論文 / 著書情報
Article / Book Information

Title	Q-Gaussian based spectral subtraction for robust speech recognition
Authors	Hilman F. Pardede, Koichi Shinoda, Koji Iwano
Citation	InterSpeech2012, , , , Tue.P5c.07
Pub. date	2012, 9
Copyright	(c) 2012 International Speech Communication Association, ISCA
DOI	http://dx.doi.org/

Q-Gaussian based spectral subtraction for robust speech recognition

Hilman F. Pardede[1], *Koichi Shinoda*[1], *Koji Iwano*[2]

[1]Department of Computer Science, Tokyo Institute of Technology, Tokyo, Japan
[2]Faculty of Environmental and Information Studies, Tokyo City University, Yokohama, Japan
hilman@ks.cs.titech.ac.jp, shinoda@cs.titech.ac.jp, iwano@tcu.ac.jp

Abstract

Spectral subtraction (SS) is derived using maximum likelihood estimation assuming both noise and speech follow Gaussian distributions and are independent from each other. Under this assumption, noisy speech, speech contaminated by noise, also follows a Gaussian distribution. 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 non-extensive statistics to represent the distribution of noisy speech and derive a new spectral subtraction method based on it. In our analysis, the q-Gaussian distribution fits the noisy speech distribution better than the Gaussian distribution does. Our speech recognition experiments showed that the proposed method, q-spectral subtraction (q-SS), outperformed the conventional SS method using the Aurora-2 database.

Index Terms: robust speech recognition, spectral subtraction, Gaussian distribution, q-Gaussian, maximum likelihood

1. Introduction

Currently, automatic speech recognition (ASR) is able to achieve high performance in clean environments. However, its performance in noisy environments is still low. Spectral subtraction (SS) which removes additive noise from noisy speech, is often utilized to improve the robustness of speech recognition against noise [1]. Spectral subtraction is derived based on an *extensive* framework. In an extensive framework, we assume that the sub-systems of a system are independent from each other, and thus, the additivity between them holds. In spectral subtraction, it is assumed that noise and speech are uncorrelated. Under this assumption, we assume that both speech and noise spectra follow Gaussian distributions, and thus, noisy speech will also follow a Gaussian distribution. By maximizing the likelihood of the noisy speech distribution, the spectral subtraction formula can be derived [2].

The extensive framework however, fails to explain some phenomena in *complex* systems. In a complex system, we do not know about the sub-systems and their relations. In such a system, the extensive property does not hold. Therefore, it is often called a "non-extensive system". A speech pattern is a complex system. In clean speech, various long-term correlations exist among its different spectral components in complex ways in various time scales. Short-time speech spectra do not follow Gaussian distributions [3] but show heavy-tailed distributions instead. Laplace [4] and Gamma [5] distributions are often used to model the speech distribution instead of the Gaussian distributions. When speech is corrupted with noise, the use of a short-time window in signal processing will also introduce a cross-term, which exists when the speech and noise spectra overlap in time-frequency space. Thus, noisy speech short-time spectra are likely to follow heavy-tailed distributions and not Gaussian distributions.

Therefore, it is not surprising that spectral subtraction may not give sufficiently high performance when noise and speech are correlated. A weighting factor is often introduced to improve its performance. However, this factor is decided heuristically.

Recently, a theory of non-extensive statistics has been introduced to explain several phenomena in complex systems [6]. 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. The q-Gaussian distribution has successfully represented many phenomena in complex systems in statistical mechanics, economics, finance, biology, astronomy and machine learning.

In this paper, we derive spectral subtraction in a non-extensive framework. In this framework, we still assume that noise and speech follow Gaussian distributions, but we allow noise and speech to be correlated. Accordingly, the distribution of noisy speech follows q-Gaussian. We derive q-spectral subtraction in a similar way as spectral subtraction is derived using maximum likelihood.

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

2. Spectral subtraction

Spectral subtraction is a popular method to remove additive noise from noisy speech in the spectral domain, assuming the noise spectrum is known. Let $y(t)$ denote noisy speech consisting of clean speech $x(t)$ and additive noise $n(t)$. By taking the short-time fourier transform of the signals, we obtain their spectral representation.

