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 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, Xf, of clean speech is a complex random variable that follows a Gaussian distribution with zero
mean and variance $\sigma(f)$. Similarly, a spectral component of
noise signal, $N_f$, is also a complex random variable that has
a Gaussian distribution with zero mean and variance $\tau(f)$. We
also assume that $X_f$ and $N_f$ are statistically independent,
and hence, noisy speech, $Y_f$, also follows a zero mean
Gaussian distribution with the probability density of $Y_f$ is
given by:

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

(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:

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

(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:

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

(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 sub-
traction (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 vari-
ants of spectral subtraction proposed to improve its perfor-
ance. One popular variant of spectral subtraction is the
one proposed by Berouti et al. [6]. They introduce an over-
subsection factor, $\alpha$, and the spectral subtraction formula
becomes:

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

(4)

Since the introduction of $\alpha$ makes the subtraction nonlinear,
it is called nonlinear spectral subtraction (NSS). Zhu and
Alwan [13] reported that the use of $\alpha$ 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 $\alpha$ is determined heuristically. There
exists no consistent ways to optimize $\alpha$.

Many variants of NSS have been proposed [7, 14, 15]. They are basically modifications of the Eq. (4). In this paper, we determine $\alpha$ for each frequency bin, denoting $\alpha_f$, using the following relation [6]:

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

(5)

Parameter $\alpha_0$ is the desired value of $\alpha$ at 0 dB SNR. It is
usually set between 4 and 6. In this paper we use $\alpha_0 = 4$. $\Phi_f$
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:

$$\Phi_f = 10 \log_{10} \gamma_f,$$

(6)

where:

$$\gamma_f = \frac{|Y_f|^2}{|N_f|^2}.$$  

(7)

In this paper, we denote NSS of which $\alpha_f$ is determined
using Eq. (5) as NSS_n. When a single and constant $\alpha_f$ 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 the-
ory generalizes Boltzmann-Gibbs statistics by utilizing $q$-
exponential function:

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

(8)

and its inverse, $q$-logarithmic function:

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

(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:

$$S_q = -k \int p_q(x) \log_q p_q(x).$$

(10)

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

A $q$-Gaussian distribution can be obtained by maximiz-
ing the Tsallis entropy in a similar way as a Gaussian dis-
tribution can be derived from Shannon entropy. The density
function for a $q$-Gaussian distribution with zero mean and
variance $\lambda_q$ is defined by:

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

(11)

where $A_q$ is a normalization term and defined as:

$$A_q = \begin{cases} 
\frac{1}{\sqrt{\pi}} & q = 1, \\
\frac{1}{\sqrt{\Gamma(\frac{1}{q})}} \sqrt{\frac{1-q}{\pi}} & -\infty < q < 1
\end{cases}$$

(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
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 \( Y_R = \text{Re}(Y_f) \) and \( Y_I = \text{Im}(Y_f) \) be the real and imaginary parts of the speech spectrum respectively. Assuming both \( Y_R \) and \( Y_I \) follow \( q \)-Gaussian and are identically distributed with variance \( \nu_q(f)/2 \). Then, the probability density functions for \( Y_R \) and \( Y_I \) are as follow:

\[
    P_q(Y_R) = \frac{\sqrt[2]{2A_qB_q}}{\sqrt[2]{\nu_q(f)}} \exp_q\left(-\frac{2B_q^2|Y_R|^2}{\nu_q(f)}\right), \quad (13)
\]

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

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:

\[
    P_q(Y_f) = \frac{2A^2_qB^2_q}{\nu_q(f)} \exp_q\left(-\frac{2B^2_q|Y_f|^2}{\nu_q(f)}\right). \quad (15)
\]

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:

\[
    \hat{\sigma}_q(f) = \frac{2(2-q)}{3-q} \sqrt[2]{\nu_q(f)} \frac{|Y_f|^2 - \tau_q(f)}{\nu_q(f)} \quad (16)
\]

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

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

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 \( \nu(q) = \frac{2(2-q)}{3-q} \), we can rewrite Eq. (17) as follows:

\[
    |\hat{X}_f|^2 = \nu(q) |Y_f|^2 - |N_f|^2. \quad (18)
\]

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

\[
    \frac{1}{\nu(q)} |\hat{X}_f|^2 = |Y_f|^2 - \frac{1}{\nu(q)} |N_f|^2. \quad (19)
\]

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

\[
    \alpha = \frac{1}{\nu(q)}, \quad (20)
\]

