PAPER

Spectral Subtraction Based on Non-extensive Statistics for Speech Recognition

Hilman PARDEDE[†*a)], ***Nonmember***, **Koji IWANO**[††b)], ***Member***, ***and*** **Koichi SHINODA**[†c)], ***Senior Member***

SUMMARY Spectral subtraction (SS) is an additive noise removal method which is derived in an extensive framework. In spectral subtraction, it is assumed that speech and noise spectra follow Gaussian distributions and are independent with each other. Hence, noisy speech also follows a Gaussian distribution. Spectral subtraction formula is obtained by maximizing the likelihood of noisy speech distribution with respect to its variance. However, it is well known that noisy speech observed in real situations often follows a heavy-tailed distribution, not a Gaussian distribution. In this paper, we introduce a q-Gaussian distribution in the non-extensive statistics to represent the distribution of noisy speech and derive a new spectral subtraction method based on it. We found that the q-Gaussian distribution fits the noisy speech distribution better than the Gaussian distribution does. Our speech recognition experiments using the Aurora-2 database showed that the proposed method, q-spectral subtraction (q-SS), outperformed the conventional SS method.

key words: *robust speech recognition, spectral subtraction, Gaussian distribution, q-Gaussian, maximum likelihood*

1. Introduction

The performance of speech recognition degrades significantly in the presence of background noise. Spectral subtraction (SS) is often implemented to remove the additive background noise [1]. Spectral subtraction (SS) is one popular method to remove additive noise [2]. It is basically a variance estimator which is derived in the extensive framework. In this framework, speech and noise spectra are assumed to follow Gaussian distributions and are uncorrelated with each other. Hence, the noisy speech spectra also follow Gaussian distributions. The spectral subtraction formula is derived by maximizing the likelihood of the noisy speech distribution [3].

Even though the distribution of the speech spectrum may approximate a Gaussian distribution when a very long window is employed, it does not follow the Gaussian distribution for short-time window [4] but show a heavy-tailed distributions instead. Therefore, it is not surprising that spectral subtraction has limitations and may not give sufficiently high performance [5]. A weighting factor is often introduced to improve its performance as in nonlinear spectral subtraction (NSS) [6], [7]. However, this factor is decided heuristically. Several other distributions such as Laplace [8] and Gamma [9], [10] distributions have been used instead of Gaussian distributions.

Recently, a theory of non-extensive statistics has been introduced to explain several phenomena in complex systems [11]. This framework uses Tsallis entropy, which is a generalization of Shannon entropy. By maximizing Tsallis entropy, a q-Gaussian distribution can be obtained. This distribution can represent a heavy-tailed distribution. It has successfully represented many phenomena in complex systems in statistical mechanics, economics, finance, biology, astronomy and machine learning.

In this paper, we propose q-spectral subtraction (q-SS) [12], which is a spectral subtraction method derived in the non-extensive statistics. In this method, we assume noisy speech spectrum follows a q-Gaussian distribution and derive q-SS in a similar way as spectral subtraction is derived. We further analyze the performance of q-SS under various conditions in more detail and derive a way to optimize the parameter in q-SS in this paper.

The remainder of this paper is organized as follows. In Sect. 2, we explain how the spectral subtraction is derived. We briefly describe the q-Gaussian distribution in Sect. 3. In Sect. 4, our proposed method, q-spectral subtraction, is explained. The experimental setup and results are described and discussed in Sects. 5 and 6 respectively. Section 7 concludes this paper.

2. Spectral Subtraction

2.1 Derivation

Spectral subtraction is perhaps one of the most popular methods to remove additive background noise. In spectral subtraction, the estimate of the clean speech spectrum is obtained by simply subtracting the noisy spectrum with the estimate of the noise spectrum. From the statistical point of view, SS is a variance estimator assuming noisy speech spectrum follows a Gaussian distribution [3].

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

Manuscript received March 8, 2013.

[†] The authors are with the Tokyo Institute of Technology, Tokyo, 152–8552 Japan.

[††] The author is with the Tokyo City University, Tokyo, 158-0087 Japan.

[*] Presently, with the Research Center for Informatics-Indonesian Institute of Science.

a) E-mail: hilman@ks.cs.titech.ac.jp

b) E-mail: iwano@tcu.ac.jp

c) E-mail: shinoda@cs.titech.ac.jp

DOI: 10.1587/transinf.E96.D.1774

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:

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

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:

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

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 $|N_f|^2 = \tau(f)$. Equation (2) becomes:

$$|\hat{X}(f)|^2 = |Y_f|^2 - |N_f|^2. \tag{3}$$

Equation (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).

