論文 / 著書情報
Article / Book Information

Title	Generalized-Log Spectral Mean Normalization for Speech Recognition
Authors	Hilman Pardede, Koichi Shinoda
Citation	INTERSPEECH, Vol. , No. , pp. 1645-1648,
Pub. date	2011, 8
Copyright	(c) 2011 International Speech Communication Association, ISCA
DOI	http://dx.doi.org/

Generalized-Log Spectral Mean Normalization for Speech Recognition

Hilman F. Pardede, Koichi Shinoda

Department of Computer Science, Tokyo Institute of Technology, Tokyo, Japan

hilman@ks.cs.titech.ac.jp, shinoda@cs.titech.ac.jp

Abstract

Most compensation methods for robust speech recognition against noise assume independency between speech, additive and convolutive noise. However, the nonlinear nature distortion caused by noise may introduce correlation between noise and speech. To tackle this issue, we propose generalized-log spectral mean normalization (GLSMN) in which log spectral mean normalization (LSMN) is carried out in the q-logarithmic domain. Experiments on the Aurora-2 database show that GLSMN improved speech recognition accuracies by 20% compared to cepstral mean normalization (CMN) in mel-frequency domain.

Index Terms: robust speech recognition, q-logarithmic function, generalized-log spectral mean normalization

1. Introduction

The performance of automatic speech recognition (ASR) systems significantly degrades in real environments due to mismatch between training and testing conditions. Main causes of the mismatch are background noise and channel distortion which is additive and/or convolutive in time domain. Common signal or feature enhancement methods such as spectral subtraction (SS) [1], CMN [2], or more sophisticated methods such as vector taylor series (VTS) [3] have shown some success in improving the robustness of speech recognition. These methods assume that speech, additive and convolutive noise are uncorrelated.

The use of the $\log$ power spectrum has been attractive because it approximately follows the Gaussian distribution [4] and transforms convolutive relation in time domain to an additive one. However, the use of the logarithmic function may cause several problems. One of them is the nonstationary nature of noise, thus Gaussian representation may not be ideal for noise. This nonlinear property may cause some interaction between speech, additive and convolutive noise, hence the additive relation between speech and convolutive noise in the $\log$ spectrum may not hold anymore.

The q-logarithmic function has been extensively studied as a generalization of the natural logarithm. One popular implementation of the q-logarithmic function is Tsallis entropy in statistical mechanics. Tsallis entropy is an extension of Shannon entropy and has been proven effective for complex systems in physics such as cosmic, fractal and chaotic systems [5]. Tsallis entropy for equiprobable states in thermodynamics, with W as number of states and k is Boltzmann constant, is formulated as follows:

$$S_q = k \log_q(W). \tag{1}$$

Here, the notation $\log_q$ does not mean the logarithm with base q, but the generalization of the natural logarithm with parameter q. Unlike the $\log$ function, the q-logarithm does not follow the additivity property. Therefore, Tsallis entropy also does not follow the additivity property of Shannon entropy. So, if there are two subsystems, A and B, that are independent in the sense of probability theory, but the nature of the system allows interaction between subsystems, then the total entropy of the system follows the nonadditivity property as follows:

$$S_q(A+B) = S_q(A) + S_q(B) + qS_q(A)S_q(B). \tag{2}$$

In Tsallis entropy, the parameter q is used as a measure of complexity or correlation that exists in such systems. The physical meaning of q is not always known and finding it is an active area of research.

The q-logarithmic function and Tsallis entropy have also been used in robust speech recognition studies. Generalized cepstrum was proposed by Kobayashi and Imai [6] using the q-logarithmic function as a homomorphic operator for cepstral analysis. Rufiner et al. [7] proposed using Tsallis entropy as an additional feature to MFCC as complexity measures of speech signals.

In this paper, we propose generalized-log spectral mean normalization (GLSMN), which is an extension of LSMN to the q-logarithmic domain. The q-logarithmic function is used as the intermediate domain between the $\log$ and linear domain. By doing so, we can implement a nonlinear normalization that may able to capture the nonlinear relation between speech and noise, which is not well represented in CMN or LSMN.

This paper is organized as follows. The q-logarithmic function is briefly described in Section 2. In Section 3, the details of our compensation scheme, including the implementation of GLSMN, are explained. The experimental results are described in Section 4, followed by a discussion based on our findings. Section 6 concludes this paper.