Based on Eq. (20), \( \alpha \) 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 \( \alpha \) 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

\[
    P_q(Y_f) = \frac{\sqrt[2]{2A_qB_q}}{\sqrt[2]{\nu_q(f)}} \exp_q\left(-\frac{2B_q^2|Y_f|^2}{\nu_q(f)}\right). \quad (14)
\]
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 $\Phi_f$ 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, $P_q(Y_R)$, which we denote as $\tilde{q}$, is selected using the following formula:

$$\tilde{q} = \arg\min_q \frac{1}{n} \sum_{i=1}^{n} \left( S(Y_R) - P_q(Y_R) \right)^2. \quad (21)$$

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:

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

where $\Psi_f(m)$ is the clean speech signal-to-noise ratio (in dB), i.e. the a priori SNR. It is different from $\Phi_f$ of Eq. (7), which is the ratio between noisy speech and noise. It is calculated using the following formula:

$$\Psi_f(m) = 10\log \xi_f(m), \quad (23)$$

where:

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

In practice, $\xi_f(m)$ is unknown and needs to be estimated. We obtain $\xi_f(m)$, the estimate of $\xi_f$, 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 $Y_f(m) = \{Y_f(m), Y_f(m-1), \ldots, Y_f(m-L+1)\}$, assuming the noise variance, $\tau(f)$ is known. By assuming the statistical independence of the $L$ observations and using Gaussian model, we obtain the likelihood function:

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

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

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

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

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

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

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

where:

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

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:

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

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

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

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:

\[ \zeta_f^\text{rel}(m) = \frac{\zeta_f(m) - \zeta_f^\text{min}(m)}{\zeta_f^\text{max}(m) - \zeta_f^\text{min}(m)}, \]

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

\[ |\hat{N}(m,k)|^2 = \begin{cases} |\hat{N}(m-1)|^2 & \text{if } \zeta_f^\text{rel}(m) < T, \\ |\hat{N}(m)|^2 & \text{else}, \end{cases} \]

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.

\[ |\hat{X}_f |^2 = \beta |Y_f |^2 \quad \text{if } |\hat{X}_f |^2 < \beta |Y_f |^2. \]

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 1st-order and 2nd-order derivatives, \( \Delta \) log energy and \( \Delta \Delta \) 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.

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. \( \nu(q) \), approaches zero. Therefore, it is very likely that \( \nu(q)|Y_f|^2 \sim |N_f|^2 \). In this condition, the clean speech estimate, \( |\hat{X}_f|^2 \), 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-
significantly 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.

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 ≤ q ≤ 1.9, 1 ≤ α ≤ 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 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 Aurora-2 database.

<table>
<thead>
<tr>
<th>Methods</th>
<th>Noise Types</th>
<th>SNR (dB)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>20</td>
</tr>
<tr>
<td>q-SS_c (q=1.88)</td>
<td>Subway</td>
<td>97.5</td>
</tr>
<tr>
<td>q-SS_m</td>
<td></td>
<td>96.6</td>
</tr>
<tr>
<td>q-SS_c (q=1.88)</td>
<td>Babble</td>
<td>91.0</td>
</tr>
<tr>
<td>q-SS_m</td>
<td></td>
<td>91.3</td>
</tr>
<tr>
<td>q-SS_c (q=1.88)</td>
<td>Car</td>
<td>97.9</td>
</tr>
<tr>
<td>q-SS_m</td>
<td></td>
<td>97.3</td>
</tr>
<tr>
<td>q-SS_c (q=1.88)</td>
<td>Exhibition</td>
<td>95.7</td>
</tr>
<tr>
<td>q-SS_m</td>
<td></td>
<td>95.7</td>
</tr>
<tr>
<td>q-SS_c (q=1.88)</td>
<td>Restaurant</td>
<td>88.9</td>
</tr>
<tr>
<td>q-SS_m</td>
<td></td>
<td>89.1</td>
</tr>
<tr>
<td>q-SS_c (q=1.88)</td>
<td>Street</td>
<td>97.2</td>
</tr>
<tr>
<td>q-SS_m</td>
<td></td>
<td>96.7</td>
</tr>
<tr>
<td>q-SS_c (q=1.88)</td>
<td>Airport</td>
<td>91.4</td>
</tr>
<tr>
<td>q-SS_m</td>
<td></td>
<td>92.3</td>
</tr>
<tr>
<td>q-SS_c (q=1.88)</td>
<td>Station</td>
<td>95.4</td>
</tr>
<tr>
<td>q-SS_m</td>
<td></td>
<td>94.5</td>
</tr>
</tbody>
</table>