2.2 Nonlinear Spectral Subtraction

The simplicity of spectral subtraction comes with a price. The inaccuracy of noise estimation causes distortions and information losts 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. [6]. They introduce an oversubtraction factor, α, and the spectral subtraction formula becomes:

$$|\hat{X}_f|^2 = |Y_f|^2 - \alpha |N_f|^2. \tag{4}$$

Since the introduction of α makes the subtraction nonlinear, it is called nonlinear spectral subtraction (NSS). Zhu and Alwan [13] 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 [7], [14], [15]. They are basically modifications of the Eq. (4). In this paper, we determine α for each frequency bin, denoting α_f, using the following relation [6]:

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

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

$$\Phi_f = 10 \log \gamma_f, \tag{6}$$

where:

$$\gamma_f = \frac{|Y_f|^2}{|\hat{N}_f|^2}. \tag{7}$$

In this paper, we denote NSS of which α_f is determined using Eq. (5) as NSS_n. When a single and constant α_f is used for all spectra, we denote it as NSS_c.

3. *Q*-Gaussian Distribution

Recently, Tsallis has introduced a theory of non-extensive statistics in the field of statistical mechanics [11]. This theory generalizes Boltzmann-Gibbs statistics by utilizing q-exponential function:

$$\exp_q(x) = (1 + (1-q)x)^{\frac{1}{1-q}}, \tag{8}$$

and its inverse, q-logarithmic function:

$$\log_q(x) = \frac{x^{1-q} - 1}{1-q}. \tag{9}$$

These functions asymptotically approach exponential and natural logarithmic functions respectively as q approaches 1. They are non-extensive when $q \neq 1$ [16]. In the non-extensive framework, entropy is redefined:

$$S_q = -k \int p_i(x) \log_q p_i(x). \tag{10}$$

This entropy is called Tsallis entropy. It is a generalization of Shannon entropy.

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 zero mean and variance λ_q is defined by:

$$P_q(X) = \frac{A_q B_q}{\sqrt{\lambda_q}} \exp_q\left(-\frac{B_q^2 |X|^2}{\lambda_q}\right), \tag{11}$$

where A_q is a normalization term and defined as:

$$A_q = \begin{cases} \dfrac{\Gamma\left(\frac{5-3q}{2-2q}\right)}{\Gamma\left(\frac{2-q}{1-q}\right)} \sqrt{\dfrac{1-q}{\pi}} & -\infty < q < 1 \\ \dfrac{1}{\sqrt{\pi}} & q = 1 \\ \dfrac{\Gamma\left(\frac{1}{q-1}\right)}{\Gamma\left(\frac{3-q}{2q-2}\right)} \sqrt{\dfrac{q-1}{\pi}} & 1 < q < 3, \end{cases} \tag{12}$$

and B_q is a scaling factor and in a normalized distribution $B_q = \frac{1}{\sqrt{3-q}}$. Figure 1 shows the probability distributions

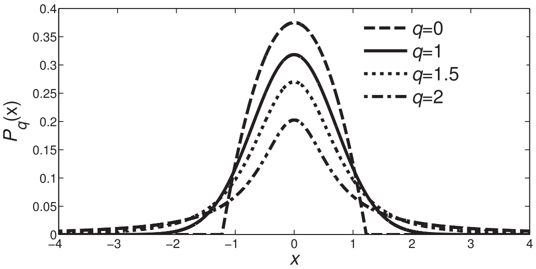

Fig. 1 q-Gaussian distribution for several q.

of q-Gaussian for several q-values. The q-Gaussian distribution is a compact support distribution when $q < 1$ and a heavy-tailed distribution when $1 < q < 3$. It is identical with the Gaussian distribution when $q = 1$.

In this non-extensive framework, the q-value is used to represent the degree of complexity [17] of a system. However, up to our knowledge, an automatic method to optimize q does not yet exist. In the implementation, it is usually chosen empirically.

4. Q-Spectral Subtraction

4.1 Derivation

When speech and noise are Gaussian random variables, noisy speech can still be a Gaussian even when speech and noise are correlated [18]. However, the short-time speech spectra are not likely to follow Gaussian distributions [4], [19]. 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 [20]–[22] or the sum of independent N q-Gaussian random variables for small number of N [23]. 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 [24], [25].

