論文 / 著書情報
Article / Book Information

Title	Feature normalization based on non-extensive statistics for speech recognition
Author	Hilman F. Pardede, Koji Iwano, Koichi Shinoda
Journal/Book name	Speech Communication, vol. 55, , pp. 587-599
Issue date	2013, 3
URL	http://www.journals.elsevier.com/speech-communication
DOI	http://dx.doi.org/10.1016/j.specom.2013.02.004
Note	このファイルは著者（最終）版です。 This file is author (final) version.

Feature normalization based on non-extensive statistics for speech recognition

Hilman F. Pardede[a,], Koji Iwano[b], Koichi Shinoda[a]

[a]Department of Computer Science, Graduate School of Information Science and Engineering, Tokyo Institute of Technology Ookayama 2-12-1, Meguro-ku, Tokyo, 152-8552 Japan

[b]Faculty of Environmental and Information Studies, Tokyo City University Ushikubo-nishi 3-3-1, Tsuzuki-ku, Yokohama, 224-8551 Japan

---

Abstract

Most compensation methods to improve the robustness of speech recognition systems in noisy environments such as spectral subtraction, CMN and MVN, rely on the fact that noise and speech spectra are independent. However, the use of limited window in signal processing may introduce a cross-term between them, which deteriorates the speech recognition accuracy. To tackle this problem, we introduce the q-logarithmic (log) spectral domain of non-extensive statistics and propose q-log spectral mean normalization (q-LSMN) which is an extension of log spectral mean normalization (LSMN) to the q-domain. The recognition experiments on a synthesized noisy speech database, the Aurora-2 database, showed that q-LSMN was consistently better than the conventional normalization methods, CMN, LSMN, and MVN. Furthermore, q-LSMN was even more effective when applied to a real noisy environment in the CENSREC-2 database. It significantly outperformed ETSI AFE front-end.

Keywords: robust speech recognition, normalization, q-logarithm, non-extensive statistics

1. Introduction

Current automatic speech recognition (ASR) systems are able to achieve good performance in quiet environments. However, their performance significantly degrades in noisy environments. The speech features are altered in the presence of noise. This causes a mismatch between quiet training conditions and recognition conditions, which are noisy. Environmental noises 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.

Robust speech recognition against noise has been an active area of research for the last few decades. A number of methods have been developed in this field. Their examples are spectral subtraction (Boll, 1979), vector Taylor series (VTS) (Moreno et al., 1996) and parallel model combination (Gales and Young, 1996). All these methods are based on an *extensive* statistics in which additivity holds.

Common features used for speech recognition, such as Mel frequency cepstral coefficient (MFCC) and perceptual linear prediction (PLP), are derived from 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 5-30 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:

$$y(t) = s(t) * h(t) + n(t). \tag{1}$$

Denoting $x(t) = s(t) * h(t)$, we can write (1) as the following:

$$y(t) = x(t) + n(t). \tag{2}$$

By taking STFT, we can represent (2) in the frequency domain as follows:

$$Y(m,k) = X(m,k) + N(m,k), \tag{3}$$

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

$$X(m,k) = |X(m,k)| \exp\left(j\theta_X(m,k)\right), \tag{4}$$

$$N(m,k) = |N(m,k)| \exp\left(j\theta_N(m,k)\right). \tag{5}$$

$|X(m,k)|$, $|N(m,k)|$ are the magnitude spectra, and $\theta_X(m,k)$, $\theta_N(m,k)$ are the phase spectra of filtered speech, i.e. the clean speech signal affected by convolutive noise only, and additive noise respectively. From (3), we obtain the power spectral representation of noisy speech as follows:

$$\begin{aligned} |Y(m,k)|^2 =& |X(m,k) + N(m,k)|^2 \\ =& |X(m,k)|^2 + |N(m,k)|^2 \\ &+ 2\mathrm{Re}\left[X(m,k)N^*(m,k)\right], \end{aligned} \tag{6}$$

where $N^*(m,k)$ is the complex conjugate of $N(m,k)$. Substituting (4) and (5) into (6), we obtain:

$$\begin{aligned}|Y(m,k)|^2 =&|X(m,k)|^2 + |N(m,k)|^2 \\ &+ 2|X(m,k)||N(m,k)| \cos\left(\theta_X(m,k) - \theta_N(m,k)\right), \end{aligned} \tag{7}$$

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. (7) is called a cross-term, which depends on the phase difference between speech and noise. This cross-term is ignored in most robust speech recognition methods under the assumption that speech and noise are uncorrelated. Although this assumption is generally valid since speech and noise are statistically independent, it does not hold when applying a short-time window (20-30 ms). Several studies have shown that the cross-term does exist in short-time power spectra (Kadambe and Boudreaux-Bartels, 1992; Jeong and Williams, 1992). The cross-term has been shown to significantly degrade the performance of speech recognition (Deng et al., 2004; Faubel et al., 2008). In addition, it is well known that a speech pattern is a complex system. In a speech pattern, various long-term correlations exist among its different spectral components in complex ways in various time scales. As a consequence, the additive relation between speech components and convolutive noise may not hold in the log spectral domain.

It is common to combine additive noise removal methods such as spectral subtraction with feature normalization methods such as cepstral mean normalization (CMN) (Furui, 1981) to remove both additive and convolutive noise. But as previously explained, speech and convolutive noise are not ad-

ditive in general, and thus the cross term exist even when the additive noise spectra are completely removed.

In this paper, we propose q-log spectral mean normalization (q-LSMN) (Pardede and Shinoda, 2011), which is an extension of the log spectral mean normalization (LSMN) (Avendano and Hermansky, 1997) to the q-log spectral domain. We further investigate the effect of q-LSMN in various conditions and analyze its property in more detail in this paper.

A few studies have already employed the non-extensive statistics for speech recognition. Rufiner et al. (2004) added Tsallis entropy, which is defined in non-extensive statistics, and its relative change to the standard MFCC features for capturing the dynamics of speech signals. Kobayashi and Imai (1984) employed the q-log function as the spectral smoother for speech features so that they are more robust especially in lower frequency regions. Ito et al. (2000) also implemented the q-log function to provide a forward masking scale for the dynamic cepstrums. In contrast to these studies, we use the q-log function to model non-additivity in noisy speech features.

The remainder of this paper is organized as follows: In Section 2, we describe some previous studies of robust speech recognition related to our study. In Section 3, we explain the influence of the cross-term in the speech features. In Section 4, we briefly review non-extensive statistics and its q-log function. In Section 5, we explain our proposed method. In Section 6, we describe the details of spectral subtraction which we implemented to remove additive noise. In Section 7, we describe the experimental setup to evaluate our proposed method. In Section 8, we present and discuss our experimental results. Section 9 concludes this paper.

