早稲田大学大学院 基幹理工学研究科

博 士 論 文 概 要

論 文 題 目

Nonstationary Signal Processing and Confidence Weighting for Robust Speech Recognition

非定常信号処理と信頼度重みづけによるロバスト音声認識

申 請 者

Lingnan	G E
葛	菱南

情報・ネットワーク専攻 ヒューマンインタフェース研究

2010 年 12 月

The Automatic Speech Recognition (ASR) is one of the key technologies in the field of the Artificial Intelligence (AI), moreover ASR is the most natural and convenient human-machine interface. Because of the advance of the computer hardware, software technology, inspiration of the huge demands, and especially the researchers' continuous efforts, ASR has won the proud achievements in both the theoretical researches and practical applications since it appearing at 50's last century.

However, the human speech is the outcome of accumulation of culture and productive activities for the thousands of years, which makes that human speech becomes quite various and complex. Hence the ASR evolves not smoothly. A recognition system is always hard to deal with various noise disturb even if it is an excellent system. Besides, the system is hard to deal with vary speaker with different age, sex, dialect, accents, speaker's feeling and spoken habit. Seaker would raise his tone under a noisy environment, so called Lombard effect. Such situations make researches for adaptability. Today, ASR has met with barriers of robustness and adaptability. ASR theory and system need new break through.

We think that the essential and natural character of acoustic speech signal is a complex and various random processes which is the result of alternating and overlapping continuously stationary processes and/or nonstationary processes. However, the current theory and technique of ASR cannot be fit with the natural character of acoustic speech digital signal. The features used in current recognition theory and system of ASR are obtained based on certain stationary and linear hypothesis. The parameter model and recognition model are immutable and frozen as a system is running. For example, log spectrum, LPC-Cep or Mel-Frequency Cepstral Coefficient (MFCC) with some fixed order as recognition feature, and a score based on Hidden Markov Model with Gaussian mixture density (GMD-HHM) (Rabiner and Juang, 1986) as recognition criterion are used. Thus, the natural character of complex and dynamical random process is not sufficiently considered in the current systems, which is perhaps the important reason that robustness has become one of the focuses of researches in ASR, especially in noisy environments. Consequently considerable attention has been paid to developing statistical theory and techniques to deal with noisy ASR (Jelinek, 2001) in the past three decades.

Widely used LPC-Cep and MFCC are not the best parameters. Since LPC-Cep based on AR (Auto-Regression model), some hypotheses of stability and linearity paradigm for digital signals are needed. Meanwhile, the order of AR model has to be selected high enough to simplify ARMA (AR-Moving Average) into AR，in theory being AR(∞),. Consequently, the enough-high-dimension must bring expensive both computational space and time. Although utilization of MFCC usually has got better result than LPC-Cep, MFCC based on short-time Fourier transform will not lose the condition of harmonics, or stationarity in the case of considering random situation.

Consequently we cannot distinguish some nonstationary signals, especially some consonants, in a parameter space with dimension as high as 50 yet, including 16-order-MFCC and their first two differences. So we need a change of basis features for ASR.

We also prove nonstationarity of speech signal process and weakness of theses used features by several statistical tests and recognition experiments in Sec. 2.2 and Sec. 3.3.2.

The thesis tries to improve the robustness in following three aspects:

I. finding some indexes to describe such natural character of the complex and dynamical speech process;

II. exploring more available mixed 2-order parameter model;

III. getting more powerful approach to deal with noise affectation.

They are concerning to nonstationarity measure with the help of a statistical test, nonlinear Doubly Random Time Series (DRTS) and its parameter estimate, and Approach of Feature with Confident Weight (AFCW) as an improvement on current noisy ASR in a new direction, respectively.

In chapter 2, this thesis selects the more rapid and simple inverted sequence test of two non-parameter tests suggested by Ge, L. et al. (2004) to get nonstationarity measure and dynamically adapt to the acoustic speech. This measure is also applied to determine some key points cutting syllable and segment speech, to supplement new features of the signal, and to pick up the parameter models.

Then in next chapter 3, a type of DRTS (Doubly Random Time Series), AR(p)-MA(q), is introduced. The DRTS suggested by Ge, L. et al. (2004) and Ge, Y. et al. (2001) is, in fact, a powerful 2-order time series that covers AR. Bayesian estimate for a type of DRTS is given in details and how to organize efficiently and automatically Bayesian estimate and the moment estimation is suggested here for a recognition system.