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 $Y_R = \mathrm{Re}(Y_f)$ and $Y_I = \mathrm{Im}(Y_f)$ 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:

$$P_q(Y_R) = \frac{\sqrt{2}A_q B_q}{\sqrt{\nu_q(f)}} \exp_q\left(-\frac{2B_q^2|Y_R|^2}{\nu_q(f)}\right), \tag{13}$$

$$P_q(Y_I) = \frac{\sqrt{2}A_q B_q}{\sqrt{\nu_q(f)}} \exp_q\left(-\frac{2B_q^2|Y_I|^2}{\nu_q(f)}\right). \tag{14}$$

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 [10]. Then, the distribution for noisy speech is formulated as follows:

$$P_q(Y_f) = \frac{2A_q^2 B_q^2}{\nu_q(f)} \exp_q\left(-\frac{2B_q^2|Y_f|^2}{\nu_q(f)}\right). \tag{15}$$

Equation (15) is identical with Eq. (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:

$$\hat{\sigma}_q(f) = \frac{2(2-q)}{3-q}|Y_f|^2 - \tau_q(f). \tag{16}$$

Since $\sigma_q(f) = |X_f|^2$ and $\tau_q(f) = |N_f|^2$, Eq. (16) becomes:

$$|\hat{X}_f|^2 = \frac{2(2-q)}{3-q}|Y_f|^2 - |N_f|^2. \tag{17}$$

Equation (17) is the q-spectral subtraction (q-SS) formula. It is the same as LSS when $q = 1$.

4.2 Relation to Nonlinear Spectral Subtraction

In this section, we relate q-SS with NSS in Eq. (4). Denoting $\upsilon(q) = \frac{2(2-q)}{3-q}$, we can rewrite Eq. (17) as follows:

$$|\hat{X}_f|^2 = \upsilon(q)|Y_f|^2 - |N_f|^2. \tag{18}$$

By dividing Eq. (18) with $\upsilon(q)$, we obtain:

$$\frac{1}{\upsilon(q)}|\hat{X}_f|^2 = |Y_f|^2 - \frac{1}{\upsilon(q)}|N_f|^2. \tag{19}$$

Since scaling does not affect the performance of speech recognition, we can relate α in Eq. (4) with Eq. (19) as follows:

$$\begin{aligned}\alpha &= \frac{1}{\upsilon(q)},\\ &= \frac{3-q}{2(2-q)}.\end{aligned} \tag{20}$$

Based on Eq. (20), 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.

4.3 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

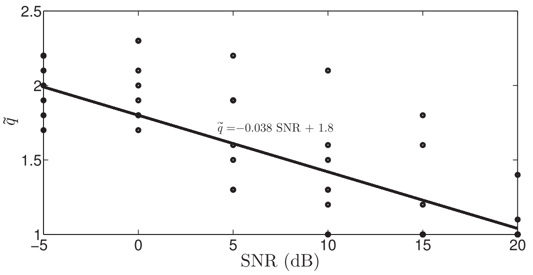

Fig. 2 The scatter plot between $\tilde{q}$, the q-value that fits the empirical distribution of noisy speech based on the MMSE criterion, and the SNR condition. The solid line is the relation between $\tilde{q}$ and the SNR based on linear regression method.

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* database, i.e. the ratio between clean speech and noise. This is different from Φ_f in Eq. (6) which is the *a posteriori* SNR. Based on the data, we obtain its variance and $S_i(Y_R)$ where $i = 1, 2, \ldots, n$ 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:

$$\tilde{q} = \underset{q}{\operatorname{argmin}} \frac{1}{n} \sum_{i=1}^{n} \left(S(Y_{R_i}) - P_q(Y_{R_i}) \right)^2 . \tag{21}$$

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

Figure 2 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 −5 dB) and $q = 1$ when the SNR is high (higher than 20 dB). These are the same limits used in NSS. The relation between the SNR and $\tilde{q}$ is formulated as:

$$\tilde{q} = \begin{cases} 1 & \text{if } \Psi_f(m) \geq 20\,\text{dB}, \\ -0.038\Psi_f(m) + 1.8 & \text{if } -5\,\text{dB} \leq \Psi_f(m) < 20\,\text{dB}, \\ 1.88 & \text{if } \Psi_f < -5\,\text{dB}, \end{cases} \tag{22}$$