2. Related studies

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 clean speech features in noisy speech by removing noise. Whereas, model compensation methods adapt 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 models each time a new type of noise is introduced. In this section, we first introduce several methods in extensive frameworks, and we describe several variants that take into account the correlation between speech and noise.

2.1. Feature enhancement

Spectral subtraction (Boll, 1979) is a popular method to remove additive noise in the spectral domain. In spectral subtraction, it is assumed that the noise spectra are known. The clean speech estimate is obtained by simply subtracting the noise spectra from the noisy speech spectra. Its performance 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. CMN (Furui, 1981) is a well established method to remove convolutive noise. It subtracts 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. Viikki and Laurila (1998) proposed mean variance normalization (MVN) which is an extension of CMN by including variance normalization.

Many advanced methods have been proposed in recent years. Moreno et al. (1996) introduced vector Taylor series (VTS) to approximate the mismatch caused by noise. In this method, the parameters of noise distribution (the mean and/or the variance) and the density distribution of noisy speech are first calculated from the distribution of clean speech and that of the observed noisy speech. The clean speech estimate is then obtained using MMSE criterion. In ETSI advanced front end (AFE) (ETSI standard doc., 2002), the two-stage mel-warped Wiener filter (Agarwal and Cheng, 1999) is implemented for removing additive noise, and blind equalization (BE) (Mauuary, 1996) is employed for removing convolutive noise.

2.2. Model compensation approach

The simplest way to adapt to the noisy environment 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, 1996) is a popular model adaptation method. In PMC, the distribution of noisy speech is estimated by using the distribution of clean speech and noise. A HMM for noisy speech is created by combining a clean speech HMM and a noise HMM. VTS is another model compensation method (Acero et al., 2000; Kim et al., 1998), which uses the statistics of noise obtained from the

Taylor expansion.

2.3. Dealing with the cross-term

Since speech and noise spectra are independent from each other, they are additive in the power spectral domain. However, the power spectrum of noisy speech includes a cross-term between them in practice, hence the additivity does not generally hold.

Several approaches have been proposed to compensate the cross-term. In spectral subtraction methods, Zhu and Alwan (2002) found that the signal-to-noise ratio (SNR)-based factor in nonlinear spectral subtraction (Berouti et al., 1979; Lockwood and Boudy, 1992) related to the missing cross-term. Some other approaches estimated the phase difference between speech and noise. Deng et al. (2004) included the cross-term in the derivation of their MMSE-based feature enhancement method, where the phase difference was estimated by assuming it followed Gaussian distribution. Li et al. (2009) also included the cross-term in VTS-based model compensation where the phase difference was a fixed value determined empirically.

3. Non-additivity in noisy speech features

In this section, we examine the effect of the cross-term in the speech features. In the log spectral domain, assuming that noise spectra are completely removed, speech and convolutive noise should be additive. However, this relation does not hold and hence the cross term between them exist in practice.

For this section, we omit the subscripts m and k. Let $|\hat{N}|^2$ be the estimate of noise power spectrum. We can obtain $|\hat{X}|^2$, which is the estimation of $|X|^2$,

as the following:

$$|\hat{X}|^2 = |Y|^2 - |\hat{N}|^2. \tag{8}$$

Substituting (7) into (8) and denoting $\theta = \theta_X - \theta_N$, we obtain:

$$|\hat{X}|^2 = |X|^2 + \left(|N|^2 - |\hat{N}|^2\right) + 2|X||N|\cos\theta. \tag{9}$$

If we assume the noise spectrum is estimated correctly, that is $|N|^2 - |\hat{N}|^2$ is zero, Eq. (9) becomes:

$$|\hat{X}|^2 = |X|^2 + 2|X||N|\cos\theta. \tag{10}$$

Meanwhile, in the power spectral domain:

$$|X|^2 = |S|^2|H|^2, \tag{11}$$

where $|S|^2$ and $|H|^2$ are the power spectra of clean speech and convolutive noise respectively. We can rewrite Eq.(10) as follows:

$$|\hat{X}|^2 = |S|^2|H|^2\chi, \tag{12}$$

where:

$$\chi = 1 + 2\frac{|N|}{|S||H|}\cos\theta. \tag{13}$$

By taking the log of of Eq. (12) we obtain:

$$\hat{\mathbf{x}} = \mathbf{s} + \mathbf{h} + \boldsymbol{\chi}, \tag{14}$$

where $\hat{\mathbf{x}}$, $\mathbf{s}$, $\mathbf{h}$, and $\boldsymbol{\chi}$ are the log of $|\hat{X}|^2$, $|S|^2$, $|H|^2$, and χ respectively. From (14), it can be seen that even though the additive noise spectra are completely removed, the cross-term still exists. The cross-term introduces $\boldsymbol{\chi}$ in the log spectral domain. Hence, the additive relation between speech and convolutive noise does not hold in this domain.

The cross-term in Eq. (13) is zero when the distance between the speech and noise components in frequency space is greater than the frequency resolution in STFT (Kadambe 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. Several studies have reported that it deteriorates the performance of ASR. Deng et al. (2004) and Zhu and Alwan (2002) showed that the clean speech estimate was not able to achieve the same accuracies as the clean speech when the cross-term was ignored, even when the noise spectra were perfectly estimated. Furthermore, Evans et al. (2006) reported that the cross-term significantly reduced the performance of speech recognition especially in the low SNR conditions.

In the conventional feature normalization methods, a normalized log spectrum is defined by:

$$\tilde{\mathbf{s}} = \mathbf{s} - \overline{\mathbf{s}}. \tag{15}$$

By doing the same process to $\hat{\mathbf{x}}$, we obtain:

$$\begin{aligned}\tilde{\hat{\mathbf{x}}} &= \hat{\mathbf{x}} - \overline{\hat{\mathbf{x}}} \\ &= (\mathbf{s} - \overline{\mathbf{s}}) + \left(\mathbf{h} - \overline{\mathbf{h}}\right) + (\boldsymbol{\chi} - \overline{\boldsymbol{\chi}}), \end{aligned} \tag{16}$$

where:

$$\overline{\boldsymbol{\chi}} = \frac{1}{M} \sum_{m=1}^{M} \boldsymbol{\chi}. \tag{17}$$

Since we assume that convolutive noise is stationary and $\mathbb{E}\{\cos\theta\} = 0$, then $\mathbf{h} = \overline{\mathbf{h}}$ and $\overline{\boldsymbol{\chi}} = 0$, and Eq. (16) becomes:

$$\tilde{\hat{\mathbf{x}}} = \tilde{\mathbf{s}} + \boldsymbol{\chi}. \tag{18}$$

From Eq. (18), it is clear that the cross-term cannot be removed even when the spectral subtraction and LSMN are ideally conducted.