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

Fig. 6 Performance comparison of q-SS_c and NSS_c for 1 ≤ q ≤ 2 and its equivalent α value.
filter. It is formulated as follows:

$$|\hat{X}_f| = H_{q-SS}(f)|Y_f|.$$  \hfill (35)

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

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

Equation (36) can be expressed in term of the a posteriori SNR, $\gamma_f$, as follows:

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

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

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

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

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

Equation (39) can be expressed in term of the a posteriori SNR, $\gamma_f$, as follows:

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

We can rewrite Eq. (40) as follows:

$$H_{NSS} = \left(\frac{\gamma_f - \alpha}{\gamma}\right)^{0.5}.$$  \hfill (41)

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

$$H_{NSS} = \left(\frac{1}{\frac{v(q)}{\gamma}}\right)^{0.5}.$$  \hfill (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 $v(q)^{0.5}$. In speech recognition, scaling down (attenuation) of the signals did not affect its performance. Therefore, the performance of speech recognition would not be affected much by the attenuation at the high SNR conditions. However, flooring occurred at a higher SNR for $q$-SS than for NSS. For instance, for $\alpha = 3$, i.e. $q = 1.8$, flooring 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 several spectral subtraction methods for each SNR conditions. The 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 $\alpha$ (NSS_c) and adaptive $\alpha$ (NSS_a). 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$_q$ for each noise type and SNR condition.

<table>
<thead>
<tr>
<th>SNR (dB)</th>
<th>Subway</th>
<th>Babble</th>
<th>Car</th>
<th>Exhibition</th>
<th>Restaurant</th>
<th>Street</th>
<th>Airport</th>
<th>Station</th>
</tr>
</thead>
<tbody>
<tr>
<td>Clean</td>
<td>98.0</td>
<td>98.2</td>
<td>98.2</td>
<td>98.1</td>
<td>98.0</td>
<td>98.2</td>
<td>98.2</td>
<td>98.1</td>
</tr>
<tr>
<td>20</td>
<td>96.1</td>
<td>91.5</td>
<td>96.5</td>
<td>95.3</td>
<td>89.6</td>
<td>95.8</td>
<td>91.0</td>
<td>94.4</td>
</tr>
<tr>
<td>15</td>
<td>93.4</td>
<td>87.8</td>
<td>95.0</td>
<td>91.7</td>
<td>85.9</td>
<td>93.6</td>
<td>89.2</td>
<td>92.8</td>
</tr>
<tr>
<td>10</td>
<td>86.3</td>
<td>79.6</td>
<td>88.7</td>
<td>85.5</td>
<td>77.0</td>
<td>83.9</td>
<td>82.9</td>
<td>86.7</td>
</tr>
<tr>
<td>5</td>
<td>70.8</td>
<td>62.9</td>
<td>69.3</td>
<td>65.0</td>
<td>61.7</td>
<td>66.4</td>
<td>67.8</td>
<td>71.4</td>
</tr>
<tr>
<td>0</td>
<td>39.1</td>
<td>35.0</td>
<td>33.3</td>
<td>31.6</td>
<td>36.6</td>
<td>36.2</td>
<td>41.5</td>
<td>38.2</td>
</tr>
<tr>
<td>−5</td>
<td>17.4</td>
<td>14.5</td>
<td>13.2</td>
<td>13.7</td>
<td>15.5</td>
<td>15.1</td>
<td>18.3</td>
<td>16.2</td>
</tr>
</tbody>
</table>

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.

<table>
<thead>
<tr>
<th>Methods</th>
<th>SNR</th>
<th>Average (0–20 dB)</th>
</tr>
</thead>
<tbody>
<tr>
<td>No Compensation</td>
<td>95.9</td>
<td>71.3</td>
</tr>
<tr>
<td>LSS</td>
<td>94.5</td>
<td>67.3</td>
</tr>
<tr>
<td>NSS$_{q}$ (α = 5.5)</td>
<td>94.4</td>
<td>70.1</td>
</tr>
<tr>
<td>NSS$_q$</td>
<td>94.1</td>
<td>28.7</td>
</tr>
<tr>
<td>q-SS$_{q}$ (q = 1.88)</td>
<td>95.1</td>
<td>17.0</td>
</tr>
<tr>
<td>q-SS$_{q}$n</td>
<td>94.6</td>
<td>28.7</td>
</tr>
</tbody>
</table>

Fig. 7  The comparison of the attenuation curve of q-SS$_{q}$ and NSS$_{q}$ for several values of $q$ and their respective $\alpha$ 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 $\alpha$ 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 $q$ 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