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. (7), which is the ratio between noisy speech and noise. It is calculated using the following formula:

$$\Psi_f(m) = 10 \log \xi_f(m), \tag{23}$$

where:

$$\xi_f(m) = \frac{|X_f|^2}{|N_f|^2}. \tag{24}$$

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 [26]. In the maximum likelihood method, $\xi_f(m)$ is estimated based on L past observations of the noisy speech spectra $\mathbf{Y}_f(m) = \left\{Y_f(m), Y_f(m-1), \ldots, Y_f(m-L+1)\right\}$, 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:

$$p(\mathbf{Y}_f(m)|\sigma(f), \tau(f)) = \prod_{i=0}^{L-1} \frac{1}{\phi\left(\sigma(f) + \tau(f)\right)} \exp\left(-\frac{Y_f^2(m-i)}{\sigma(f) + \tau(f)} \right), \tag{25}$$

where $\sigma(f)$ is the clean speech variance. By maximizing Eq. (25) with respect to $\sigma(f)$, we obtain:

$$\hat{\sigma}(f) = \max\left(\frac{1}{L} \sum_{i=0}^{L-1} Y_f^2(m-j) - \tau_f(m), 0 \right). \tag{26}$$

By dividing both side of Eq. (26) by $\tau_f(m)$, we obtain:

$$\hat{\xi}_f(m) = \max\left[\frac{1}{L} \sum_{i=0}^{L-1} \gamma_f(m-j) - 1, 0 \right], \tag{27}$$

where $\gamma_f(m)$ is the *a posteriori* SNR given in Eq. (7). In practice, Eq. (27) is replaced by recursive operation:

$$\hat{\xi}_f(m) = \max\left[\bar{\gamma}_f(m) - 1, 0 \right], \tag{28}$$

where:

$$\bar{\gamma}_f(m) = a\bar{\gamma}_f(m-1) + (1-a)\frac{\gamma_f(m)}{b}, \tag{29}$$

In this paper, we apply $a = 0.725$ and $b = 2$. We denote q-SS that use the relation in Eq. (22) as q-SS_m.

4.4 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 [27] to estimate noise. In this method, the noise power spectrum is pre-estimated using the following formula:

$$|\tilde{N}_f(m)|^2 = \gamma|\hat{N}_f(m-1)|^2 + \frac{1-\gamma}{1-\lambda}\left(|\ddot{Y}_f(m)|^2 - |\ddot{Y}_f(m)|^2 \right), \tag{30}$$

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

$$|\ddot{Y}_f(m)|^2 = \delta|\ddot{Y}_f(m-1)|^2 + (1-\delta)\,|Y_f(m)|^2. \tag{31}$$

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 [28]. 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:

$$\zeta_f^{\mathrm{rel}}(m) = \frac{\zeta_f(m) - \zeta_f^{\min}(m)}{\zeta_f^{\max}(m) - \zeta_f^{\min}(m)}, \tag{32}$$

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:

$$|\hat{N}(m,k)|^2 = \begin{cases} |\hat{N}_f(m-1)|^2 & \text{if } \xi_f^{\mathrm{rel}}(m) < T, \\ |\tilde{N}_f(m)|^2 & \text{else,} \end{cases} \tag{33}$$

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

4.5 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.

$$|\hat{X}_f|^2 = \beta|Y_f|^2 \qquad \text{if } |\hat{X}_f|^2 < \beta|Y_f|^2. \tag{34}$$

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

5. Experimental Setup

Our proposed method was evaluated in speech recognition experiments using the Aurora-2 database [29]. In this database, eight types of noise: subway, babble, car, exhibition hall, restaurant, street, airport and train station, were added to clean speech artificially. It has two training conditions: clean-condition and multi-condition. In this paper, we used the clean condition training data for training the acoustic model. For testing, this database provides three test sets: A, B and C where noise is added at SNRs of 20 dB, 15 dB, 5 dB, 0 dB and −5 dB.

We used 38 dimensional MFCC features: 12 static features, their 1^{st}-order and 2^{nd}-order derivatives, Δ log energy and ΔΔ log energy. An HMM-based decoder is used for speech recognition. Each digit is modeled by an HMM with 16 states, left-to-right, with three Gaussian mixtures for each state.

For evaluation measure, we used a word accuracy rate. For the Aurora-2 database, the average accuracy denotes the average over SNR 0 dB to 20 dB.

6. Experiment Results and Discussions

6.1 *Q*-Gaussian Representation of Noisy Speech

We investigated the q-Gaussianity of noisy speech in a similar way explained in Sect. 4.3. 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.

