論文 / 著書情報
Article / Book Information

論題(和文)	
Title(English)	Spectral Subtraction Based on q-Gaussian Assumption for Noise Robust Speech Recognition
著者(和文)	パーデデ ヒルマン フェルディナンデス, 篠田 浩一, 岩野 公司
Authors(English)	Hilman F. Pardede, Koichi Shinoda, Koji Iwano
出典(和文)	日本音響学会 2012年 春季研究発表会 講演論文集, Vol. , No. , pp. 21-22
Citation(English)	2012 Spring Meeting ASJ, Vol. , No. , pp. 21-22
発行日 / Pub. date	2012, 3

Spectral Subtraction Based on q-Gaussian Assumption for Noise Robust Speech Recognition*

☆ Hilman F. Pardede[1], Koichi Shinoda[1], Koji Iwano[2]
([1]Tokyo Institute of Technology, [2]Tokyo City University)

1 Introduction

Spectral subtraction is used to remove additive noise and improves the robustness of speech recognition systems [1]. It is derived by maximizing the likelihood of noisy speech distribution assuming noise and speech are Gaussian distributions and statistically independent from each other. However, noise and speech could be correlated, for instance when short time fourier transform (STFT) is used. Therefore, the Gaussian assumption may not applicable.

Tsallis entropy was proposed as a generalization of Shannon entropy [2]. It has successfully implemented to complex systems in many areas in physics such as cosmology, fractals and self-gravitating systems. The q-Gaussian distribution was derived by maximizing Tsallis entropy [3] in similar fashion as Gaussian distribution can be derived my maximizing Shannon entropy. This distribution is a generalized form of the Gaussian distribution. By using the q-Gaussian distribution, we can consider that noise and speech can be correlated [4].

2 q-Spectral Subtraction

We assume that a DFT component, X, at a frequency bin f, is a complex random variable that follows Gaussian distribution with zero mean and variance σ. Similarly, each DFT component of noise signal, N, is also a complex random variable that has a Gaussian distribution with zero mean and variance τ. The variance of the distributions represent the power spectrum of the observed process. We denote $|x|^2$ and $|n|^2$ are the observed power spectrum of clean speech and noise respectively, therefore $|x|^2 = \sigma$ and $|n|^2 = \tau$. We also assume that X and N are statistically independent, and noisy speech Y, also follows Gaussian distribution and the variance $\nu = \sigma + \tau$. Then, the probability density of Y is given by:

$$P(Y) = \frac{1}{\pi\nu} e^{\left(-\frac{|y|^2}{\nu}\right)}. \tag{1}$$

We would like to find the estimation of clean speech $|x|$ from an observation of $|y|$ and we know τ. Therefore, by maximizing $P(Y)$ with respect to σ, we obtain the maximum likelihood estimation of $\hat{\sigma}$ as the following:

$$\hat{\sigma} = |y|^2 - \tau \tag{2}$$

Eq. (2) is basically linear spectral subtraction (LSS) formulation. Therefore spectral subtraction can be derived using maximum likelihood principles by assuming that the speech and noise DFT components are independent Gaussian random process with the variance is unknown but deterministic.

The Gaussian assumption does not satisfy when the noise and speech are correlated. We extend the Gaussian assumption in Eq. (1) to q-Gaussian. The q-Gaussian distribution utilizes the q-exponential (q-exp) function, $e_q^x = (1 + (1-q)x)^{\frac{1}{1-q}}$, which recovers the usual exponential function when q is 1. This function introduces nonadditivity when $q \neq 1$. The q-Gaussian model for noisy speech is defined by:

$$P_q(Y) = \frac{2A_q^2 B_q^2}{\nu_q} e_q^{\left(-\frac{2B_q|y|^2}{\nu_q}\right)}, \tag{3}$$

where A_q is a normalization term:

$$A_q = \begin{cases} \frac{\Gamma\left(\frac{5-3q}{2-2q}\right)}{\Gamma\left(\frac{2-q}{1-q}\right)}\sqrt{\frac{1-q}{\pi}} & -\infty < q < 1 \\ \frac{1}{\sqrt{\pi}} & q = 1 \\ \frac{\Gamma\left(\frac{1}{q-1}\right)}{\Gamma\left(\frac{3-q}{2q-2}\right)}\sqrt{\frac{q-1}{\pi}} & 1 < q < 3, \end{cases} \tag{4}$$

and $B_q = \frac{1}{\sqrt{3-q}}$. The q-Gaussian gives the Gaussian distribution when $q = 1$.

With similar fashion as for spectral subtraction, by differentiating $P_q(Y)$ with respect to σ_q, and equating to zero we obtain the maximum likelihood estimation of $\hat{\sigma}_q$ as the following:

$$(3-q)\hat{\sigma}_q = 2(2-q)|y|^2 - (3-q)\tau_q \tag{5}$$

Since, $|x|^2 = \sigma_q$ and $|n|^2 = \tau_q$, Eq. (5) becomes:

$$|\hat{x}|^2 = 2\frac{(2-q)}{(3-q)}|y|^2 - |n|^2. \tag{6}$$

Eq. (6) is q-spectral subtraction (q-SS). It is the same with LSS when $q = 1$.