2. The q-logarithmic function

The q-logarithm is a type of one-parameter generalization of the logarithmic function, and often referred to simply as the generalized-logarithm. This logarithm is defined as follows:

$$\log_q(x) = \begin{cases} \frac{x^q - 1}{q} & \text{if } q \neq 0, \\ \log(x) & \text{if } q = 0. \end{cases} \tag{3}$$

The inverse of the q-logarithm is called the q-exponential, and is defined as the following:

$$\exp_q(x) = \begin{cases} (1+qx)^{\frac{1}{q}} & \text{if } q \neq 0, \\ \exp(x) & \text{if } q = 0, \end{cases} \tag{4}$$

where $\lim_{q \to 0} \log_q x = \log_0 x := \log x$. Thus this q-logarithmic function becomes the ordinary natural logarithm when $q = 0$. The remarkable property of this q-logarithmic

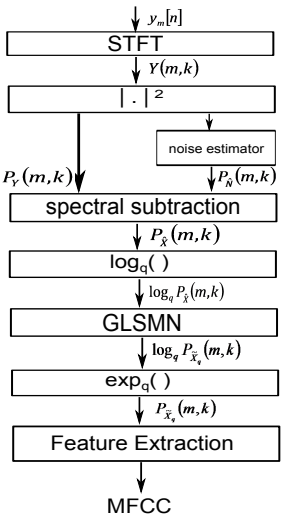


Figure 1: *Block diagram of our compensation system*

function is the nonadditivity as shown below:

$$\log_q(xy) = \log_q(x) + \log_q(y) + q\log_q(x)\log_q(y) \quad (5)$$

$$\log_q\left(\frac{x}{y}\right) = \frac{\log_q(x) - \log_q(y)}{1 + q\log_q(y)} \quad (6)$$

Eq. 5 and Eq. 6 explain the nonadditivity property of Tsallis entropy as shown in Eq. 2. The nonadditivity exists when $q \neq 0$, which is where the "complex relation" is present. When $q = 0$, Tsallis entropy becomes the same as Shannon entropy, and subsystems of the whole system are independent of each other. More details about the q-logarithm and the q-exponential and their properties can be found in [8].

3. Compensation system

3.1. Outline

In this research, the q-logarithmic function is used to provide the domain for signal compensation in the pre-processing stage before extraction of features (Fig. 1). We implement log spectral mean normalization (LSMN) in the q-logarithmic domain as described in more detail in Section 3.3. We also implement spectral subtraction (SS) before normalization.

The input speech signal $y_m[n]$ is windowed using Hamming window with a length of 25 ms with 10 ms frame shift. Here m is the frame index. After windowing, the FFT is computed and the square of the magnitude is taken to obtain the power spectrum $P_Y(m,k)$ where k is the index of the frequency bin. SS is then performed, followed by GLSMN on the output of SS. The normalized spectrum is then inverted back to the spectral domain, and a normal MFCC feature extraction is performed on the spectrum, leading the compensation system to be portable to any usual speech recognition backend. In this research, we use 23 triangle mel-filterbanks and then obtain the 12 MFCC's coefficients and log energy for feature extraction.

3.2. Spectral subtraction

Spectral subtraction (SS) is a popular method to enhance the quality of speech corrupted by additive noise. It is commonly used in speech recognition as a preprocessing stage to remove the effect of background noise. SS is formulated as follows:

$$P_{\hat{X}}(m,k) = \max(P_Y(m,k) - \alpha P_{\hat{N}}(m,k), \beta P_Y(m,k)), \quad (7)$$

where $P_{\hat{X}}(m,k)$ is the enhanced power spectrum of the k-th frequency bin of the m-th frame, $P_{\hat{N}}(m,k)$ is the estimated noise spectrum, while α and β are control parameters.

We implement several configurations of α and β. For α the values of 1.0, 2.0 and 3.0 are chosen, and we choose the values of 0.001, 0.005, 0.01, 0.05, and 0.1 as the values for β. We also implement nonlinear spectral subtraction (NSS) where a α is a function of the noisy signal-to-noise ratio (NSNR).

$$\alpha_k = \begin{cases} 1 & \text{if NSNR}_k \geq 20dB, \\ 4.75 - \frac{3}{20}\text{NSNR}_k & -5dB \leq \text{NSNR}_k < 20dB, \\ 4.75 & \text{if NSNR}_k < -5dB, \end{cases} \quad (8)$$

where NSNR_k of the k-th frequency bin is formulated as:

$$\text{NSNR}_k = 10\log\frac{P_Y(k)}{P_{\hat{N}(k)}}. \quad (9)$$

3.3. Generalized-log spectral mean normalization

Log spectral mean normalization (LSMN) is a technique used for channel compensation which is mathematically the same as CMN [9]. Both CMN and LSMN have been shown effective in removing the effect of linear convolutive distortion. LSMN is formulated as:

$$P_{\tilde{X}}(m,k) = \exp\left(\log P_{\hat{X}}(m,k) - \frac{1}{N}\sum_{i=1}^{N-1}\log P_{\hat{X}}(i,k)\right), \quad (10)$$

where $P_{\tilde{X}}$ is the normalized power spectrum and N is the total number of frames in an utterance.