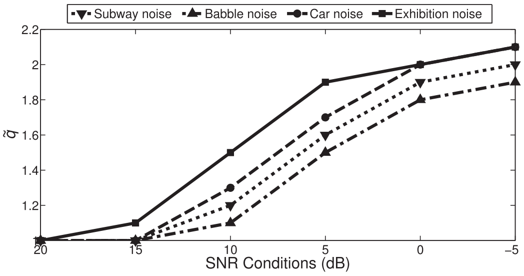

Fig. 3 The optimal q-values for different SNR conditions.

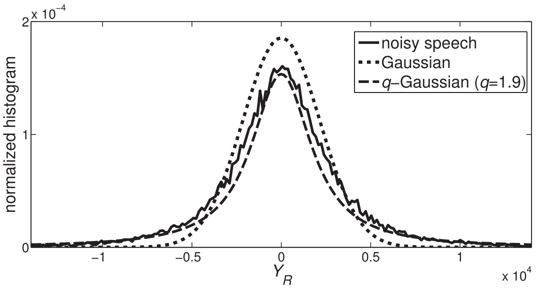

Fig. 4 Gaussian and q-Gaussian distributions fitted to the histogram of the speech data corrupted with subway noise at 0 dB SNR.

Figure 3 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 4 shows that the q-Gaussian distribution with $q = 1.9$ better fits the noisy speech than a Gaussian distribution ($q = 1$) does.

6.2 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 increment 0.01 in this region. Figure 5 shows the performance of q-SS_c for different SNR conditions on the Aurora-2 database. 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.

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 20 dB SNR. When q approaches 2, the weight factor in q-SS, i.e. $\upsilon(q)$, approaches zero. Therefore, it is very likely that $\upsilon(q)|Y_f|^2 < |N_f|^2$. 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.

We found that q-SS was significantly effective when the SNR conditions are between 0 to 10 dB. It was not sig-

Table 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

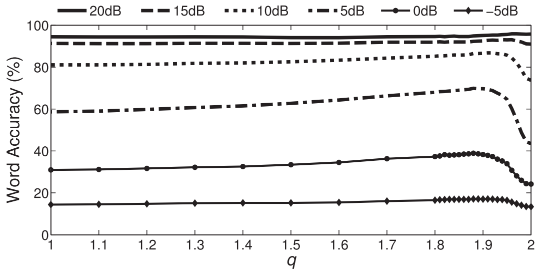

Fig. 5 Word Accuracy of q-SS_c for different SNR conditions. The averaged values over all the noise types are shown.

nificantly effective under the higher SNR conditions. We noticed some inconsistencies of our experimental results with the results in Sects. 4.3 and 6.1. When the SNR > 5 dB, the performance was better when a higher q was applied. We did not expect these results. These results suggest that the SNR may not be the only factors that can affect q or the relation between the SNR and q may 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 1 and 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. We found that 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 2 Performance comparison (Word Accuracy (%)) of q-SS_c (q = 1.88) and q-SS_m for each test set in Aurora-2. The averaged values over all the noise types are shown.

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

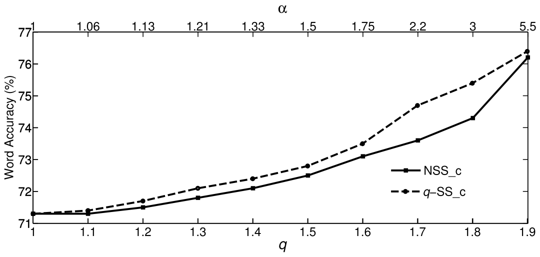

Fig. 6 Performance comparison of q-SS_c and NSS_c for $1 \leq q \leq 2$ and its equivalent α value.

6.3 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 $1 \leq q \leq 1.9$, $1 \leq \alpha \leq 5.5$. The results are shown in Fig. 6. 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

Table 3 Performance comparison (Word accuracy (%)) of q-SS_m for each noise type and SNR condition.

SNR (dB)	Noise Types							
	Subway	Babble	Car	Exhibition	Restaurant	Street	Airport	Station
Clean	**98.2**	98.1	98.0	**98.4**	**98.2**	98.1	98.0	**98.4**
20	**96.6**	91.3	**97.3**	**95.7**	89.1	**96.7**	**92.2**	94.5
15	**94.4**	**89.2**	**96.5**	**92.9**	85.6	**94.7**	**90.4**	92.7
10	**89.7**	**82.4**	**92.0**	**87.7**	79.1	**89.0**	**85.2**	**88.7**
5	**76.1**	**67.8**	**75.7**	**72.6**	**64.0**	**73.2**	**72.1**	**73.6**
0	**45.0**	**39.3**	**37.6**	**39.1**	**39.6**	**40.8**	**46.3**	**42.2**
−5	**20.3**	**16.1**	**15.3**	**15.7**	**16.1**	**17.5**	**20.3**	**18.1**

Table 4 Performance comparison (Word accuracy (%)) of LSS for each noise type and SNR condition.