Consider a spectral component at frequency f. We assume a spectral component, X_f, of clean speech is a complex random variable that follows a Gaussian distribution with zero mean and variance $\sigma(f)$. Similarly, a spectral component of noise signal, N_f, is also a complex random variable that has a Gaussian distribution with zero mean and variance $\tau(f)$. The variance of a distribution represents the power spectrum of the observed signals. We denote $|X_f|^2$ and $|N_f|^2$ as the observed power spectra of clean speech and noise respectively. Therefore, $|X_f|^2 = \sigma(f)$ and $|N_f|^2 = \tau(f)$. We also assume that X_f and N_f are statistically independent, and hence, noisy speech, Y_f, also follows Gaussian distribution with variance $\nu(f) = \sigma(f) + \tau(f)$. Then, the probability density of Y_f is

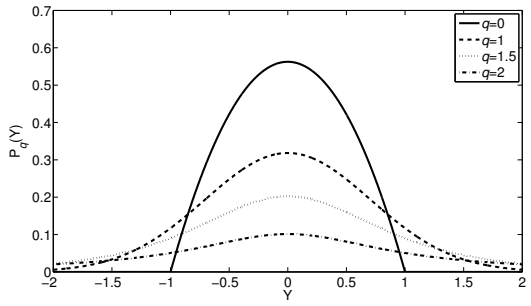

Figure 1: q-Gaussian distribution for several q

given by:

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

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

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

Since $\hat{\sigma}(f)$ is the estimated power spectrum of clean speech, $|\hat{X}_f|^2$, Eq. (2) is basically power spectral subtraction. It maintains a linear relation between noisy speech, noise and clean speech. Therefore, it is also called linear spectral subtraction (LSS).

Berouti et al. [7] introduced an oversubtraction factor, α with the original intention of reducing the effect of musical noise caused by spectral subtraction. This parameter is an SNR-dependent parameter. The spectral subtraction formula becomes:

$$|\hat{X}_f|^2 = |Y_f|^2 - \alpha(\text{SNR})|N_f|^2. \quad (3)$$

Since the introduction of α makes the subtraction nonlinear, it is called nonlinear spectral subtraction (NSS). Zhu and Alwan [8] reported that this factor also compensates for nonlinear relation between noise and speech. Since there exists no consistent ways to optimize α, it is usually determined heuristically.

3. Q-Gaussian distribution

Recently, Tsallis has introduced a theory of non-extensive statistics in the field of statistical mechanics [6]. This theory generalizes Boltzmann-Gibbs statistics by utilizing q-exponential function:

$$\exp_q(x) = (1 + (1-q)x)^{\frac{1}{1-q}}, \quad (4)$$

and its inverse, q-logarithmic function:

$$\log_q(x) = \frac{x^{1-q} - 1}{1-q}. \quad (5)$$

These functions asymptotically approach exponential and natural logarithmic functions as q approaches 1. They are non-extensive when $q \neq 1$ [9]. In the non-extensive framework, entropy is redefined:

$$S_q = -k \int p_i(x) \log_q p_i(x). \quad (6)$$

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:

$$P_q(Y) = \frac{A_q B_q}{\sqrt{\lambda_q}} \exp_q\left(-\frac{B_q^2|Y|^2}{\lambda_q}\right), \quad (7)$$

where A_q is a normalization term and defined as:

$$A_q = \begin{cases} \frac{\Gamma\left(\frac{5-3q}{2-2q}\right)}{\Gamma\left(\frac{2-q}{1-q}\right)}\sqrt{\frac{1-q}{\pi}} & -\infty < q < 1 \\ \frac{1}{\sqrt{\pi}} & q = 1 \\ \frac{\Gamma\left(\frac{1}{q-1}\right)}{\Gamma\left(\frac{3-q}{2q-2}\right)}\sqrt{\frac{q-1}{\pi}} & 1 < q < 3, \end{cases} \quad (8)$$

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 $1 < q < 3$. The q-Gaussian distribution 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 [10] of a system. However, up to our knowledge, an automatic method to optimize q does not yet exist. Usually, in the implementation, it is chosen empirically.