Based on the generalized-log property in Eq. 6, the generalized log spectral mean normalization is derived from Eq. 10:

$$P_{\tilde{X}_q}(m,k) = \exp_q\left(\frac{\log_q P_{\hat{X}}(m,k) - \frac{1}{N}\sum_{i=1}^{N-1}\log_q P_{\hat{X}}(i,k)}{1 + q\frac{1}{N}\sum_{i=1}^{N-1}\log_q P_{\hat{X}}(i,k)}\right) \quad (11)$$

where $P_{\tilde{X}_q}$ is the normalized power spectrum after GLSMN. Based on Eq. 11, it can be seen that the GLSMN formula will be the same as LSMN when $q = 0$.

4. Experiments

Our proposed approach is evaluated using the Aurora-2 database and tasks [10]. This database is an English spoken database for digit recognition with additive noise and channel distortion added artificially at various SNRs. There are three test sets in this database: A, B and C, and two training conditions: clean and multicondition training.

4.1. Experimental setup

We use 38 dimensional MFCC features for the recognition. These features consist of 12 cepstral coefficients, their 1^{st}-order and 2^{nd}-order derivatives, Δ log energy and $\Delta\Delta$ log energy. We exclude the log energy from the features.

An HMM-based speech recognizer is used for our recogni-

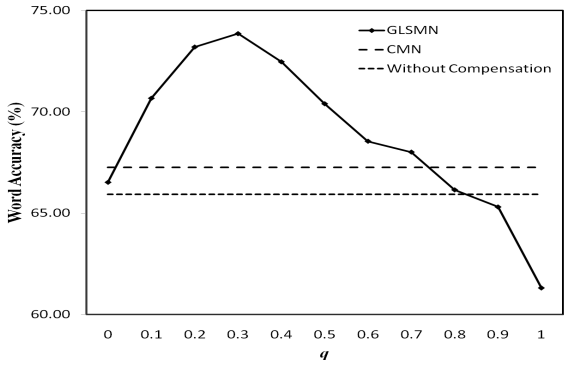


Figure 2: *Average overall word accuracies (0dB to 20dB SNR) on the Aurora-2 database, using GLSMN and CMN*

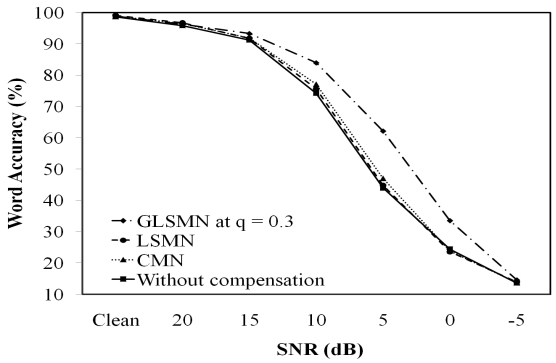


Figure 3: *Performance comparison of GLSMN and CMN for various SNR conditions*

tion experiments. Each digit is modeled by an HMM with 16 states, left-to-right, with three Gaussian mixtures for each state. Two pause models are used: "sil" model and "sp" model. The "sil" model consists of 3 states, each with 6 mixture components. The "sp" model consists of a single state which is tied with the middle state of the "sil" model.

4.2. Experimental results

4.2.1. GLSMN vs CMN

We conducted several experiments to evaluate the performance of GLSMN compared to CMN. To find the optimum value of q, we varied q from 0.0 to 1.0 with increments of 0.1. The experimental results are shown in Fig. 2. This figure shows the average word accuracies of test set A, B and C for 0-20 dB SNR condition using clean condition training. Generally, using q between 0.1 and 0.7 GLSMN outperforms CMN. The best performance is achieved at $q = 0.3$, where we achieve 20.1% relative improvement compared to CMN, and 23.3% compared to without compensation. As seen in Fig. 3, GLSMN show superiority mainly for low SNR conditions ($15dB$ to $0dB$).