SNR (dB)	Noise Types							
	Subway	Babble	Car	Exhibition	Restaurant	Street	Airport	Station
Clean	97.9	**98.3**	**98.4**	98.3	97.9	**98.3**	**98.4**	98.3
20	96.3	**92.0**	96.1	**95.8**	**91.4**	95.6	91.9	**95.0**
15	93.3	88.5	93.7	92.6	**88.6**	92.7	89.8	92.7
10	83.0	81.0	83.3	82.6	**80.0**	81.9	83.5	84.9
5	62.1	62.3	56.6	56.1	63.1	58.7	66.4	63.6
0	31.6	34.6	26.1	25.9	36.0	32.5	39.9	32.9
−5	16.6	14.1	12.5	12.3	15.2	14.2	17.7	15.2

filter. It is formulated as follows:

$$|\hat{X}_f| = H_{q\text{-SS}}(f)|Y_f|, \tag{35}$$

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

$$H_{q\text{-SS}} = \left(\frac{\upsilon(q)|Y_f|^2 - |N_f|^2}{|Y_f|^2}\right)^{0.5}. \tag{36}$$

Equation (36) can be expressed in term of the *a posteriori* SNR, γ_f, as follows:

$$H_{q\text{-SS}} = \left(\frac{\upsilon(q).\gamma_f - 1}{\gamma}\right)^{0.5}. \tag{37}$$

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

$$\text{Attenuation} = 20\log_{10} H_{q\text{-SS}}. \tag{38}$$

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

$$H_{\text{NSS}} = \left(\frac{|Y_f|^2 - \alpha|N_f|^2}{|Y_f|^2}\right)^{0.5}. \tag{39}$$

Equation (39) can be expressed in term of the *a posteriori* SNR, γ_f, as follows:

$$H_{\text{NSS}} = \left(\frac{\gamma_f - \alpha}{\gamma}\right)^{0.5}. \tag{40}$$

We can rewrite Eq. (40) as follows:

$$H_{\text{NSS}} = \left(\frac{\alpha\left(\frac{\gamma_f}{\alpha} - 1\right)}{\gamma_f}\right)^{(0.5)} \tag{41}$$

Since $\alpha = \frac{1}{\upsilon(q)}$, we can write Eq. (41) as follows:

$$H_{\text{NSS}} = \left(\frac{1}{\upsilon(q)}\right)^{0.5} H_{q\text{-SS}} \tag{42}$$

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

Figure 7 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 5 dB, the attenuation was −2.22 dB with NSS and −5.23 dB with q-SS. As we can see from Eq. (42), the transfer function of q-SS was the transfer function of NSS attenuated with a factor $\upsilon(q)^{0.5}$. 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 occurred at 2 dB SNR for NSS and 3 dB 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. Tables 3, 4, and 5 show the performance of LSS, NSS_n, and q-SS_m respectively for each combination of noise type and SNR. The best performance for each type of noise and each SNR was printed in bold. We found that q-SS_m was the best among them in average.

Table 6 summarizes the performance of several spectral subtraction methods for each SNR conditions. 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. These results also confirmed that the parameter q can be controlled using the SNR information.

Table 5 Performance comparison (Word accuracy (%)) of NSS_n for each noise type and SNR condition.

SNR (dB)	Noise Types							
	Subway	Babble	Car	Exhibition	Restaurant	Street	Airport	Station
Clean	98.0	98.2	98.2	98.1	98.0	98.2	98.2	98.1
20	96.1	91.5	96.5	95.3	89.6	95.8	91.0	94.4
15	93.4	87.8	95.0	91.7	85.9	93.6	89.2	**92.8**
10	86.3	79.6	88.7	85.5	77.0	85.9	82.9	86.7
5	70.8	62.9	69.3	65.0	61.7	66.4	67.8	71.4
0	39.1	35.0	33.3	31.6	36.6	36.2	41.5	38.7
−5	17.4	14.5	13.2	13.7	15.5	15.1	18.3	16.2

Table 6 The performance comparison (Word accuracy (%)) of several spectral subtraction methods for each SNR conditions. The average values over all noise types are shown.

