IEEE Internat. Conf. on Acoustics, Speech and Signal Processing*, pages 441–444 vol.1.
- Itahashi, S. (1990). Recent speech database projects in Japan. *Proc. IEEE Internat. Conf. Spoken Language Processing*, pages 1081–1084.
- Ito, Y., Matsumoto, H., and Yamamoto, K. (2000). Forward masking on a generalized logarithmic scale for robust speech recognition. *Proc. Interspeech*, pages 530–533.
- Jeong, J. and Williams, W. (1992). Mechanism of the cross-terms in spectrograms. *IEEE Trans. Signal Process.*, 40(10):2608–2613.
- Jiulin, D. (2007). Nonextensivity and the power-law distributions for the systems with self-gravitating long-range interactions. *Astrophys. Space Sci.*, 312:47–55.
- Kadambe, S. and Boudreaux-Bartels, G. (1992). A comparison of the existence of ‘cross terms’ in the Wigner distribution and the squared magnitude of the wavelet transform and the short-time Fourier transform. *IEEE Trans. Signal Process.*, 40(10):2498 – 2517.
- Kamath, S. and Loizou, P. (2002a). A multi-band spectral subtraction method for enhancing speech corrupted by colored noise. *IEEE Internat. Conf. Acoustics, Speech and Signal Processing*, 4:IV–4164.
- Kamath, S. and Loizou, P. (2002b). Enhancement of speech corrupted by acoustic noise. *Proc. IEEE Internat. Conf. on Acoustics, Speech and Signal Processing*, 4:IV–4164.
- Kim, C. and Stern, R. M. (2009). Feature extraction for robust speech recognition using a power-law nonlinearity and power-bias subtraction. *Proc. Interspeech*, pages 28–31.
- Kim, D. Y., Un, C. K., and Kim, N. S. (1998). Speech recognition in noisy environments using first-order vector Taylor series. *Speech Commun.*, 24(1):39 – 49.
- Kitaoka, N., Akahori, I., and Nakagawa, S. (2001). Speech recognition under noisy environments using spectral subtraction with smoothing of time direction and real-time cepstral mean normalization. *Proc. Interspeech*, pages 159–162.

- Kobayashi, T. and Imai, S. (1984). Spectral analysis using generalized cepstrum. *IEEE Trans. Acoust. Speech Signal Process.*, 32(5):1087 – 1089.
- Kozaki, M. and Sato, A.-H. (2008). Application of the beck model to stock markets: Value-at-risk and portfolio risk assessment. *Physica A*, 387(56):1225 – 1246.
- Lee, K.-F. (1988). On large-vocabulary speaker-independent continuous speech recognition. *Speech Commun.*, 7(4):375 – 379.
- Li, J., Deng, L., Yu, D., Gong, Y., and Acero, A. (2009). A unified framework of HMM adaptation with joint compensation of additive and convolutive distortions. *Computer Speech and Language*, 23(3):389 – 405.
- Lim, J. (1979). Spectral root homomorphic deconvolution system. *IEEE Trans. Acoust. Speech Signal Process.*, 27(3):223 – 233.
- Lockwood, P. and Alexandre, P. (1994). Root adaptive homomorphic deconvolution schemes for speech recognition in noise. *IEEE Internat. Conf. Acoustics, Speech and Signal Processing*, i:I/441 –I/444 vol.1.
- Lockwood, P. and Boudy, J. (1992). Experiments with a nonlinear spectral subtractor (NSS), hidden Markov models and the projection, for robust speech recognition in cars. *Speech Commun.*, 11(2-3):215 – 228.
- Lu, Y. and Loizou, P. C. (2008). A geometric approach to spectral subtraction. *Speech Commun.*, 50:453–466.
- Macho, D., Mauuary, L., Noé, B., Cheng, Y. M., Ealey, D., Jouvét, D., Kelleher, H., Pearce, D., and Saadoun, F. (2002). Evaluation of a noise-robust DSR front-end on Aurora databases. *Proc. Interspeech*.
- Makhoul, J. (1975). Spectral linear prediction: Properties and applications. *IEEE Trans. Acoust. Speech Signal Process.*, 23(3):283 – 296.
- Malacarne, L. C., Mendes, R. S., and Lenzi, E. K. (2001). q -exponential distribution in urban agglomeration. *Phys. Rev. E.*, 65:017106.