In noisy ASR, a recognition system performs mostly poorly. In the last recent years, Missing Feature Approaches (MFAs) have shown their progresses in noisy ASR (e.g. Cerisara et al., 2007; Barker et al., 2000; Cooke et al., 2001; Penevey and Drygajlo 2000). However, hard mask based on selecting threshold value and binary classification is rough and risk. Soft masks, including ones improved by Bayesian technology (Seltzer and Richard, 2004), meet with difficulty essentially to exploring joint or conditional probability density functions (*p.d.f.*s) of reliable and unreliable feature components. Sigmoid function employed by the most of soft mask works is not reasonable *p.d.f.* and only is substitute. In our opinion, it is false since it's one-peaky and symmetric. Barker et al. also wrote that "In practice the noise estimation error is only likely to be Gaussian if we have a good model of the noise. The missing data approach however attempts to avoid employing noise dependent models. In the current work we employ a simple stationary noise estimate for all

noise types. For nonstationary noises the error in the estimate is likely to have a non-Gaussian distribution. Accepting this, we have not attempted to compute ideal fuzzy masks, but have instead generated a mask of values between 0 and 1 by compressing x with a simple sigmoid function". Meanwhile, MFAs cannot be straightforwardly employed only in cepstral domain and then lost recognition accuracy. Although, accurately identifying missing parts remains a very challenging task, we also need some change in the idea.

So following, in the chapter 4, we extend the notion of Confident Weight (CW) (Ge, Y. et al., 2004) to four classes and set up the corresponding criteria to replace the thresholds and *p.d.f.*s needed by MFA. We give a novel analysis of confident degrees of feature component in each subband based on four criteria: Energy, Signal-to-Noise Ration (SNR), statistics on SNR, and statistics on a new Rate between signal and noise. And then we suggest four classes of confidence weights in this section. Renewal AFCW describes the noise effect in a more precise way, gives an available reliable degree of feature components, and simply deals with reliable components and unreliable ones under a uniform framework. Furthermore, AFCW can be extended easily in cepstral domain and consequently AFCW in this time brings the system to more significant improvement.

AFCWs suggested in this chapter significantly enhance real speech signals under an inverse noisy environment, including stationary and nonstationary noises. Any AFCW does need neither threshold nor joint *p.d.f.* of reliable component and unreliable one, and the reliabilities of all components are finely computed under a uniform framework and the effects of noise are described simply and accurately.

Under the framework of GMD-HMM with Demi-syllable Unit (DSU) (Ge, Y. et al., 1992), experimental results in chapter 5 show that proposed approaches could improve the recognition accuracy significantly in adverse environment, including stationary and nonstationary noisy environments. We prefer select DSU as recognition unit in our experiments, which make us focus on dealing with nonstationary part of a syllable and avoid the difficulty of cutting out a consonant from a syllable. Since the correct recognition rate of a vowel syllable is much higher than a consonant one, we focus the recognition techniques on Consonant-Vowel (CV)-unit consisting of the whole consonant and the head part of connected vowel of Chinese syllable.

Since the essential character of speech signal, the ASR theory and system should possess the ability of self-adaptable in their inner. Now we can set up dynamic recognition theory and system with self-adaptive ability in the inner, but not in the output, of the system based on proposed technique and stochastic process theory discussed in the thesis, which is different from those auto-adaptable theory and system that need some adaptable process based on outer environment. The system ability is enhanced greatly with the help of proposed features and Bayesian

estimation.

The last chapter is our conclusions and expected further researches.

We summarize my works as following:

1) Based on statistical theory and techniques, combing with the different technique, do data-mining and propose a group of new features, called nonstationarity measure, which could be used to describe the nonstationarity of the speech signal to be recognized and develop its other uses esp. We prove that nonstationarity measure is important supplement to traditional features, like LPC-Cep and Mel-Cep in both statistical theory and recognition experiments.

2) Explore a type of DRTS (Doubly Random Time Series DRTS), essentially being a mixed square random model, and give simple moment estimation and Bayesian one of model parameters with point-view of RCA (random coefficient regression model) , upon which set a new group of features, DRTS Group of Features, to supplement again to traditional features.

3) Join both two new feature groups, lower-order traditional features in the final Extended Features, which raises robustness and adaptability of recognition system.

4) Put forward a proposal of CW (Confident Weight) in four meanings and AFCW (Approach of Feature with Confident Weight) to handle noise effects in unified framework and accurate manner, which significantly enhance speech and improve robustness and adaptability and don't need to find the joint or conditional probability density functions of reliable and unreliable feature components on spectrum domain and calculating their integral mean except determination of thresholds having risk, like MDT. And give also realization of proposed approaches in popular cepstum domain.