4. Review of non-extensive statistics

In this section, we briefly describe the concept of non-extensive statistics. Tsallis (1988) introduced a theory of non-extensive statistics in the field of statistical mechanics. This theory generalizes Boltzmann-Gibbs statistics by utilizing the q-exponential (q-exp) function and its inverse, the q-log function. The q-exp function is defined as the following:

$$\exp_q(x) = (1 + (1-q)x)^{\frac{1}{1-q}}, \tag{19}$$

and the the q-log is defined as:

$$\log_q(x) = \frac{x^{1-q} - 1}{1 - q}. \tag{20}$$

These functions asymptotically approach exponential and natural logarithmic functions as q approaches 1. The q-log function for real number x is illustrated in Fig. 1. It can be seen that $\log_q(x)$ varies with parameter q and approaches $\log(x)$ when q is close to 1.

A special property of q-log function is its pseudo-additivity, which is introduced when $q \neq 1$ (Nivanen et al., 2003; Borges, 2004):

$$\log_q(xy) = \log_q(x) + \log_q(y) + (1 - q)\log_q(x)\log_q(y), \tag{21}$$

$$\log_q\left(\frac{x}{y}\right) = \frac{\log_q(x) - \log_q(y)}{1 + (1 - q)\log_q(y)}. \tag{22}$$

We can see that the q-log function is extensive when $q = 1$.

In this framework, entropy is redefined:

$$\begin{aligned} S_q &= -k \sum_{i=1}^{W} p_i^q \log_q p_i \\ &= k\frac{1 - \sum_{i=1}^{W} p_i^q}{q - 1}, \end{aligned} \tag{23}$$

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 subsystem's 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 subsystems of a complex system X. Then:

$$S_q(X) = S_q(A) + S_q(B) + (1-q)S_q(A)S_q(B). \tag{24}$$

The third term of Eq. (24) represents the correlation between A and B, the degree of which depends on the choice of q.

The theory of nonextensive statistics 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 Wlodarczyk, 2002), and self-gravitating systems (Jiulin, 2007). This theory has also been successfully applied to the other fields such as finance and economics (Gradojevic and Gencay, 2011), biological system (Moret, 2011), medicine (Bezerianos et al., 2003; Weili et al., 2009), and pattern recognition (Zhang and Wu, 2008).

One of the challenges in non-extensive statistics is to explain the physical meaning of the q, which is still not clear and highly debated. To our knowledge, an automatic method for optimizing q has not been found. It is empirically determined to fit the non-additive phenomena.

5. q-log spectral mean normalization

5.1. Framework

In Section 3, it was 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 speech recognition performance. 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 non-extensive framework 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 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-LMSN). 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.

The block diagram of our front-end is shown in Fig. 2. 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 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 by using our approach.

5.2. q-LSMN

In Eq. (18), 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 $|S(m,k)|^2$:

$$|\tilde{S}(m,k)|^2 = \frac{|S(m,k)|^2}{\exp\left[\frac{1}{M}\sum_{m=1}^{M}\mathbf{s}(m,k)\right]}. \tag{25}$$

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

$$\overline{\mathbf{s}}_q(k) = \frac{1}{M}\sum_{m=1}^{M}\mathbf{s}_q(m,k). \tag{26}$$

By normalizing the spectrum with its arithmetical mean of the q-logarithm

of $|S(m,k)|^2$, we obtain:

$$|\tilde{S}(m,k)|^2 = \frac{|S(m,k)|^2}{\exp_q\left[\frac{1}{M}\sum_{m=1}^{M}\mathbf{s}_q(m,k)\right]}. \tag{27}$$

For $q = 1$, Eq. (27) is the same as Eq. (25). By taking the q-log of Eq.(27) we obtain:

$$\tilde{\mathbf{s}}_q(m,k) = \frac{\mathbf{s}_q(m,k) - \overline{\mathbf{s}}_q(k)}{1 + (1-q)\overline{\mathbf{s}}_q(k)}. \tag{28}$$

Eq. (28) 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:

$$Z = |S|^2\chi. \tag{29}$$

We can rewrite Eq. (12) as follows:

$$|\hat{X}|^2 = Z|H|^2. \tag{30}$$

By taking the q-log of Eq. (30), we obtain:

$$\hat{\mathbf{x}}_q = \mathbf{z}_q + \mathbf{h}_q + (1-q)\mathbf{z}_q\mathbf{h}_q. \tag{31}$$

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. (28), and assuming convolutive noise

is stationary, we obtain:

$$\begin{aligned}\tilde{\hat{\mathbf{x}}}_q &= \frac{(\mathbf{z}_q - \overline{\mathbf{z}}_q)\,(1 + (1-q)\mathbf{h}_q)}{(1 + (1-q)\overline{\mathbf{z}}_q)\,(1 + (1-q)\mathbf{h}_q)} \\ &= \frac{\mathbf{z}_q - \overline{\mathbf{z}}_q}{1 + (1-q)\overline{\mathbf{z}}_q} \\ &= \tilde{\mathbf{z}}_q, \end{aligned} \tag{32}$$

where:

$$\overline{\mathbf{z}}_q = \frac{1}{M}\sum_{m=1}^{M} \mathbf{z}_q. \tag{33}$$

From Eq. (32) it is clear that q-LSMN is robust against convolutive noise. By taking the q-log of Eq. (29) we obtain:

$$\mathbf{z}_q = \mathbf{s}_q + \boldsymbol{\chi}_q + (1-q)\mathbf{s}_q\boldsymbol{\chi}_q, \tag{34}$$

where $\boldsymbol{\chi}_q$ is the q-log of χ. Since $\mathbb{E}\{\cos\theta\} = 0$, $\mathbb{E}\{\chi\} = 0$. 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 $0 \leq q \leq 1$. Therefore, the long term average of $\mathbf{z}_q$ equals:

$$\overline{\mathbf{z}}_q = \overline{\mathbf{s}}_q. \tag{35}$$

By substituting (34) and (35) to (32), we obtain:

$$\begin{aligned}\tilde{\hat{\mathbf{x}}}_q &= \frac{\mathbf{s}_q + \boldsymbol{\chi}_q + (1-q)\mathbf{s}_q\boldsymbol{\chi}_q - \overline{\mathbf{s}}_q}{1 + (1-q)\overline{\mathbf{s}}_q} \\ &= \frac{(\mathbf{s}_q - \overline{\mathbf{s}}_q) + \boldsymbol{\chi}_q\left(1 + (1-q)\mathbf{s}_q\right)}{1 + (1-q)\overline{\mathbf{s}}_q} \\ &= \tilde{\mathbf{s}}_q + \frac{1 + (1-q)\mathbf{s}_q}{1 + (1-q)\overline{\mathbf{s}}_q}\boldsymbol{\chi}_q. \end{aligned} \tag{36}$$

From (36), we can see that normalizing in the q-log spectral domain introduces a weight to $\boldsymbol{\chi}_q$. For the spectral valleys, i.e. $\mathbf{s}_q < \overline{\mathbf{s}}_q$, this weight is smaller than 1, hence the cross-term is smaller. On the other hand, for the spectral peaks, i.e. $\mathbf{s}_q > \overline{\mathbf{s}}_q$, 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.

