Available online at www.sciencedirect.com

SciVerse ScienceDirect

Speech Communication 55 (2013) 587–599

SPEECH COMMUNICATION

www.elsevier.com/locate/specom

Feature normalization based on non-extensive statistics for speech recognition

Hilman F. Pardede [a,*], Koji Iwano [b], Koichi Shinoda [a]

[a] *Department of Computer Science, Graduate School of Information Science and Engineering, Tokyo Institute of Technology, Ookayama 2-12-1, Meguro-ku, Tokyo 152-8552, Japan*

[b] *Faculty of Environmental and Information Studies, Tokyo City University, Ushikubo-nishi, 3-3-1, Tsuzuki-ku, Yokohama 224-8551, Japan*

Received 15 June 2012; received in revised form 1 February 2013; accepted 16 February 2013
Available online 24 February 2013

Abstract

Most compensation methods to improve the robustness of speech recognition systems in noisy environments such as spectral subtraction, CMN, and MVN, rely on the fact that noise and speech spectra are independent. However, the use of limited window in signal processing may introduce a cross-term between them, which deteriorates the speech recognition accuracy. To tackle this problem, we introduce the *q*-logarithmic (*q*-log) spectral domain of non-extensive statistics and propose *q*-log spectral mean normalization (*q*-LSMN) which is an extension of log spectral mean normalization (LSMN) to this domain. The recognition experiments on a synthesized noisy speech database, the Aurora-2 database, showed that *q*-LSMN was consistently better than the conventional normalization methods, CMN, LSMN, and MVN. Furthermore, *q*-LSMN was even more effective when applied to a real noisy environment in the CENS-REC-2 database. It significantly outperformed ETSI AFE front-end.

Keywords: Robust speech recognition; Normalization; *q*-Logarithm; Non-extensive statistics

1. Introduction

Current automatic speech recognition (ASR) systems are able to achieve good performance in quiet environments. However, their performance significantly degrades in noisy environments. The speech features are altered in the presence of noise. This causes a mismatch between quiet training conditions and recognition conditions, which are noisy. Environmental noises are classified into two categories: additive noise and convolutive noise. Examples of additive noise are street noise, train noise, computer fan, and the voice of other persons. Examples of convolutive noise are reverberation and channel distortions.

Robust speech recognition against noise has been an active area of research for the last few decades. A number of methods have been developed in this field. Their examples are spectral subtraction (Boll, 1979), vector Taylor series (VTS) (Moreno et al., 1996) and parallel model combination (Gales and Young, 1996). All these methods are based on an *extensive* statistics in which additivity holds.

Common features used for speech recognition, such as Mel frequency cepstral coefficient (MFCC) and perceptual linear prediction (PLP), are derived from short-time power spectra of speech. In short-time processing such as short-time Fourier transform (STFT), the speech signal is processed in blocks over which speech is assumed to be stationary. This block of speech is called a frame. The length of a frame is usually between 5 and 30 ms. In the time domain, the relation between noisy speech $y(t)$, clean speech $s(t)$, additive noise $n(t)$ and convolutive noise $h(t)$ can be written as the following:

$$y(t) = s(t) * h(t) + n(t). \tag{1}$$

[*] Corresponding author. Tel.: +81 3 5734 3481.
E-mail address: hilman@ks.cs.titech.ac.jp (H.F. Pardede).

0167-6393/$ - see front matter © 2013 Elsevier B.V. All rights reserved.
http://dx.doi.org/10.1016/j.specom.2013.02.004

Denoting $x(t) = s(t) * h(t)$, we can write (1) as the following:

$$y(t) = x(t) + n(t). \tag{2}$$

By taking STFT, we can represent (2) in the frequency domain as follows:

$$Y(m,k) = X(m,k) + N(m,k), \tag{3}$$

where k is the index of frequency bin (a total frequency components $K = 256$) and m is the frame index and:

$$X(m,k) = |X(m,k)| \exp(j\theta_X(m,k)), \tag{4}$$

$$N(m,k) = |N(m,k)| \exp(j\theta_N(m,k)). \tag{5}$$

$|X(m,k)|$, $|N(m,k)|$ are the magnitude spectra, and $\theta_X(m,k)$, $\theta_N(m,k)$ are the phase spectra of filtered speech, i.e. the clean speech signal affected by convolutive noise only, and additive noise respectively. From (3), we obtain the power spectral representation of noisy speech as follows:

