論文 / 著書情報
Article / Book Information

題目(和文)	
Title(English)	Robust Speech Recognition Based on Non-extensive Statistics
著者(和文)	ﾋﾙﾏﾝﾊﾟﾙﾃﾞｯﾄﾞ
Author(English)	Hilman Pardede
出典(和文)	学位:博士(工学), 学位授与機関:東京工業大学, 報告番号:甲第9248号, 授与年月日:2013年6月30日, 学位の種別:課程博士, 審査員:篠田 浩一
Citation(English)	Degree:Doctor (Engineering), Conferring organization: Tokyo Institute of Technology, Report number:甲第9248号, Conferred date:2013/6/30, Degree Type:Course doctor, Examiner:Koichi Shinoda
学位種別(和文)	博士論文
Category(English)	Doctoral Thesis
種別(和文)	論文要旨
Type(English)	Summary

（博士課程）
Doctoral Program

論 文 要 旨

THESIS SUMMARY

専攻： Department of	Computer Science 専攻	申請学位（専攻分野）： Academic Degree Requested	博士 Doctor of (Engineering)
学籍番号： Student ID Number		指導教員（主）： Academic Advisor(main)	篠田 浩一
学生氏名： Student's Name	Hilman Ferdinandus Pardede	指導教員（副）： Academic Advisor(sub)	

要旨（英文 800 語程度）
Thesis Summary (approx.800 English Words)

The existence of environmental noise introduces a mismatch between clean training conditions and recognition conditions which are noisy, hence significantly degrades the accuracy of speech recognition systems. A number of methods have been developed to reduce this mismatch and improve the robustness of speech recognition in noisy environments. Their examples are spectral subtraction, cepstral mean normalization (CMN), vector Taylor series (VTS), and parallel model combination (PMC). These methods are derived on an extensive framework, which is based on Boltzmann-Gibbs statistics and Shannon entropy. Under this framework, it is assumed that noise and speech are uncorrelated. Therefore, the linear addition between speech and noise holds. It is well known that Gaussian distributions maximize Shannon entropy under appropriate constraints. For that reason, they are used to model speech and noise. Accordingly, noisy speech is also assumed to follow a Gaussian distribution.

A speech pattern however 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. Therefore, it is not surprising that the short-time speech spectra do not follow Gaussian distributions, but show heavy-tailed distributions instead. For this reason, it is very likely that the short-time spectra of noisy speech do not follow Gaussian distributions either. In addition, several studies have shown that the cross-term exists in the short-time power spectra of noisy speech. This cross-term has been shown to significantly degrade the performance of speech recognition. These inaccurate assumptions hamper the performance of the robust speech recognition methods. It is found that the same accuracy as when using clean speech cannot be obtained even when noise spectra are accurately estimated.

Recently, a theory of non-extensive statistics has been introduced to explain several phenomena in complex systems. This framework is a generalization of Boltzmann-Gibss statistics. This framework relies on two non-extensive functions, q-exponential (q exp) and q-logarithmic (q -log) functions. In this framework, the entropy is redefined, which is called Tsallis entropy. This entropy is a generalization of Shannon entropy and a non-extensive one. Also in this

framework, heavy-tailed type distributions can be derived. One of them is called the q-Gaussian distribution. The q-Gaussian distribution is a generalization of the Gaussian distribution. This framework has successfully represented many phenomena in complex systems in statistical mechanics, economics, finance, biology, astronomy and machine learning.

In this thesis, we implement the framework of the non-extensive statistics for robust speech recognition to solve the problems of the cross-term and non-Gaussianity of noisy speech. We propose two methods: q-log spectral mean normalization (q-LSMN) and q-spectral subtraction (q-SS),

Our first method, q-LSMN, is a feature normalization method. Feature normalization is a well established method to remove convolutive noise. Conventional feature normalization methods are usually performed either in the cepstral domain or the log spectral domain. In the conventional approaches, the logarithmic function is utilized to transform a convolutive relation between speech and convolutive noise in the time domain into an additive one in the log spectral or cepstral domain. However, the cross-term introduces nonlinearity between speech and convolutive noise. It cannot be removed by conventional methods. In q-LSMN, we utilize the q-log function and its properties and introduce a new domain, q-log spectral domain. By doing so, we can naturally represent noisy speech including the cross-term. We derive a feature normalization technique in this domain, which is an extension of the log spectral mean normalization (LSMN) to the q-log spectral domain. Our method not only normalizes speech features but also removes the cross-term. Our experiments on a synthesized noisy database, Aurora-2, and a real noisy environment database, CENSREC-2, confirmed that q-LSMN outperformed the traditional feature normalization methods.

Our second method, q-SS, is a spectral subtraction method, which is a popular method for removing additive noise. In spectral subtraction, it is assumed that noise and speech are uncorrelated and follow Gaussian distributions. Thus, noisy speech also follows a Gaussian distribution. Spectral subtraction is derived by maximizing the likelihood of noisy speech with respect to its variance. In q-SS, we assume that noisy speech follows the q-Gaussian distribution. We then derive q-SS in a similar way as spectral subtraction is derived. We found that the q-Gaussian distribution fits the noisy speech distribution better than the Gaussian distribution does. Our speech recognition experiments confirmed the effectiveness of q-SS compared to the conventional spectral subtraction method on the Aurora-2 and the CENSREC-2 tasks.