6. Spectral Subtraction

We remove additive noise by implementing spectral subtraction (SS). It is formulated as follows:

$$|\hat{X}(m,k)|^2 = \max(|Y(m,k)|^2 - \alpha_m|\hat{N}(m,k)|^2, \beta|Y(m,k)|^2), \tag{37}$$

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):

$$\alpha_m = \begin{cases} 1 & \text{if } \mathrm{NSNR}_m \geq 20\mathrm{dB}, \\ 4 - \frac{3}{20}\mathrm{NSNR}_m & \text{if } -5\mathrm{dB} \leq \mathrm{NSNR}_m < 20\mathrm{dB}, \\ 4.75 & \text{if } \mathrm{NSNR}_m < -5\mathrm{dB}, \end{cases} \tag{38}$$

where NSNR_m is formulated as:

$$\mathrm{NSNR}_m = 10\log\frac{\sum_{k=1}^{K}|Y(m,k)|^2}{\sum_{k=1}^{K}|\hat{N}(m,k)|^2}. \tag{39}$$

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:

$$|\tilde{N}(m,k)|^2 = \gamma|\hat{N}(m-1,k)|^2 + \frac{1-\gamma}{1-\lambda}\left(|\ddot{Y}(m,k)|^2 - |\ddot{Y}(m-1,k)|^2\right), \quad (40)$$

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

$$|\ddot{Y}(m,k)|^2 = \delta|\ddot{Y}(m-1,k)|^2 + (1-\delta)\,|Y(m,k)|^2. \quad (41)$$

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:

$$\xi_{\text{rel}}(m,k) = \frac{\xi(m,k) - \xi_{\min}(m,k)}{\xi_{\max}(m,k) - \xi_{\min}(m,k)}, \quad (42)$$

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:

$$|\hat{N}(m,k)|^2 = \begin{cases} |\hat{N}(m-1,k)|^2 & \text{if } \xi_{\text{rel}}(m,k) < T, \\ |\tilde{N}(m,k)|^2 & \text{else,} \end{cases} \quad (43)$$

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

7. Experimental Setup

7.1. 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.

7.1.1. 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.

7.1.2. The CENSREC-2 database

This database 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 was 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 low and high speed conditions and the HF microphones were used. In this condition, both channel and additive noise are the sources of mismatch.

7.2. 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 3 Gaussian mixtures. Two pause models were used: "sil" and "sp". The "sil" model consisted of 3 states. Each state in the "sil" model has 6 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.

8. 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.

8.1. 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 them was calculated:

$$D_{\mathrm{LS}} = \frac{1}{KM} \sum_{m=1}^{M} \sum_{k=1}^{K} \sqrt{\left(\log \frac{|S(m,k)|^2}{|\hat{S}(m,k)|^2} \right)^2}. \tag{44}$$

Fig 3 shows the LSD 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. 37 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. (18). 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.

Fig. 4 and 5 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 $\mathbf{s}_q < \overline{\mathbf{s}}_q$ 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.

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 6. We found that the performance of speech recognition improved 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.

Table 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. (10), 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).

8.2. Evaluation of q-LSMN without spectral subtraction

Figure 7 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 8 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 8.1. Fig. 9 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 10 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 11 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 ($q = 1$). Larger improvements were obtained in Condition 3 and 4 where there are channel differences between training and testing.

Table 2 summarizes 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.

8.3. 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. 7 and 10). For both databases, the optimum q-value was different from that without SS. When SS was implemented, $q = 0.8$ was 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.

Table 3 summarizes 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 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 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.

9. 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.

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. While we proved that the use of q different from 1 is effective in robust speech recognition, the meaning of q in speech processing is still not clear and should be investigated in future.

The optimal q-value may be different in different frequency bands since noise affects each frequency band differently. While we used the same values for all the bands in this study, the optimization of q-values to each band might be promising. Another interesting topic would be the implementation of other compensation methods in the q-log domain.

Acknowledgement

This work is supported by Grant in Aid for Challenging Exploratory Research No. 24650079

References

Acero, A., Deng, L., Kristjansson, T., 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., Cheng, Y.M., 1999. Two-stage mel-warped wiener filter for robust speech recognition. Proc. IEEE Workshop on Automatic Speech Recognition and Understanding , 12–15.

Avendano, C., Hermansky, H., 1997. On the effects of short-term spectrum smoothing in channel normalization. IEEE Trans. Speech Audio Process. 5, 372–374.

Berouti, M., Schwartz, R., 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., 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, 113–120.

Borges, E.P., 2004. A possible deformed algebra and calculus inspired in nonextensive thermostatistics. Physica A, 340, 95–101.

Cohen, I., 2005. Relaxed statistical model for speech enhancement and a priori snr estimation. IEEE Trans. Speech Audio Process. 13, 870 – 881.

Deng, L., Droppo, J., 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, 133 – 143.

Doblinger, G., 1995. Computationally efficient speech enhancement by spectral minima tracking in subbands. Proc. Eurospeech , 1513–1516.

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., 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., Klakow, D., 2008. A phase-averaged model for the relationship between noisy speech, clean speech and noise in the log-mel domain. Proc. Interspeech , 553–556.

Furui, S., 1981. Cepstral analysis technique for automatic speaker verification. IEEE Trans. Acoust. Speech Signal Process. 29, 254 – 272.

Gales, M., Young, S., 1996. Robust continuous speech recognition using parallel model combination. IEEE Trans. Speech Audio Process. 4, 352 –359.

Gradojevic, N., Gencay, R., 2011. Financial applications of nonextensive entropy [applications corner]. IEEE Signal Process. Mag. 28, 116–141.

Hermansky, H., 1990. Perceptual linear predictive (PLP) analysis of speech. J. Acoust. Soc. Amer. 87, 1738–1752.

Hirsch, H., 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., Pearce, D., 2000. The aurora experimental framework for the performance evaluation of speech recognition systems under noisy conditions. Proc. ISCA ITRW ASR2000 , 181–188.