---

*耐雑音音声認識のための q-Gaussian 仮定に基づくスペクトルサブトラクション, ヒルマン パーデデ [1], 篠田 浩一 [1], 岩野公司 [2] ([1] 東京工業大学,[2] 東京都市大学)

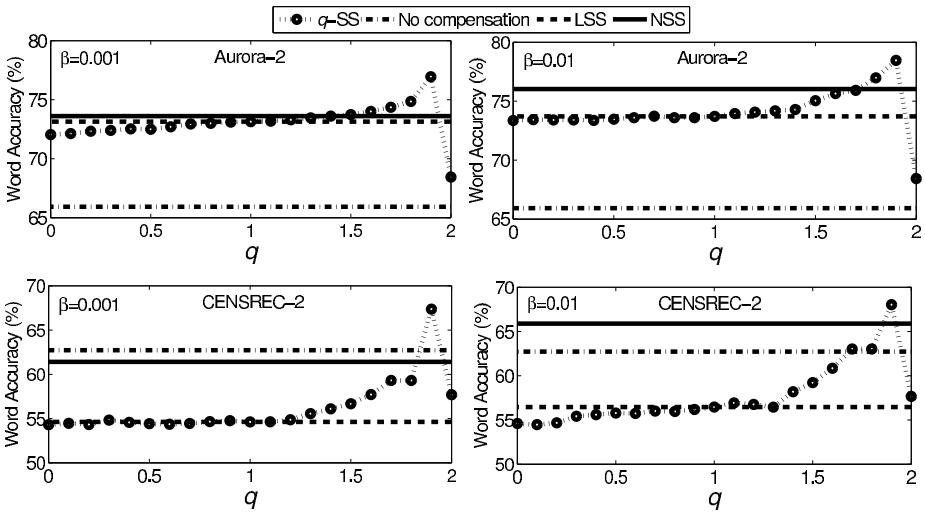

Fig. 1: Performance comparison (Word Accuracy) of q-SS with LSS and NSS

We notice that q-SS has similarity with nonlinear spectral subtraction (NSS):

$$|\hat{x}|^2 = |y|^2 - \alpha|n|^2 \tag{7}$$

where the oversubtraction factor, α, is introduced as function of SNRs as a tunable parameter. However, the value of α is decided intuitively. By implementing the q-Gaussian distribution, we can explain that the nonlinear effect of additive noise on the speech spectrums.

In practice, $|n|^2$ is not known and should be estimated. We implemented the minimum tracking algorithm [7] to find $|\hat{n}|^2$. Since noise is usually non-stationary, we implement a simple voice activity detector (VAD) [8] for noise updating.

We compare q-SS with both LSS and NSS. For NSS, the parameter α is determined by the following relation:

$$\alpha = \begin{cases} 1 & \text{if NSNR} \geq 20\text{dB}, \\ 4.75 - \frac{3}{20}\text{NSNR} & \text{if } -5\text{dB} \leq \text{NSNR} < 20\text{dB}, \\ 4.75 & \text{if NSNR} < -5\text{dB}, \end{cases} \tag{8}$$

where NSNR is the noisy signal to noise ratio which is calculated for each frame. To avoid negative value of the enhanced speech, $|\hat{x}|$, that we obtain from spectral subtraction, we applied regularization $|\hat{x}|^2 = \beta|y|^2$ when $|\hat{x}|^2 < \beta|y|^2$. In this paper, we set 0.001 and 0.01 for β.

3 Experiments

Our proposed method was evaluated in speech recognition experiment using two databases, the Aurora-2 [5] and the CENSREC-2 [6]. We used MFCC as features which were obtained using 23 triangle mel-filterbanks. For recognition, 38 dimensional MFCC features were used which consist of 12 static features, their 1st and 2nd derivatives, Δ log energy and $\Delta\Delta$ log energy. For recognition systems we implemented HMM-based system [5].

Figure 1 shows the word accuracies of q-SS, LSS and NSS. For q-SS, we varied q from 0 to 2. We noticed that the performance of q-SS improved when $q > 1$. The implementation of spectral subtraction seemed not too good for CENSREC-2 database. This may due to non-stationary noise in real environment are not accurately estimated. We found that implementation of NSS achieved better word accuracies than LSS for both databases. The performance of q-SS was gradually closer to NSS when $q > 1$. When $q = 1.9$, q-SS was better than NSS.

4 Conclusions

We have derived q-spectral subtraction algorithm based on the q-Gaussian distribution, which represents well the nonlinear effect of additive noise on speech. We found the similarity of our method with nonlinear spectral subtraction. Our experimental results showed that our method is better than nonlinear spectral subtraction when q is 1.9. For future work, we are interested on extending the q-Gaussian assumption to other methods for robust speech recognition such as minimum mean squared error (MMSE) based method.

References

[1] D. V. Compernolle, Computer Speech and Language, vol. 3 (2), 151 – 167, 1989.
[2] C. Tsallis, J. Stat. Phys., vol. 52, pp. 479–487, 1988.
[3] C. Tsallis, Milan J. Math., vol. 73, 145 – 176, 2005.
[4] C. Vignat, *et al.*, J. Phys. A, vol. 40 (45), F969, 2007.
[5] H. Hirsch, *et al.*, in Proc. ISCA ITRW ASR2000, 181 – 188, 2000.
[6] S. Nakamura, *et al.*, in Proc. Interspeech, 2330 – 2333, 2006.
[7] G. Doblinger, in Proc. Eurospeech, 1513–1516, 1995.
[8] H. Hirsch, *et al.*, in Proc. ICASSP, 1, 153 – 156, 1995.