- Martin, R. (2002). Speech enhancement using MMSE short time spectral estimation with gamma distributed speech priors. *Proc. IEEE Internat. Conf. on Acoustics, Speech and Signal Processing*, 1:I-253 –I-256.
- Martin, R. (2005). Speech enhancement based on minimum mean-square error estimation and supergaussian priors. *IEEE Trans. Speech Audio Process.*, 13(5):845 – 856.
- Mauuary, L. (1996). Blind Equalization for robust telephone based speech recognition. *Proc. EUSIPCO*.
- McAulay, R. and Malpass, M. (1980). Speech enhancement using a soft-decision noise suppression filter. *IEEE Trans Acoust.*, 28(2):137 – 145.
- McAuley, J., Ming, J., Stewart, D., and Hanna, P. (2005). Subband correlation and robust speech recognition. *IEEE Trans. Speech Audio Process.*, 13(5):956 – 964.
- Ming, J., O’Boyle, P., McMahon, J., and Smith, F. (1996). Speech recognition using a strong correlation assumption for the instantaneous spectra. *Proc. Internat. Conf. Spoken Language Processing*, 2:1061 –1064 vol.2.
- Mohamed, A., Dahl, G., and Hinton, G. (2012). Acoustic modeling using deep belief networks. *IEEE Trans. Audio Speech Lang. Processing*, 20(1):14 –22.
- Molau, S., Pitz, M., Schluter, R., and Ney, H. (2001). Computing Mel-frequency cepstral coefficients on the power spectrum. *IEEE Internat. Conf. Acoustics, Speech and Signal Processing*, 1:73 –76 vol.1.
- Moreira, D., Albuquerque, E., da Silva, L., and Galvo, D. (2008). Low-temperature specific heat spectra considering nonextensive long-range correlated quasiperiodic DNA molecules. *Physica A*, 387(22):5477 – 5482.
- Moreno, P., Raj, B., and Stern, R. (1996). A vector Taylor series approach for environment-independent speech recognition. *Proc. IEEE Internat. Conf. on Acoustics, Speech and Signal Processing*, 2:733 –736.
- Moret, M. (2011). Self-organized critical model for protein folding. *Physica A*, 390(17):3055 – 3059.

- Moyano, L. G. and Anteneodo, C. (2006). Diffusive anomalies in a long-range Hamiltonian system. *Phys. Rev. E*, 74:021118.
- Myers, C. and Rabiner, L. (1981). A level building dynamic time warping algorithm for connected word recognition. *IEEE Trans. Acoust. Speech Signal Process.*, 29(2):284 – 297.
- Nakamura, S., Fujimoto, M., and Takeda, K. (2006). CENSREC2: Corpus and evaluation environments for in car continuous digit speech recognition. *Proc. Interspeech*, pages 2330–2333.
- Nivanen, L., Méhauté, A. L., and Wang, Q. (2003). Generalized algebra within a nonextensive statistics. *Rep. Math. Phys.*, 52(3):437 – 444.
- Oikonomou, T., Provata, A., and Tirnakli, U. (2008). Nonextensive statistical approach to non-coding human DNA. *Physica A*, 387(11):2653 – 2659.
- Olemskoi, A., Shuda, I., and Borisjuk, V. (2010). Generalization of multifractal theory within quantum calculus. *Europhys. Lett.*, 89(5):50007.
- Oppenheim, A. and Schafer, R. (1968). Homomorphic analysis of speech. *IEEE Trans. Acoust.*, 16(2):221 – 226.
- Pawate, B., McMahan, M., Wiggins, R., Doddington, G., and Rajasekaran, P. (1987). Connected word recognizer on a multiprocessor system. *IEEE Internat. Conf. on Acoustics, Speech and Signal Processing*, 12:1151 – 1154.
- Picone, J. (1993). Signal modeling techniques in speech recognition. *Proceedings of the IEEE*, 81(9):1215 – 1247.
- Plastino, A., Plastino, A., Plastino, A., Miller, H., and Uys, H. (2004). Foundations of nonextensive statistical mechanics and its cosmological applications. *Astrophys. Space Sci.*, 290:275–286.
- Poruba, J. (2002). Speech enhancement based on nonlinear spectral subtraction. *Proc. IEEE Internat. Conf. Devices, Circuits and Systems*, pages T031–1 – T031–4.
- Rabiner, L. (1989). A tutorial on hidden Markov models and selected applications in speech recognition. *Proceedings of the IEEE*, 77(2):257 – 286.

