論文 / 著書情報
Article / Book Information

論題(和文)	
Title(English)	A Compensation Technique Using q-Logarithm for Noisy 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. 19-20
Citation(English)	2012 Spring Meeting ASJ, Vol. , No. , pp. 19-20
発行日 / Pub. date	2012, 3

A Compensation Technique Using q-Logarithm for Noisy Speech Recognition *

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

1 Introduction

The existence of noise significantly degrades the performance of automatic speech recognition (ASR). In most methods for robust ASR, it is assumed that noise and speech signals are uncorrelated. This assumption, however, may not hold when short time Fourier transform (STFT) is used. In STFT, noise and speech can be correlated, and such correlation generate their cross-term. This cross-term also deteriorates the ASR accuracy [1].

Tsallis [2] proposed Tsallis entropy as generalization of Shannon entropy. Tsallis entropy has been successfully implemented to complex systems in physics such as cosmology, fractal and self-gravitating systems. The q-logarithmic (q-log) function is used in Tsallis entropy instead of the usual natural logarithmic function. The pseudo-additivity properties of this function are used to explain the nonextensive phenomena in complex systems.

In this paper, we introduce the q-log spectral domain to speech normalization. The aim is to deal with non-additivity in log spectral domain which is caused by the cross-term. We extend the log spectral mean normalization (LSMN) to the q-log spectral domain. We call this method the q-log spectral mean normalization (q-LSMN).

2 q-Log Spectral Mean Normalization

In power spectral domain, the relation between noisy speech, $|Y|^2$, clean speech, $|X|^2$, additive, $|N|^2$, and convolutive noise, $|H|^2$, is formulated by:

$$|Y(k)|^2 = |X(k)|^2|H(k)|^2 + |N(k)|^2 + \underbrace{2|X(k)||H(k)||N(k)|\cos\theta}_{\text{Cross-term}}, \quad (1)$$

where k is the frequency index, θ is the phase difference between XH and N. Additive noise can be removed using spectral subtraction (k is dropped for simplicity):

$$|\hat{X}|^2 = |Y|^2 - |\hat{N}|^2 = |X|^2|H|^2 + 2|X||H||N|\cos\theta + \varepsilon, \quad (2)$$

where $\varepsilon = |N|^2 - |\hat{N}|^2$ is inaccuracy in noise estimation. Eq. (2) shows the limitation of spectral subtraction. Even when the noise spectrum can perfectly be estimated, spectral subtraction does not remove the cross-term. In many compensation methods for robust ASR, the cross-term is assumed to be zero. However, the limited window used in STFT causes the cross-term to likely exist. This cross-term significantly degrades the ASR accuracy.

Assuming $\varepsilon = 0$, we can rewrite Eq. (2):

$$|\hat{X}|^2 = |X|^2|H|^2\left(1 + 2\frac{|N|}{|X||H|}\cos\theta\right). \quad (3)$$

In log spectral domain, Eq. (3) becomes:

$$\hat{\mathbf{x}} = \mathbf{x} + \mathbf{h} + \chi_{\text{term}}, \quad (4)$$

where $\hat{\mathbf{x}}$, $\mathbf{x}$ and $\mathbf{h}$ is the log spectral of $|\hat{X}|^2$, $|X|^2$ and $|H|^2$ respectively. χ_{term} represents the effect of the cross-term in log spectral domain:

$$\chi_{\text{term}} = \log\left(1 + 2\frac{|N|}{|X||H|}\cos\theta\right). \quad (5)$$

By similar approach as in non-extensive theory, we implement the q-log function and introduce the q-log spectral domain to consider that noise and speech could be correlated. The q-log of x is defined by $\log_q(x) = \frac{x^{1-q}-1}{1-q}$. This function recovers the natural logarithmic function when $q = 1$, but it introduces the nonadditivity when $q \neq 1$ [3]. The inverse of the q-log, q-exp, is defined by $\exp_q(x) = (1 + (1-q)x)^{\frac{1}{1-q}}$.

We would like to show that application of the q-logarithm to the conventional assumption $|\hat{X}|^2 = |X|^2|H|^2$ can easily deal with the cross-term. In the q-log domain, this equation becomes:

$$\hat{\mathbf{x}}_q = \mathbf{x}_q + \mathbf{h}_q + \kappa_q, \quad (6)$$

where $\hat{\mathbf{x}}_q$, $\mathbf{x}_q$ and $\mathbf{h}_q$ are the q-log spectrums of $|\hat{X}|^2$, $|X|^2$ and $|H|^2$ respectively. The q-log domain introduces the correlation factor κ_q:

$$\kappa_q = (1-q)\mathbf{x}_q\mathbf{h}_q. \quad (7)$$

This value is determined by the choice of q. If speech and noise are uncorrelated, $q = 1$ and $\kappa_q = 0$. If they are

---

*雑音下音声認識のための q-対数関数を用いた補正手法, 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)

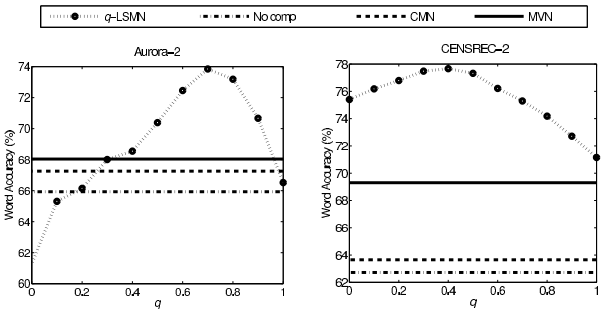

Fig. 1: Performance comparison (Word Accuracy %) of q-LSMN with CMN and MVN

correlates, then $q \neq 1$ is required. We argue that there is an appropriate q-log domain such that $\kappa = \chi_{\text{term}}$, where $\kappa = \log\left(\exp_q(\kappa_q)\right)$.