4. Q-Spectral subtraction

In this section we derive our proposed method. We assume that the spectral component of noisy speech follows the q-Gaussian distribution with variance $\nu_q(f)$. 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. 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:

$$P_q(Y_R) = \frac{\sqrt{2}A_q B_q}{\sqrt{\nu_q(f)}} \exp_q\left(-\frac{2B_q^2|Y_R|^2}{\nu_q(f)}\right), \quad (9)$$

$$P_q(Y_I) = \frac{\sqrt{2}A_q B_q}{\sqrt{\nu_q(f)}} \exp_q\left(-\frac{2B_q^2|Y_I|^2}{\nu_q(f)}\right). \quad (10)$$

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 [11]. The distribution for noisy speech can be formulated as follows:

$$P_q(Y_f) = \frac{2A_q^2 B_q^2}{\nu_q(f)} \exp_q\left(-\frac{2B_q^2|Y_f|^2}{\nu_q(f)}\right). \quad (11)$$

Equation (11) 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:

$$\hat{\sigma}_q(f) = \frac{2(2-q)}{3-q}|Y_f|^2 - \tau_q(f). \quad (12)$$

Since, $|X_f|^2 = \sigma_q(f)$ and $|N_f|^2 = \tau_q(f)$, Eq. (12) becomes:

$$|\hat{X}_f|^2 = \frac{2(2-q)}{3-q}|Y_f|^2 - |N_f|^2. \quad (13)$$

Eq. (13) is the q-spectral subtraction (q-SS) formula. This method will be the same as LSS when $q = 1$.

It can be seen from Eq. (13) that q-SS closely related to NSS. In q-SS, a factor of $\frac{2(2-q)}{3-q}$ is introduced. By dividing Eq. (13) with this factor we obtain:

$$\frac{3-q}{2(2-q)}|\hat{X}_f|^2 = |Y_f|^2 - \frac{3-q}{2(2-q)}|N_f|^2. \quad (14)$$

Since scaling does not affect the performance of speech recognition, we can relate α in Eq. (3) with Eq. (14):

$$\alpha = \frac{3-q}{2(2-q)}. \quad (15)$$

Thus, our q-SS formulation gives a consistent way to estimate the control parameter α in NSS.

5. Experiments

5.1. Experimental setup

Our proposed method was evaluated in speech recognition experiments using the Aurora-2 database [12]. 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 $\Delta\Delta$ 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.

5.2. Evaluation procedure

In this paper, we implemented the minimum tracking algorithm [13] for estimating the noise spectrum, $|\hat{N}_f|^2$. We also implemented the voice activity detector (VAD) in [14] for noise updating.

For our evaluation, we also implemented the conventional NSS method [7], in which the control parameter α is determined in a heuristic way as:

$$\alpha = \begin{cases} 1 & \text{if NSNR}_f \geq 20\text{dB}, \\ \alpha_0 - \frac{3}{20}\text{NSNR}_f & \text{if } -5\text{dB} \leq \text{NSNR}_f < 20\text{dB}, \\ 4.75 & \text{if NSNR}_f < -5\text{dB}. \end{cases} \quad (16)$$

Parameter α_0 is the desired value of α at 0 dB SNR. Usually it is set between 4 to 6. In this paper we use $\alpha_0 = 4$. NSNR$_f$ is the noisy signal to noise ratio:

$$\text{NSNR}_f = 10\log\frac{|Y_f|^2}{|\hat{N}_f|^2}. \quad (17)$$

To avoid having negative values in the estimate of the clean speech spectrum, $|\hat{X}_f|^2$, we applied the following flooring rule:

$$|\hat{X}_f|^2 = \beta|Y_f|^2 \qquad \text{if } |\hat{X}_f|^2 < \beta|Y_f|^2. \quad (18)$$

Parameter β is usually set between 0.1 to 0.001. We set $\beta = 0.01$. This rule is applied for the three spectral subtraction methods, LSS, NSS and q-SS.