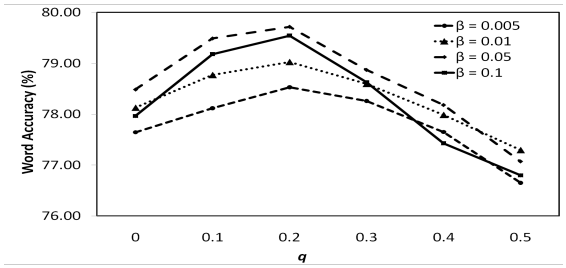
4.2.2. Combination with spectral subtraction

One factor that affects the performance of SS on speech recognition is the choice of parameter α and β. This can be seen from our experimental results shown in Table 1.

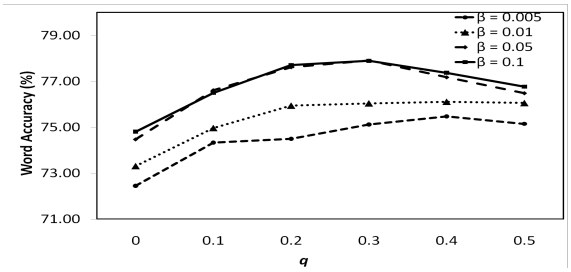
The overall word accuracies when GLSMN is performed after SS, is given in Fig. 4. The combination of SS and GLSMN outperforms the combination of SS and LSMN using the same

Table 1: *Overall average of word accuracy (set A, B, and C) (%) of the Aurora 2 database.*

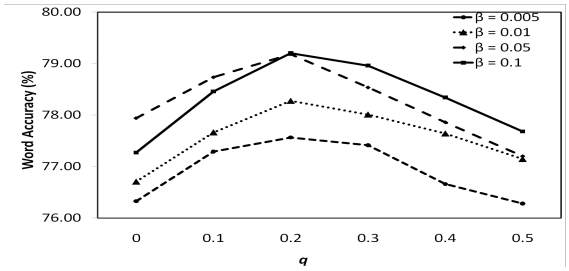
β	NSS	$\alpha = 1.0$	$\alpha = 2.0$	$\alpha = 3.0$
0.001	72.4	72.3	72.5	72.6
0.005	75.2	73.3	74.8	75.4
0.01	76.5	74.3	75.7	76.6
0.05	77.6	74.7	76.5	77.7
0.1	77.6	75.1	77.1	77.8



(a) α as function of NSNR



(b) $\alpha = 1.0$



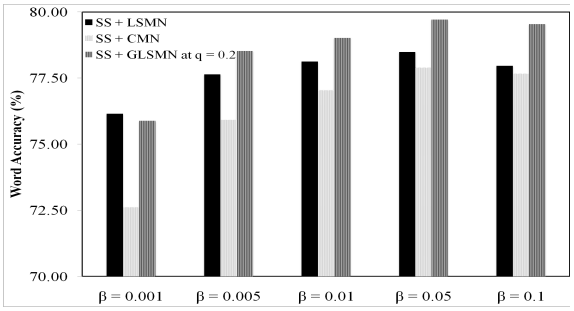
(c) $\alpha = 2.0$

Figure 4: *The performance of GLSMN when q is varied from 0.0 to 0.5 when it is combined with SS.*

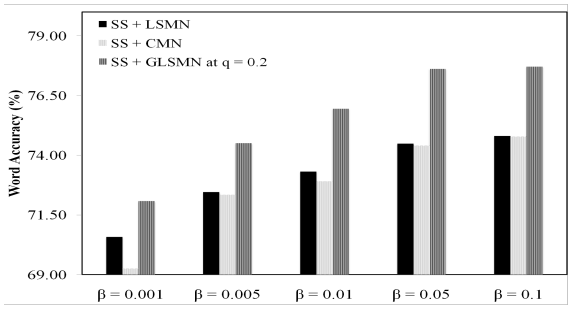
configuration of SS. As seen in Fig. 4, the best improvement is obtained when q is between 0.2 and 0.3. Combination of SS and GLSM at $q = 0.2$ achieves 5.6% relative improvement on average for all SS configurations. Compared to the combination of SS and CMN, the combination of SS and GLSMN is also generally better as shown in Fig. 5. A relative improvement of 9.9% was achieved at $q = 0.2$.