Itahashi, S., 1990. Recent speech database projects in japan. Proc. IEEE Internat. Conf. Spoken Language Processing , 1081–1084.

Ito, Y., Matsumoto, H., Yamamoto, K., 2000. Forward masking on a generalized logarithmic scale for robust speech recognition. Proc. Interspeech , 530–533.

Jeong, J., Williams, W., 1992. Mechanism of the cross-terms in spectrograms. IEEE Trans. Signal Process. 40, 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., 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, 2498 –2517.

Kim, C., Stern, R.M., 2009. Feature extraction for robust speech recognition using a power-law nonlinearity and power-bias subtraction. Proc. Interspeech , 28–31.

Kim, D.Y., Un, C.K., Kim, N.S., 1998. Speech recognition in noisy environments using first-order vector taylor series. Speech Commun. 24, 39 – 49.

Kobayashi, T., Imai, S., 1984. Spectral analysis using generalized cepstrum. IEEE Trans. Acoust. Speech Signal Process. 32, 1087 – 1089.

Li, J., Deng, L., Yu, D., Gong, Y., Acero, A., 2009. A unified framework of hmm adaptation with joint compensation of additive and convolutive distortions. Computer Speech and Language 23, 389 – 405.

Lockwood, P., 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, 215 – 228.

Mauuary, L., 1996. Blind Equalization for robust telephone based speech recognition. Proc. EUSIPCO .

McAuley, J., Ming, J., Stewart, D., Hanna, P., 2005. Subband correlation and robust speech recognition. IEEE Trans. Speech Audio Process. 13, 956 – 964.

Ming, J., O'Boyle, P., McMahon, J., 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.

Moreno, P., Raj, B., 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, 3055 – 3059.

Nakamura, S., Fujimoto, M., Takeda, K., 2006. Censrec2: corpus and evaluation environments for in car continuous digit speech recognition. Proc. Interspeech , 2330–2333.

Nivanen, L., Méhauté, A.L., Wang, Q., 2003. Generalized algebra within a nonextensive statistics. Rep. Math. Phys. 52, 437 – 444.

Olemskoi, A., Shuda, I., Borisyuk, V., 2010. Generalization of multifractal theory within quantum calculus. Europhys. Lett. 89, 50007.

Pardede, H.F., Shinoda, K., 2011. Generalized-log spectral mean normalization for speech recognition. Proc. Interspeech , 1645–1648.

Plastino, A., Plastino, A., Plastino, A., Miller, H., Uys, H., 2004. Foundations of nonextensive statistical mechanics and its cosmological applications. Astrophys. Space Sci. 290, 275–286.

Rufiner, H.L., Torres, M.E., Gamero, L., 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., Hall, J.L., 1979. Optimizing digital speech coders by exploiting masking properties of the human ear. J. Acoust. Soc. Amer. 66, 1647–1652.

Tsallis, C., 1988. Possible generalization of boltzmann-gibbs statistics. J. Stat. Phys. 52, 479–487.

Viikki, O., Laurila, K., 1998. Cepstral domain segmental feature vector

normalization for noise robust speech recognition. Speech Commun. 25, 133 – 147.

Weili, S., Yu, M., Zhanfang, C., Hongbiao, Z., 2009. Research of automatic medical image segmentation algorithm based on tsallis entropy and improved pcnn. Proc. Internat. Conf. Mechatronics and Automation , 1004 –1008.

Wilk, G., Wlodarczyk, Z., 2002. Application of nonextensive statistics to particle and nuclear physics. Physica A 305, 227 – 233.

Zhang, Y., Wu, L., 2008. Pattern recognition via pcnn and tsallis entropy. Sensors 8, 7518–7529.

Zhu, Q., Alwan, A., 2002. The effect of additive noise on speech amplitude spectra: a quantitative analysis. IEEE Signal Process. Lett. 9, 275 – 277.

Table 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

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

Methods	Aurora-2				CENSREC-2				
	Set A	Set B	Set C	Ave.	Cond 1	Cond 2	Cond 3	Cond 4	Ave.
No compensation	65.8	68.6	60.9	**65.9**	85.5	78.7	46.2	40.4	**62.7**
q-LSMN ($q = 0.7$)	72.2	76.9	71.1	**73.8**	88.7	84.4	70.1	57.9	**75.3**
q-LSMN ($q = 0.4$)	67.2	71.7	64.8	**68.5**	88.6	85.2	73.5	63.3	**77.7**
LSMN	64.7	70.2	62.8	**66.5**	88.3	83.3	64.4	48.7	**71.1**
CMN	65.7	70.5	63.9	**67.3**	85.9	77.7	52.9	38.9	**63.6**
MVN	68.3	69.3	64.9	**68.0**	85.5	82.5	63.2	46.0	**69.3**

Table 3: The word accuracy (%) of q-LSMN and the other methods after spectral subtraction for the Aurora-2 task and the CENSREC-2 task.

Methods	Aurora-2				CENSREC-2				
	Set A	Set B	Set C	Ave.	Cond 1	Cond 2	Cond 3	Cond 4	Ave.
SS	79.0	76.7	75.9	**77.5**	86.7	79.9	53.2	47.0	**66.7**
SS + q-LSMN ($q = 0.8$)	78.5	80.7	78.1	**79.3**	89.0	85.9	75.9	68.1	**79.7**
SS + q-LSMN ($q = 0.5$)	75.8	78.0	74.4	**76.4**	89.2	86.1	78.5	71.6	**81.4**
SS + LSMN	76.2	79.3	75.2	**77.2**	88.9	84.7	72.7	62.2	**77.1**
SS + CMN	76.2	78.4	75.2	**76.9**	86.7	79.7	59.5	47.6	**68.4**
SS + MVN	74.3	74.3	71.8	**73.8**	86.7	83.0	65.9	51.8	**71.8**
ETSI AFE	80.1	82.1	79.5	**80.8**	85.6	80.3	59.9	53.3	**69.8**

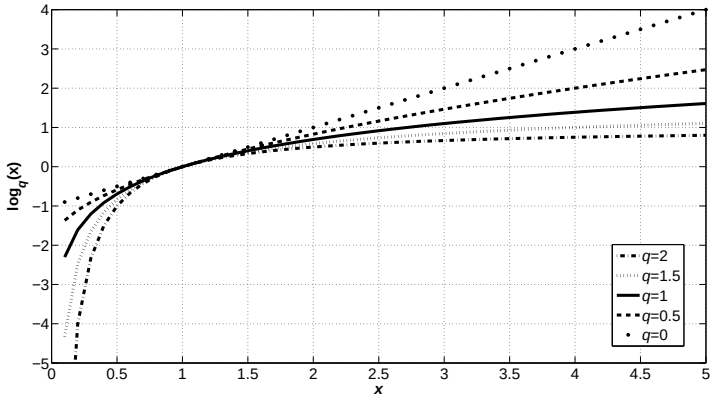

Figure 1: The q-logarithmic function of real variable x for different q-values.