- Rufiner, H. L., Torres, M. E., Gamero, L., and Milone, D. H. (2004). Introducing complexity measures in nonlinear physiological signals: Application to robust speech recognition. *Physica A*, 332:496 – 508.
- Schroeder, M. R., Atal, B. S., and Hall, J. L. (1979). Optimizing digital speech coders by exploiting masking properties of the human ear. *J. Acoust. Soc. Amer.*, 66(6):1647–1652.
- Silverman, H. and Morgan, D. (1990). The application of dynamic programming to connected speech recognition. *IEEE ASSP Magazine*, 7(3):6 –25.
- Stark, A. and Paliwal, K. (2011). MMSE estimation of log-filterbank energies for robust speech recognition. *Speech Commun.*, 53(3):403 – 416.
- Stouten, V., Van hamme, H., and Wambacq, P. (2005). Effect of phase-sensitive environment model and higher order VTS on noisy speech feature enhancement. *Proc. IEEE Internat. Conf. on Acoustics, Speech and Signal Processing*, 1:433 – 436.
- Thistleton, W., Marsh, J., Nelson, K., and Tsallis, C. (2009). q -Gaussian approximants mimic non-extensive statistical-mechanical expectation for many-body probabilistic model with long-range correlations. *Cent. Eur. J. Phys.*, 7:387–394.
- Tirnakli, U., Beck, C., and Tsallis, C. (2007). Central limit behavior of deterministic dynamical systems. *Phys. Rev. E*, 75:040106.
- Touchette, H. and Beck, C. (2005). Asymptotics of superstatistics. *Phys. Rev. E*, 71:016131.
- Tsallis, C. (1988). Possible generalization of boltzmann-gibbs statistics. *J. Stat. Phys.*, 52:479–487.
- Tsallis, C. (2002). Entropic nonextensivity: a possible measure of complexity. *Chaos Solitons Fractals*, 13(3):371 – 391.
- Tsallis, C. (2009). Nonadditive entropy: The concept and its use. *Eur. Phys. J. A.*, 40:257–266. 10.1140/epja/i2009-10799-0.
- Tsallis, C. and de Albuquerque, M. (2000). Are citations of scientific papers a case of nonextensivity? *Eur. Phys. J. B.*, 13:777–780. 10.1007/s100510050097.

- Tsallis, C. and Duarte Queirós, S. M. (2007). Nonextensive statistical mechanics and central limit theorems I-Convolution of independent random variables and q -product. *Proc. AIP Conf.*, 965:8–20.
- Udrea, R. M., Vizireanu, N., Ciochina, S., and Halunga, S. (2008). Nonlinear spectral subtraction method for colored noise reduction using multi-band Bark scale. *Signal Processing*, 88(5):1299 – 1303.
- Umarov, S., Tsallis, C., and Steinberg, S. (2008). On a q -Central limit theorem consistent with nonextensive statistical mechanics. *Milan J. Math.*, 75:307–328.
- Vignat, C. and Plastino, A. (2007). Central limit theorem and deformed exponentials. *J. Phys. A*, 40(45):F969–F978.
- Viikki, O. and Laurila, K. (1998). Cepstral domain segmental feature vector normalization for noise robust speech recognition. *Speech Commun.*, 25(13):133 – 147.
- Waibel, A., Hanazawa, T., Hinton, G., Shikano, K., and Lang, K. (1989). Phoneme recognition using time-delay neural networks. *IEEE Trans. Acoust. Speech Signal Process.*, 37(3):328 –339.
- Weili, S., Yu, M., Zhanfang, C., and Hongbiao, Z. (2009). Research of automatic medical image segmentation algorithm based on Tsallis entropy and improved PCNN. *Proc. Internat. Conf. Mechatronics and Automation*, pages 1004 –1008.
- Wilbur B. Davenport, J. (1952). An experimental study of speech-wave probability distributions. *J. Acoust. Soc. Am.*, 24(4):390–399.
- Wilk, G. and Wlodarczyk, Z. (2002). Application of nonextensive statistics to particle and nuclear physics. *Physica A*, 305(1-2):227 – 233.
- William A. Kretzschmar, J. (2009). Speech as a complex system. In *The Linguistics of Speech*. Cambridge University Press.
- Woodland, P., Odell, J., Valtchev, V., and Young, S. (1994). Large vocabulary continuous speech recognition using HTK. *IEEE Internat. Conf. Acoustics, Speech and Signal Processing*, ii:II/125 –II/128 vol.2.

- Yoma, N., McInnes, F., and Jack, M. (1998). Improving performance of spectral subtraction in speech recognition using a model for additive noise. *IEEE Trans. Speech Audio Process.*, 6(6):579–582.
- Young, S. (1996). A review of large-vocabulary continuous-speech. *IEEE Signal Process. Mag.*, 13(5):45.
- Yu, D., Deng, L., Droppo, J., Wu, J., Gong, Y., and Acero, A. (2008). Robust speech recognition using a cepstral minimum-mean-square-error-motivated noise suppressor. *Trans. Audio, Speech and Lang. Proc.*, 16(5):1061–1070.
- Zhu, Q. and Alwan, A. (2002). The effect of additive noise on speech amplitude spectra: A quantitative analysis. *IEEE Signal Process. Lett.*, 9(9):275–277.