Methods	SNR						Average
	20	15	10	5	0	−5	(0–20 dB)
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**

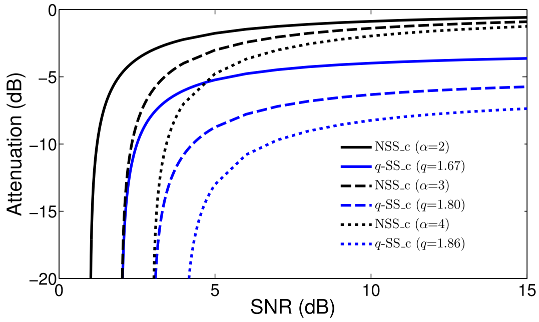

Fig. 7 The comparison of the attenuation curve of q-SS_c and NSS_c for several values of q and their respective α value.

7. 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.

Further investigation on the meaning of q should be investigated in future. While in this research no assumption was made on the clean speech distribution, it would be interesting to model clean speech spectra using the q-Gaussian distribution. Finding the relation between the change of q before and after speech are contaminated by noise also an interesting direction. It would also be interesting to apply the q-Gaussian assumption to other techniques for estimating spectrum such as the minimum mean squared error (MMSE)-based methods.

References

[1] D.V. Compernolle, "Noise adaptation in a hidden Markov model speech recognition system," Computer Speech and Language, vol.3, no.2, pp.151–167, 1989.

[2] S.F. Boll, "Suppression of acoustic noise in speech using spectral subtraction," IEEE Trans. Acoust. Speech Signal Process., vol.ASSP-27, no.2, pp.113–120, 1979.

[3] R. McAulay and M. Malpass, "Speech enhancement using a soft-decision noise suppression filter," IEEE Trans. Acoust. Speech Signal Process., vol.28, no.2, pp.137–145, April 1980.

[4] J. Wilbur B. Davenport, "An experimental study of speech-wave probability distributions," J. Acoust. Soc. Am., vol.24, no.4, pp.390–399, 1952.

[5] N. Evans, J. Mason, W. Liu, and B. Fauve, "An assessment on the fundamental limitations of spectral subtraction," Proc. IEEE Internat. Conf. on Acoustics, Speech and Signal Processing, vol.1, pp.1520–6149, May 2006.

[6] M. Berouti, R. Schwartz, and J. Makhoul, "Enhancement of speech corrupted by acoustic noise," Proc. IEEE Internat. Conf. on Acoustics, Speech and Signal Processing, vol.4, pp.208–211, April 1979.

[7] P. Lockwood and J. Boudy, "Experiments with a nonlinear spectral subtractor (nss), hidden Markov models and the projection, for robust speech recognition in cars," Speech Commun., vol.11, no.2-3, pp.215–228, 1992.

[8] B. Chen and P.C. Loizou, "A Laplacian-based MMSE estimator for speech enhancement," Speech Commun., vol.49, no.2, pp.134–143, 2007.

[9] R. Martin, "Speech enhancement using MMSE short time spectral estimation with gamma distributed speech priors," Proc. IEEE Internat. Conf. on Acoustics, Speech and Signal Processing, vol.1, pp.I–253–I–256, May 2002.

[10] R. Martin, "Speech enhancement based on minimum mean-square error estimation and supergaussian priors," IEEE Trans. Speech Au-

dio Process., vol.13, no.5, pp.845–856, Sept. 2005.

[11] C. Tsallis, "Possible generalization of Boltzmann-Gibbs statistics," J. Stat. Phys., vol.52, pp.479–487, 1988.

[12] H. Pardede, K. Shinoda, and K. Iwano, "*Q*-Gaussian based spectral subtraction for robust speech recognition," Proc. Interspeech, Tue.P5c.07, 2012.

[13] Q. Zhu and A. Alwan, "The effect of additive noise on speech amplitude spectra: a quantitative analysis," IEEE Signal Process. Lett., vol.9, no.9, pp.275–277, Sept. 2002.

[14] R.M. Udrea, N. Vizireanu, S. Ciochina, and S. Halunga, "Nonlinear spectral subtraction method for colored noise reduction using multi-band Bark scale," Signal Process., vol.88, no.5, pp.1299–1303, 2008.