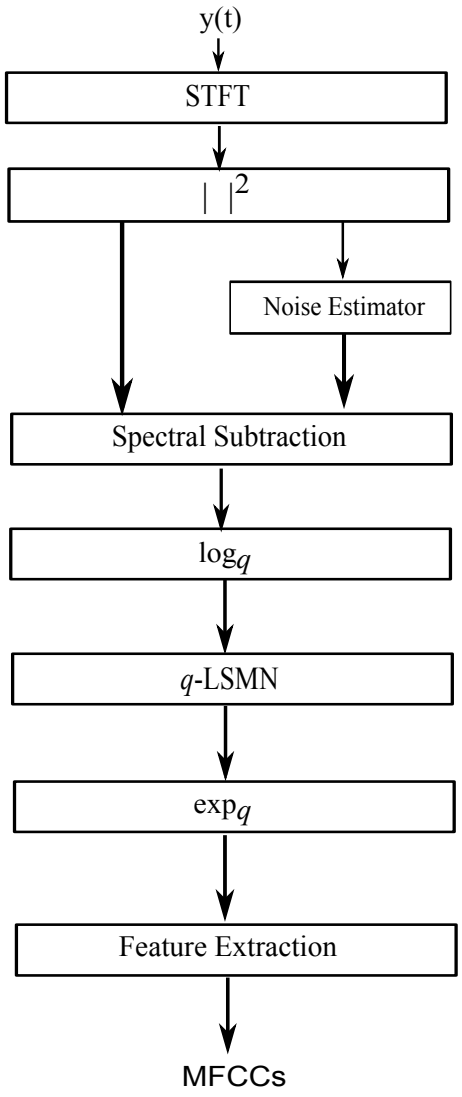

Figure 2: Block diagram of the proposed front-end.

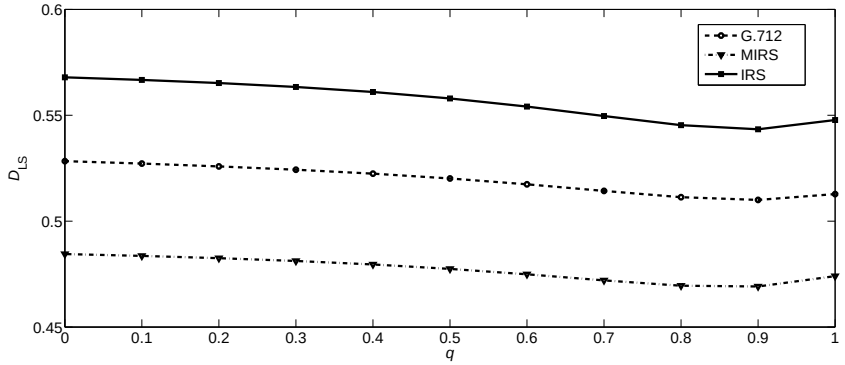

Figure 3: The log spectral distance between "filtered speech", speech affected by convolutive noise, and the clean speech after applying q-LSMN.

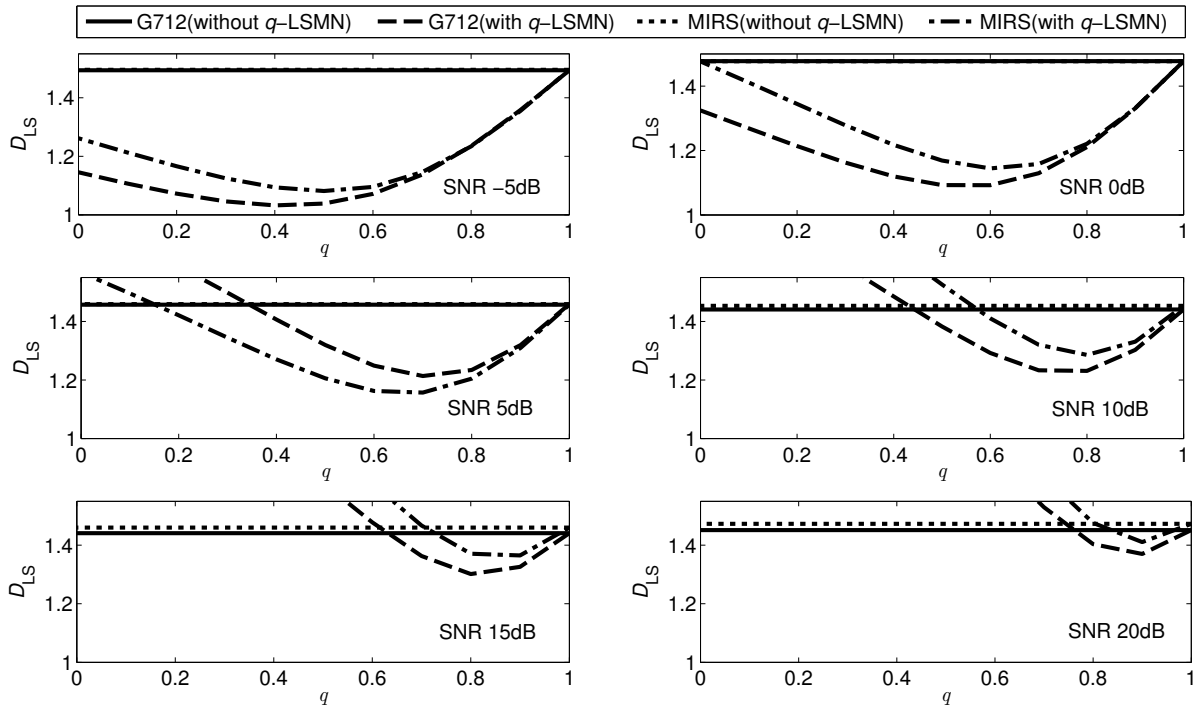

Figure 4: The log spectral distance between the clean speech estimate and the clean speech with and without applying q-LSMN for the spectral valleys. The clean speech estimate is obtained after employing spectral subtraction and normalization on the noisy speech spectra. In this figure, babble noise is used as additive noise.