5. Discussion

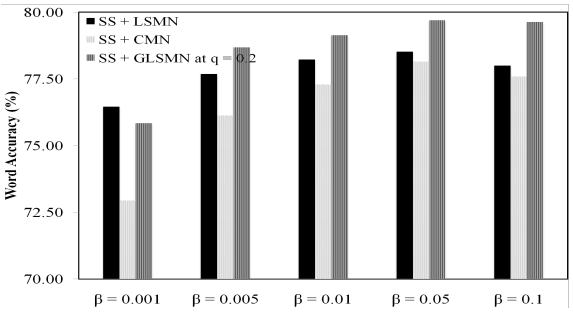
Our experiments show that GLSMN achieves higher word accuracies than LSMN and CMN. As we can see in Fig. 2, the optimum performance of GLSMN is found at around $q = 0.3$. This suggests that if we know the optimal value of q, the nor-



(a) α as function of NSNR



(b) $\alpha = 1.0$



(c) $\alpha = 3.0$

Figure 5: *Comparison of CMN and GLSMN for $q = 0.2$ when each of them is combined with SS.*

malization in the q-log spectral is better than that in the usual log spectral. These results indicate that there might exist a correlation between speech and convolutive noise which causes the system to be nonadditive in the log domain. This correlation might be introduced by the nonstationarity of the additive noise. Therefore, by having knowledge of the q to indicate that the complexity exists, we are able to further reduce the effect of noise in speech.

The q-logarithmic function is also less sensitive to the change in the low region of the power spectrum than the usual log function which is negatively large when approaching zero. Inaccuracy in the estimation of noise in the lower region of the power spectrum has a large effect in the log domain. The use of the q-logarithmic function acts as a "masking", hence the inaccuracy in noise estimation that can cause distortion in the enhanced speech, may be minimized.

The optimal value of q changes slightly when using GLSMN after SS and best performance is obtained at $q = 0.2$ for most SS configurations. This may be due to the removal of some additive noise after SS. The removal of additive noise might reduce the correlation between noise and speech, hence the optimum q value is closer to 0 ($q = 0$ corresponds to the uncorrelated case).

As in statistical mechanics, even though Tsallis entropy has shown to be successful in interpreting some physical phenomena that cannot be fully explained by Shannon entropy, the physical meaning of the q-value has not been fully explained yet by physicists. Correspondingly, in speech processing, further investigation into the relationship between the physical properties of speech and noise, and the q-values, is a topic which would be interesting.

6. Conclusions and future work

We have shown that the normalization in the q-logarithmic domain outperforms CMN and LSMN. The idea of this compensation domain comes from the assumption that the nonlinearity of noise and speech signals may introduce correlation between speech and noise, and cause a nonlinear relation in the log domain. In our experiments, our method improved the recognition accuracy by 20.1% from that obtained by CMN.

The implementation of GLSMN after SS further improved performance compared to not using SS. We obtained 9.9% relative improvement compared to the combination of SS and CMN at $q = 0.2$. The difference of the optimum q value without and with spectral subtraction might indicate the SNR as one factor affecting q.

In future work, we plan to investigate the relation between the q-value with the properties of speech and noise. We also wish to explore other compensation methods in q-logarithmic domain.

7. References

[1] S. F. Boll, "Suppression of acoustic noise in speech using spectral subtraction," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 27, no. 2, pp. 113–120, 1979.

[2] S. Furui, "Cepstral analysis technique for automatic speaker verification," *Acoustics, Speech and Signal Processing, IEEE Transactions on*, vol. 29, pp. 254 – 272, Apr. 1981.

[3] P. J. Moreno, B. Raj, and R. M. Stern, "A vector taylor series approach for environment-independent speech recognition," in *Proc. ICASSP-96*, pp. 733–736, 1996.

[4] M. J. Hunt, "Spectral signal processing for asr," in *Proc. ASRU99*, pp. 17–25, 1999.

[5] C. Tsallis, "Possible generalization of boltzmann-gibbs statistics," *Journal of Statistical Physics*, vol. 52, pp. 479–487, 1988.

[6] T. Kobayashi and S. Imai, "Spectral analysis using generalized cepstrum," *Acoustics, Speech and Signal Processing, IEEE Transactions on*, vol. 32, pp. 1087 – 1089, Oct. 1984.

[7] H. L. Rufiner, M. E. Torres, L. Gamero, and D. H. Milone, "Introducing complexity measures in nonlinear physiological signals: application to robust speech recognition," *Physica A: Statistical Mechanics and its Applications*, vol. 332, pp. 496 – 508, 2004.

[8] L. Nivanen, A. L. Mhaut, and Q. Wang, "Generalized algebra within a nonextensive statistics," *Reports on Mathematical Physics*, vol. 52, no. 3, pp. 437 – 444, 2003.

[9] C. Avendano, , C. Avendano, and H. Hermansky, "On the effects of short-term spectrum smoothing in channel normalization," *IEEE Transactions on Speech and Audio Processing*, vol. 5, no. 4, pp. 372–374, 1997.

[10] D. Pearce and H. Hirsch, "The Aurora Experimental Framework for the Performance Evaluation of Speech Recognition Systems under Noisy Conditions," in *ISCA ITRW ASR2000*, pp. 29–32, 2000.