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

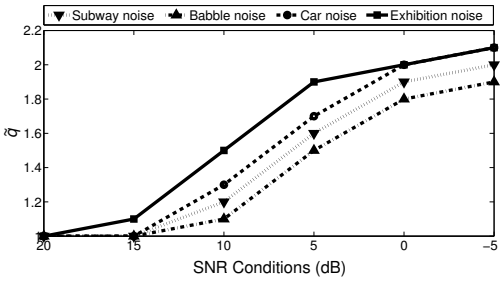

Figure 2: Estimation of q based on the mean square error for different SNR condition

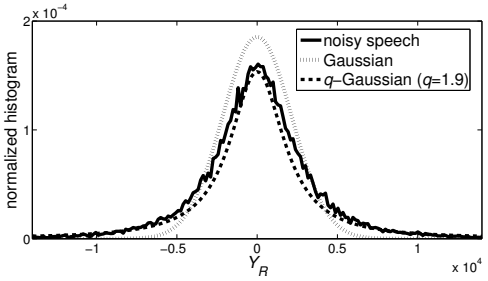

Figure 3: Gaussian and q-Gaussian distributions fitted to histograms of speech corrupted with subway noise at 0 dB SNR

5.3. Q-Gaussian representation of noisy speech

In this section, we will show how a q-Gaussian distribution better fits the noisy speech distribution and estimate the optimum q. Let $S(Y_R)$ be the empirical distribution of noisy speech. We obtain this distribution from the histograms of the real part of a DFT coefficient from 200 utterances of female speakers for each SNR condition from Test Set A of the Aurora-2 database. We only consider a single DFT coefficient (50-th coefficient) from a total of 256 coefficients. Then, we normalize the histograms so that the total area of the histograms is 1. Based on the data, we obtain its variance and $S(Y_{R_i})$ where $i = 1, 2, ..., n$ are the center point of each histogram bin. Then, we calculate the $P_q(Y_{R_i})$ for $1 \leq q < 3$ using Eq. (9). The optimum q, $\tilde{q}$, is the q-value that minimizes the mean squared error between the normalized histogram, $S(Y_{R_i})$ and the q-Gaussian distribution, $P_q(Y_{R_i})$:

$$\tilde{q} = \underset{q}{\text{argmin}}\frac{1}{n}\sum_{i=1}^{n}\left(S(Y_{R_i}) - P_q(Y_{R_i})\right)^2. \quad (19)$$

Figure 2 shows the estimated q-value for each noise conditions and for each SNR condition. As we can see, the optimum q-value is higher when the SNR is lower. Figure 3 shows that the q-Gaussian distribution with $q = 1.9$ better fits the noisy speech than a Gaussian distribution ($q = 1$) does.

5.4. Recognition results

We conducted several experiments to compare the performance of q-SS to those of LSS and NSS. Figure 4 shows the average accuracy of q-SS when the q-value is varied from 1 to 2. We found that when $1 < q \leq 1.9$, q-SS is better than LSS. The best accuracy is obtained when $q = 1.9$, where we achieved 17.9% relative improvement compared to LSS. Compared to NSS, q-SS was also better for some q-values. Figure 5 shows

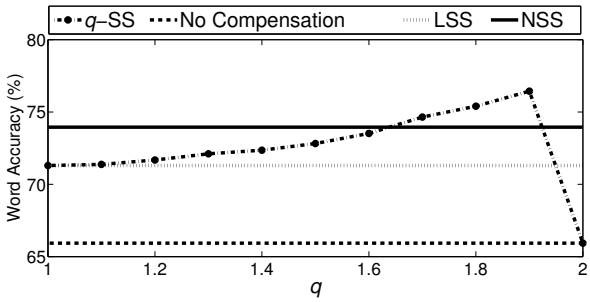

Figure 4: Performance comparison (Word Accuracy) of q-SS with LSS and NSS

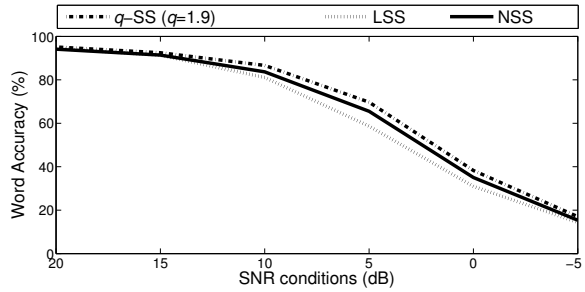

Figure 5: Performance comparison (Word Accuracy) of q-SS with LSS and NSS for different SNR conditions of the Aurora-2 database