In the q-log spectral domain, the mean of $\hat{\mathbf{x}}_q$, assuming $\mathbf{h}_q$ is constant, equals to:

$$\bar{\hat{\mathbf{x}}}_q = \bar{\mathbf{x}}_q + \mathbf{h}_q + (1-q)\bar{\mathbf{x}}_q\mathbf{h}_q. \quad (8)$$

By subtracting $\bar{\hat{\mathbf{x}}}_q$ from $\hat{\mathbf{x}}_q$, and dividing it with $\left(1+(1-q)\bar{\hat{\mathbf{x}}}_q\right)$ we obtain:

$$\tilde{\hat{\mathbf{x}}}_q = \frac{\hat{\mathbf{x}}_q - \bar{\hat{\mathbf{x}}}_q}{1+(1-q)\bar{\hat{\mathbf{x}}}_q} = \frac{\mathbf{x}_q - \bar{\mathbf{x}}_q}{1+(1-q)\bar{\mathbf{x}}_q} = \tilde{\mathbf{x}}_q. \quad (9)$$

Eq. (9) is the q-LSMN formula. It is identical with LSMN when $q = 1$. We can see that q-LSMN is robust against not only convolutive noise but also κ_q. In an appropriate domain, this process corresponds to removing the cross term that may exist.

We combine linear spectral subtraction (LSS) [4] and q-LSMN to remove the additive noise. To estimate the noise spectrum, $|\hat{N}|^2$, we use the minima tracking algorithm [5] for noise estimation and a simple voice activity detector (VAD) algorithm [6] for noise updating.

3 Experiments Setup

Our proposed method was evaluated in speech recognition experiment using two databases, the Aurora-2 [7] and the CENSREC-2 [8]. We used MFCC as features which were obtained using 23 triangle mel-filterbanks. For recognition, we used 38 dimensional MFCC features which consist of 12 static features, their 1^{st} and 2^{nd}-order derivatives, Δ log energy and $\Delta\Delta$ log energy. For ASR, we implemented HMM-based system [7].

4 Results

Fig. 1 shows the word accuracies of q-LSMN without LSS, when q is varied between 0 to 1 with increment 0.1. The best performance was achieved at $q = 0.7$ for Aurora-2, and $q = 0.4$ for the CENSREC-2. The difference in optimum q might indicate that the degree of correlation is higher in real environmental situation. For both databases, q-LSMN was consistently better than CMN and mean variance normalization (MVN). For the Aurora-2, 20.1% and 18.2% relative improvement were achieved from CMN and MVN respectively. For CENSREC-2, we obtained 38.5% and 27.2% of relative improvement from CMN and MVN.

Table 1: The comparison (word accuracy %) of q-LSMN with other normalization method, when linear spectral subtraction is used

Method	Aurora-2	CENSREC-2
No compensation	65.9	62.7
LSS + CMN	72.0	63.1
LSS + MVN	69.8	69.4
LSS + q-LSMN ($q = 0.7$)	76.8	79.5
LSS + q-LSMN ($q = 0.4$)	76.4	81.4

As shown in Table. 1, the combination of LSS with q-LSMN was consistently better than the same combination with CMN or MVN. For the Aurora-2, we gained 17.2% and 23.2% of relative improvement compared to the combination of LSS with CMN and MVN respectively. For the CENSREC-2, we achieved a relative improvement of 44.5% and 33.1%.

5 Conclusions

We have proposed q-LSMN, a normalization approach in the q-log spectral domain. The use of the q-log function introduces a correlation factor that can approximate the effect of the cross-term in the log domain. Our evaluation using two types of databases showed that q-LSMN was more robust compared to CMN and MVN. For future work, we would like to investigate the relation between the q-value and the properties of speech and noise such as SNR.

References

[1] L. Deng, *et al.*, IEEE Trans. Speech Audio Process., vol. 12, 133 – 143, 2004.
[2] C. Tsallis, J. Stat. Phys., vol. 52, pp. 479–487, 1988.
[3] L. Nivanen, *et al.*, Rep. Math. Phys., vol. 52 (3), 437 – 444, 2003.
[4] D. V. Compernolle, Computer Speech and Language, vol. 3 (2), 151 – 167, 1989.
[5] G. Doblinger, in Proc. Eurospeech, 1513–1516, 1995.
[6] H. Hirsch and C. Ehrlicher, in Proc. ICASSP, 1, 153 – 156, 1995.
[7] H. Hirsch and D. Pearce, in Proc. ISCA ITRW ASR2000, 181 – 188, 2000.
[8] S. Nakamura, *et al.*, in Proc. Interspeech, 2330 – 2333, 2006.