$$\begin{aligned} |Y(m,k)|^2 &= |X(m,k) + N(m,k)|^2 \\ &= |X(m,k)|^2 + |N(m,k)|^2 \\ &\quad + 2\mathrm{Re}[X(m,k)N^*(m,k)], \end{aligned} \tag{6}$$

where $N^*(m,k)$ is the complex conjugate of $N(m,k)$. Substituting (4) and (5) into (6), we obtain:

$$\begin{aligned} |Y(m,k)|^2 &= |X(m,k)|^2 + |N(m,k)|^2 \\ &\quad + 2|X(m,k)||N(m,k)| \cos(\theta_X(m,k) - \theta_N(m,k)), \end{aligned} \tag{7}$$

where $\theta_X(m,k) - \theta_N(m,k)$ is the phase difference between $X(m,k)$ and $N(m,k)$. The last term of Eq. (7) is called a cross-term, which depends on the phase difference between speech and noise. This cross-term is ignored in most robust speech recognition methods under the assumption that speech and noise are uncorrelated. Although this assumption is generally valid since speech and noise are statistically independent, it does not hold when applying a short-time window (20–30 ms). Several studies have shown that the cross-term does exist in short-time power spectra (Kadambe and Boudreaux-Bartels, 1992; Jeong and Williams, 1992). The cross-term has been shown to significantly degrade the performance of speech recognition (Deng et al., 2004; Faubel et al., 2008). In addition, it is well known that a speech pattern is a complex system. In a speech pattern, various long-term correlations exist among its different spectral components in complex ways in various time scales. As a consequence, the additive relation between speech components and convolutive noise may not hold in the log spectral domain.

It is common to combine additive noise removal methods such as spectral subtraction with feature normalization methods such as cepstral mean normalization (CMN) (Furui, 1981) to remove both additive and convolutive noise. But as previously explained, speech and convolutive noise are not additive in general, and thus the cross term exist even when the additive noise spectra are completely removed.

In this paper, we propose q-log spectral mean normalization (q-LSMN) (Pardede and Shinoda, 2011), which is an extension of the log spectral mean normalization (LSMN) (Avendano and Hermansky, 1997) to the q-log spectral domain. We further investigate the effect of q-LSMN in various conditions and analyze its property in more detail in this paper.

A few studies have already employed the non-extensive statistics for speech recognition. Rufiner et al. (2004) added Tsallis entropy, which is defined in non-extensive statistics, and its relative change to the standard MFCC features for capturing the dynamics of speech signals. Kobayashi and Imai (1984) employed the q-log function as the spectral smoother for speech features so that they are more robust especially in lower frequency regions. Ito et al. (2000) also implemented the q-log function to provide a forward masking scale for the dynamic cepstrums. In contrast to these studies, we use the q-log function to model non-additivity in noisy speech features.

The remainder of this paper is organized as follows: In Section 2, we describe some previous studies of robust speech recognition related to our study. In Section 3, we explain the influence of the cross-term in the speech features. In Section 4, we briefly review non-extensive statistics and its q-log function. In Section 5, we explain our proposed method. In Section 6, we describe the details of spectral subtraction which we implemented to remove additive noise. In Section 7, we describe the experimental setup to evaluate our proposed method. In Section 8, we present and discuss our experimental results. Section 9 concludes this paper.

2. Related studies

Various methods have been proposed in the past literature for improving the robustness of speech recognition in noisy environments. Generally, they can be categorized into two groups: feature enhancement and model compensation. In feature enhancement, the aim is to estimate clean speech features in noisy speech by removing noise. Whereas, model compensation methods adapt clean speech models to noisy conditions, by considering the noise statistics. They can usually achieve better performance than feature enhancement-based methods. On the other hand, they require higher computational cost and more data than feature enhancement-based methods. In addition, they should update models each time a new type of noise is introduced. In this section, we first introduce several methods in extensive frameworks, and we describe several variants that take into account the correlation between speech and noise.

2.1. Feature enhancement

Spectral subtraction (Boll, 1979) is a popular method to remove additive noise in the spectral domain. In spectral