5) Give a proposal to construct a dynamic recognition system with self-adaptive ability in the inner, but not in the output, of the system based on based on some dynamical indexes like nonstationarity measure, and selection of more models and more parameters. These fit quite well the essential character of acoustic signal and shows it be very hopeful.

6) Do compare experiments on several recognition approaches of ASR with the help of 2-order HMM model and recognition unit of expanded Demi-Syllable Unit (DSU) , which demonstrates that our approaches suggested here ensure excellent performances.

早稲田大学　博士（工学）　学位申請　研究業績書

氏 名　葛　菱南　　印

（２０１０年１２月　現在）

種 類 別	題名、　発表・発行掲載誌名、　発表・発行年月、　連名者（申請者含む）
論文　○	Approach of feature with confident weight for robust speech recognition, Acoustical Science and Technologies, Vol.32, No.3 (to appear), Lingnan Ge, Katsuhiko Shirai, Akira Kurematsu
論文	Non-Linear Techniques For Robust Speech Recognition, 4th International Conference on Cybernetics and Information Technologies, Systems and Applications, pp 134-138, Jul 2007, IDS Number BHL79, ISBN 978-1-934272-24-4; INT INST INFORMATICS & SYSTEMICS, FL 32837 USA, Orlando, USA, Yubo Ge, Lingnan Ge, Katsuhiko Shirai
論文	Feature Parameters and Confident Weights for Robust Speech Recognition under Noisy Environment, Proc. 2nd IASTED International Conference on Computer Intelligence 2006, pp 309-313, ISBN: 978-0-88986-602-7, ACTA PRESS, IDS Number BFQ37, San Francisco, USA, Nov 2006, Yubo Ge, Lingnan Ge, Katsuhiko Shirai
論文	Robust Speech Recognition Based on Dynamical Selections, 2006 Proc. 2nd IASTED INTERNATIONAL CONFERENCE ON COMPUTATIONAL, pp 314-317, ISBN: 978-0-88986-602-7, ACTA PRESS, IDS Number BFQ3 7, San Francisco, USA, Nov 2006, Lingnan Ge, Katsuhiko Shirai, Yubo Ge
論文	Dynamic Robust Speech Recognition, ISCCSP 06 2nd International Symposium on Communications, Control and Signal Processing, ISBN: 2-908849-17-8, pp59, MOROCCO, Mar 2006, Yubo Ge, Lingnan Ge, Katsuhiko Shirai
論文	Enhancing Robustness of Speech Recognition by Approach of Feature with Confident Weight, Proc. IASTED International Conference on Artificial Intelligence and Applications 2006, pp 115-119, ISBN 0-88986-556-6, IDS Number BEA05, AUSTRIA, Feb 2006, Lingnan Ge, Katsuhiko Shirai, Yubo Ge
論文	Translation three Chapters of ‘Applied Multivariate Statistical Analysis’, R. A. Johnson, D. W. Wichern, Edi. 4), consisting of Ch8 Principle Analysis, Ch9 Correlation Analysis and Ch10 Canonical Analysis. (published by Tsinghua University Press).
論文	Approach of Feature with Confident Weight for Robust Speech Recognition, IEEE Multimedia Signal Processing IEEE Catalog No: 04TH8761C ISBN: 0-7803-8578-0；pp11-14; IDS Number：BBC43, Siena, Italy, September 2004, Yubo Ge, Jun Song, Lingnan Ge, Shirai,K.
論文	Nonlinear Random Features of Non-stationary Signals and Applications to Speech Recognition, SPECOM, Proceedings of the 9th International Conference “Speech and

早稲田大学　博士（工学）　学位申請　研究業績書

種 類 別	題名、　発表・発行掲載誌名、　発表・発行年月、　連名者（申請者含む）
	Computer”, ISBN:5-7452-0110-x pp141-145; Saint Petersburg, Russian, September 2004, Lingnan GE, Katsuhiko Shirai, Yubo GE
論文	Confidence Weighting Missing Feature Approach for Robust Speech Recognition, EURASIP, 2004 European Signal Processing Conference, pp 337-340, ISBN 3-200-00165-8; Vienna, AUSTRIA, Sep 2004, Yube Ge, Jun Song, Lingnan Ge
論文	Dynamic Techniques Based On Statistics To Recognize Non-Stationary Signals, 4th International Conference on Statistics, Mathematics and Related Fields, ISSN: 1550-3747 pp282-285, Hawaii, USA, June 2004, Lingnan Ge, Katsuhiko Shirai, Yubo Ge
講演	Dynamic Non-Linear Techniques for Speech Recognition, 日本音声音响学会 2005 年春季学会, 2005 年 03 月, Lingnan Ge, Katsuhiko Shirai,