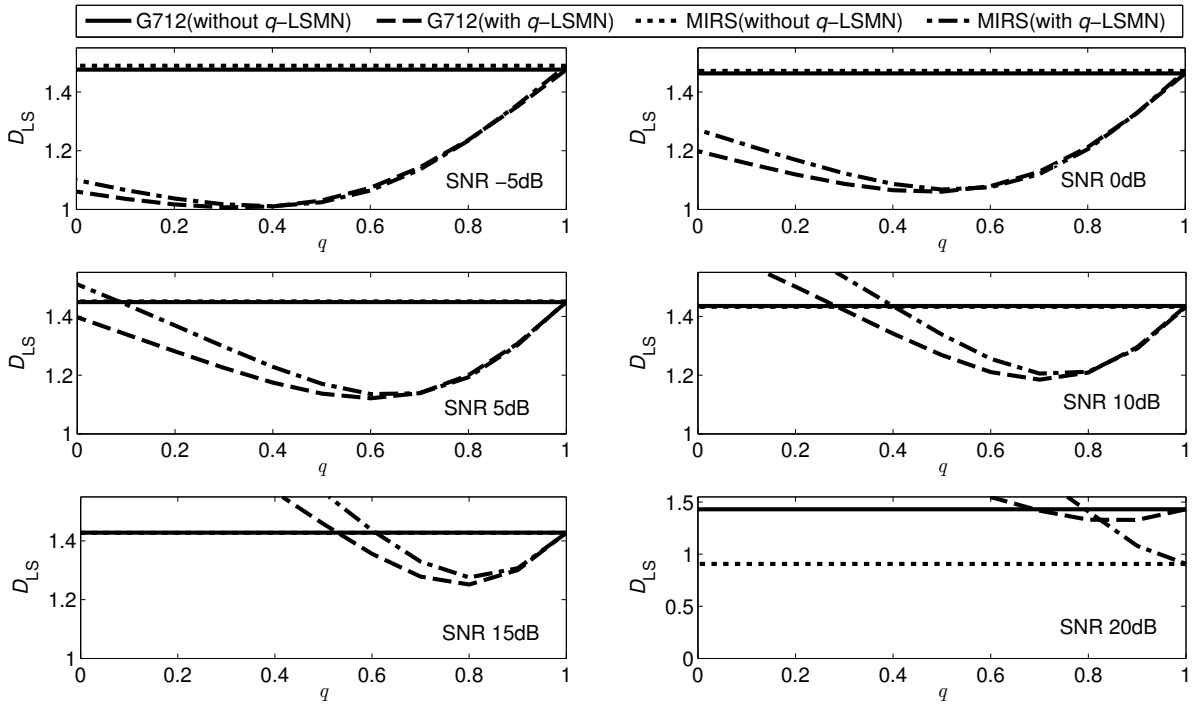

Figure 5: The log spectral distance between the clean speech estimate and the clean speech with and without applying q-LSMN for the spectral valleys. The clean speech estimate is obtained after employing spectral subtraction and normalization on the noisy speech spectra. In this figure, white noise is used as additive noise.

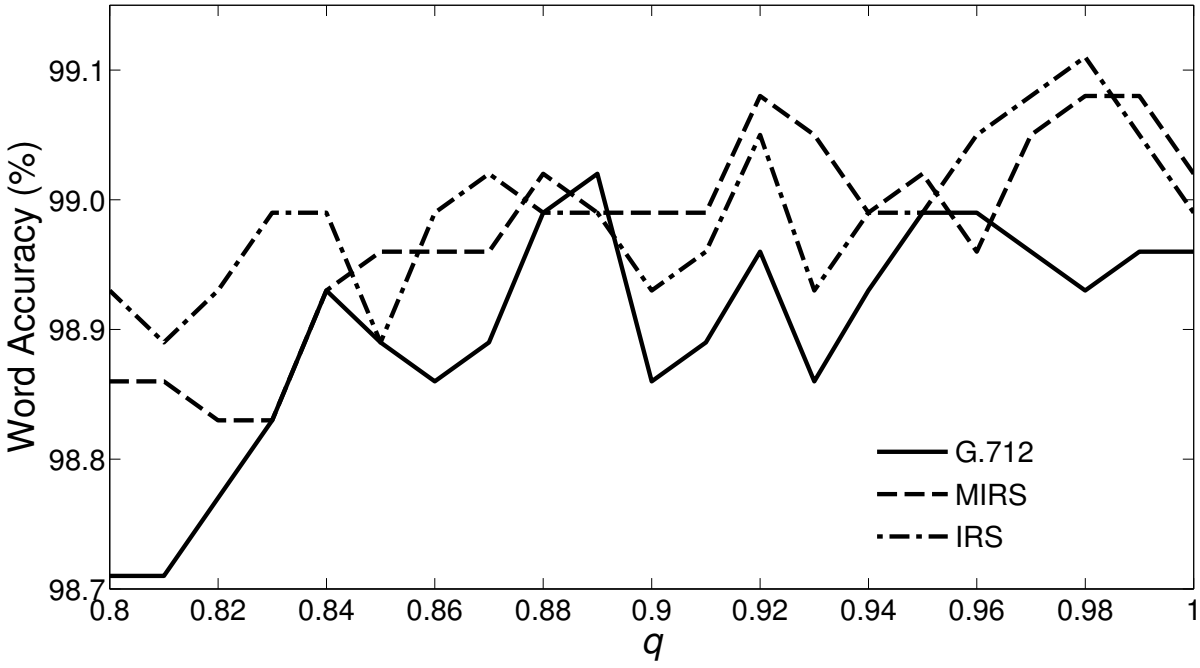

Figure 6: The word accuracies (%) of q-LSMN for filtered speech, i.e speech contaminated with convolutive noise only.

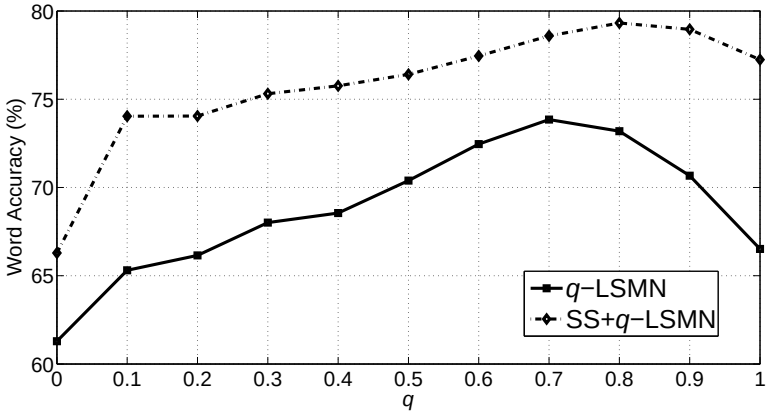

Figure 7: The word accuracies (%) of q-LSMN with and without spectral subtraction for the Aurora-2 task for different q values.