[15] S. Kamath and P. Loizou, "Enhancement of speech corrupted by acoustic noise," Proc. IEEE Internat. Conf. on Acoustics, Speech and Signal Processing, vol.4, pp.IV–4164, May 2002.

[16] L. Nivanen, A.L. Méhauté, and Q. Wang, "Generalized algebra within a nonextensive statistics," Rep. Math. Phys., vol.52, no.3, pp.437–444, 2003.

[17] C. Tsallis, "Entropic nonextensivity: A possible measure of complexity," Chaos Solitons Fractals, vol.13, no.3, pp.371–391, 2002.

[18] L.H.Y. Chen and Q.M. Shao, "Normal approximation under local dependence," The Annals of Probability, vol.32, no.3, pp.1985–2028, 2004.

[19] S. Gazor and W. Zhang, "Speech probability distribution," IEEE Signal Process. Lett., vol.10, no.7, pp.204 –207, July 2003.

[20] S. Umarov, C. Tsallis, and S. Steinberg, "On a *q*-Central Limit Theorem consistent with nonextensive statistical mechanics," Milan J. Math., vol.75, pp.307–328, 2008.

[21] C. Vignat and A. Plastino, "Central limit theorem and deformed exponentials," J. Phys. A, vol.40, no.45, pp.F969–F978, 2007.

[22] W. Thistleton, J. Marsh, K. Nelson, and C. Tsallis, "*Q*-Gaussian approximants mimic non-extensive statistical-mechanical expectation for many-body probabilistic model with long-range correlations," Cent. Eur. J. Phys., vol.7, pp.387–394, 2009.

[23] C. Tsallis and S.M. Duarte Queirós, "Nonextensive statistical mechanics and central limit theorems I-Convolution of independent random variables and q-product," Proc. AIP Conf., vol.965, pp.8–20, 2007.

[24] C. Beck and E. Cohen, "Superstatistics," Physica A, vol.322, pp.267–275, 2003.

[25] H. Touchette and C. Beck, "Asymptotics of superstatistics," Phys. Rev. E, vol.71, 016131, Jan 2005.

[26] Y. Ephraim and D. Malah, "Speech enhancement using a minimum-mean square error short-time spectral amplitude estimator," IEEE Trans. Acoust. Speech Signal Process., vol.ASSP-32, no.6, pp.1109–1121, Dec 1984.

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

[28] H. Hirsch and C. Ehrlicher, "Noise estimation techniques for robust speech recognition," Proc. IEEE Internat. Conf. on Acoustics, Speech and Signal Processing,, vol.1, pp.153–156, 1995.

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

Hilman Pardede received his B.E degree in electrical engineering from University of Indonesia in 2004. He received his M.E degree in information and communication technology from University of Western Australia in 2009. He is currently working toward the Ph.D. degree at Tokyo Institute of Technology, Japan. His research interests include speech recognition, speech enhancement, and signal processing. He is a student member of International Speech Communication Association (ISCA).

Koji Iwano received the B.E. degree in information and communication engineering in 1995, and the M.E. and Ph.D. degrees in information engineering respectively in 1997 and 2000 from the University of Tokyo. He is currently an Associate Professor at Tokyo City University, Faculty of Environmental and Information Studies. His research interests are in speech information processing, such as speech recognition, speaker verification, and speech synthesis. He is a member of the IEEE, International Speech Communication Association (ISCA), the Information Processing Society of Japan (IPSJ), and the Acoustical Society of Japan (ASJ).

Koichi Shinoda received his B.S. in 1987 and his M.S. in 1989, both in physics, from the University of Tokyo. He received his D.Eng. in computer science from the Tokyo Institute of Technology in 2001. In 1989, he joined NEC Corporation, Japan, and was involved in research on automatic speech recognition. From 1997 to 1998, he was a visiting scholar with Bell Labs, Lucent Technologies, in Murray Hill, NJ. From June 2001 to September 2001, he was a Principal Researcher with Multimedia Research Laboratories, NEC Corporation. From October 2001 to March 2002, he was an Associate Professor with the University of Tokyo. He is currently an Associate Professor at the Tokyo Institute of Technology. His research interests include speech recognition, statistical pattern recognition, and human interfaces. Dr. Shinoda received the Awaya Prize from the Acoustic Society of Japan in 1997 and the Excellent Paper Award from the Institute of Electronics, Information, and Communication Engineers IEICE in 1998. He is an Associate Editor of Computer Speech and Language. He is a member of IEEE, ACM, ASJ, IPSJ, and JSAI.