the performance of q-SS for different SNR conditions when $q = 1.9$. The performance of q-SS was better for all SNR conditions than LSS, especially for the conditions 0dB to 15dB SNR. From Fig. 6, the same optimum q-value was found for almost all SNR conditions except 20 dB SNR.

As shown in Eq. (15), we can relate the nonlinear factor in q-SS with the oversubtraction factor, α, in NSS. When $q = 1.9$, we obtain $\alpha = 5.5$. The results when α is fixed at 5.5 is shown in Table 1. The slight difference between q-SS and NSS is because of the flooring process is not scaled as well.

6. Conclusions

We have derived q-spectral subtraction based on the q-Gaussian distribution assumption for noisy speech. The q-Gaussian distribution has been shown to fit noisy speech better than a Gaussian distribution. Our speech recognition results showed that our method is better than nonlinear spectral subtraction when q is 1.9. It gives a consistent way to estimate the control parameter α in NSS from the spectra of observed noisy speech.

We plan to investigate how to optimize q. We are also interested in extending the q-Gaussian assumption to other techniques used in robust speech recognition such as the minimum mean squared error (MMSE)-based method.

7. 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] R. McAulay and M. Malpass, "Speech enhancement using a soft-decision noise suppression filter," *IEEE Trans Acoust,*, vol. 28, pp. 137 – 145, apr 1980.

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

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

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

[6] C. Tsallis, "Possible generalization of boltzmann-gibbs statistics," *J. Stat. Phys.*, vol. 52, pp. 479–487, 1988.

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

[8] Q. Zhu and A. Alwan, "The effect of additive noise on speech amplitude spectra: a quantitative analysis," *IEEE Signal Process. Lett.*, vol. 9, pp. 275 – 277, sep 2002.

[9] 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.

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

[11] R. Martin, "Speech enhancement based on minimum mean-square error estimation and supergaussian priors," *IEEE Trans. Speech Audio Process.*, vol. 13, pp. 845 – 856, sept. 2005.

[12] H. Hirsch and D. Pearce, "The aurora experimental framework for the performance evaluation of speech recognition systems under noisy conditions," in *ISCA ITRW ASR2000*, (Paris, France), pp. 181–188, 2000.

[13] G. Doblinger, "Computationally efficient speech enhancement by spectral minima tracking in subbands," in *Proc. Eurospeech,*, (Madrid, Spain), pp. 1513–1516, 1995.

[14] 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.

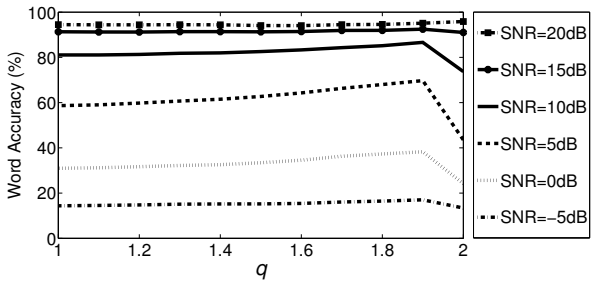

Figure 6: Performance (Word Accuracy) of q-SS for different SNR conditions of the Aurora-2 database

Table 1: Performance comparison (Word accuracy (%)) between q-SS when $q = 1.9$ and NSS when α is set at 5.5

Conditions (dB)	q-SS ($q = 1.9$)	NSS ($\alpha = 5.5$)
Clean	98.8	98.0
20	95.1	94.4
15	92.5	91.9
10	86.6	86.1
5	69.7	70.1
0	38.3	38.8
-5	17.0	16.7
Average (0-20dB)	76.4	76.2