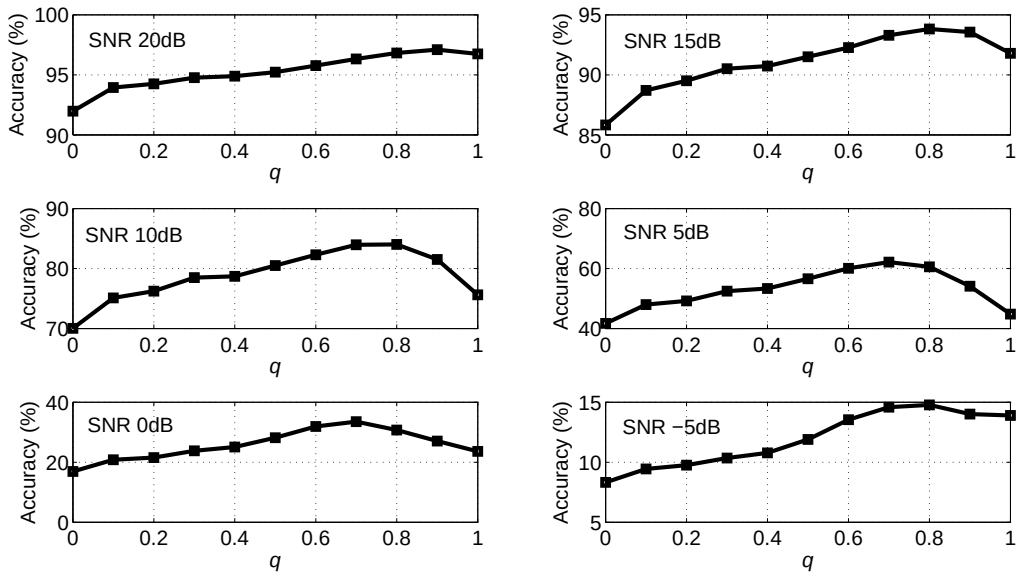

Figure 8: The word accuracies (%) of q-LSMN for various SNR conditions in the Aurora-2 task.

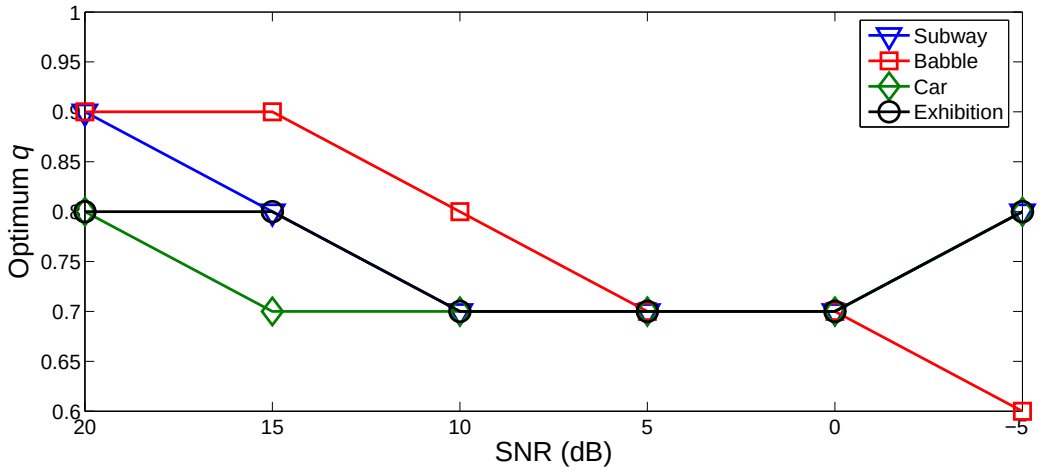

Figure 9: The optimum q-values for four types of noise in Test Set A of the Aurora-2 database for various SNR conditions.

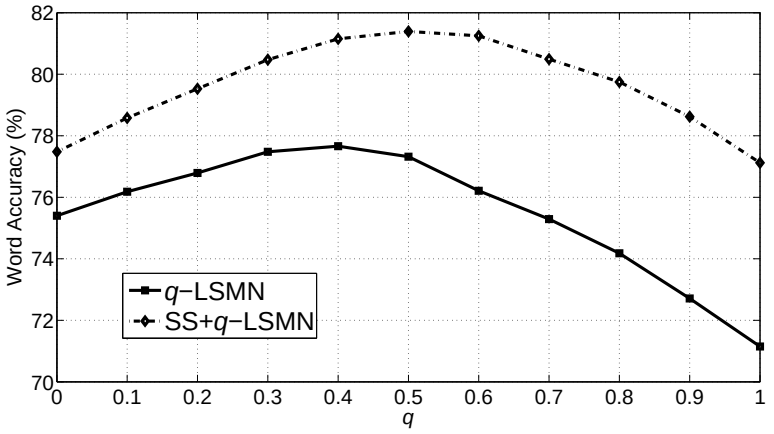

Figure 10: The word accuracies (%) of q-LSMN with and without spectral subtraction for CENSREC-2 task with different q values.

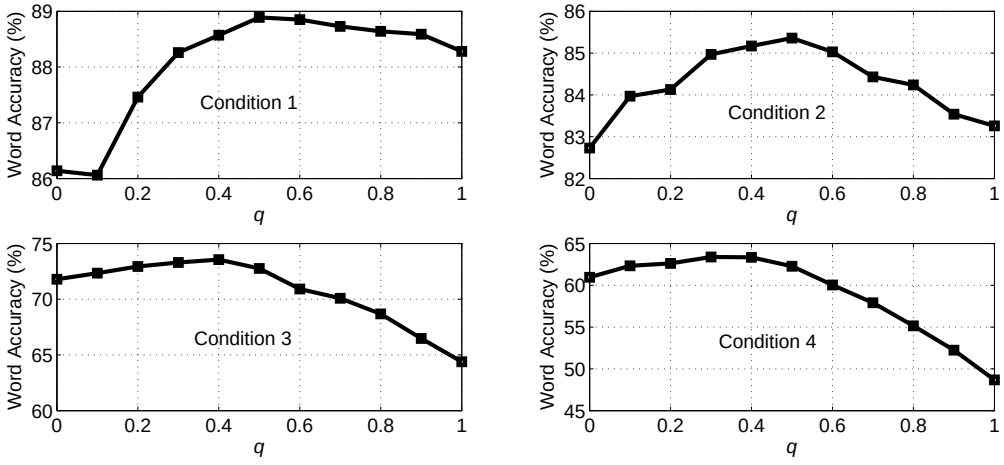

Figure 11: The word accuracies (%) of q-LSMN for CENSREC-2 task for each condition evaluation with different q values.