DOE and ANOVA based Performance Influencing Factor Analysis for Evaluation of Speech Recognition Systems

Xiangdong Wang[1, 2], Feng Xie[1, 2], Shouxun Lin[1], Yuelian Qian[1], Qun Liu[1]

[1]Institute of Computing Technology, Chinese Academy of Sciences,
Beijing 100080
[2]Graduate University of Chinese Academy of Sciences
Beijing 100085
{xdwang, xiefeng, sxlin, ylqian, liuqun}@ict.ac.cn

Abstract. In this paper, a framework of performance influencing factor analysis (PIFA) is proposed for evaluation of speech recognition systems. For each system under evaluation, the influence of various data properties (e. g. accent, SNR, and speaking rate) on system performance is analyzed to show corresponding characteristics of the system. The main idea of this approach is to design test data by means of Design of Experiments (DOE) methods and perform analysis of variance (ANOVA) and other statistical analyses on the performance of each system. To use ANOVA which is far more powerful than rank tests generally used in performance evaluation, a method is proposed for generating performance measurement data satisfying the two basic assumptions of ANOVA. The PIFA approach is applied to the 2004 HTRDP ASR Evaluations, and analysis results are reported and interpreted, which are considered helpful for pointing out virtues and deficiencies of each system and accelerating improvement.

Keywords: PIFA**,** ASR, Speech Recognition, Evaluation, DOE, ANOVA.

1 Introduction

By testing all systems on the same task with the same test corpus and performance metrics, an evaluation of speech recognition systems can show differences in performance metrics among systems and facilitate improvements by promoting the use of superior techniques. For speech recognition, it is well accepted that some data properties such as accent and SNR can influence system performance considerably and thus pose a major challenge to current techniques. So in evaluations, participants are eager to know how the systems work on these data properties and which system is the most robust, for this is of great help for improving system performance by adopting and studying the best algorithms. However, in current evaluations, little work has been done concerning these requirements. Evaluation reports from NIST [1] gave information of data properties such as SNR and speakers, but didn't point out how they affect the performance. Other researchers [6, 7] conducted experiments using test sets with different values of one data property (e. g. test sets with different

SNRs) to investigate its impact on recognition performance. But this approach is inappropriate to be used in evaluations, as it requires a number of additional test sets besides the original test corpus. Furthermore, current approaches [6, 7] only report differences in metrics (e.g. WERs), without any analysis of statistical significance of the differences.

In this paper, we propose the Performance Influencing Factor Analysis (PIFA) framework for evaluation of speech recognition systems. It is based on statistical methods such as Design of Experiment (DOE) [10] and analysis of variance (ANOVA) [11, 12], which are popular used as standard procedures in such fields as agriculture, medicine, and manufacture but seldom used in the field of speech recognition or evaluation. Under the PIFA framework, for each system, the influences of data properties (referred to as "performance influencing factors") on recognition performance is analyzed, showing whether the influence is statistically significant and giving the degree of the influence. This actually gives information of robustness to the data property for all systems, so robust algorithms can be compared and improvement can be made by adopting and developing better techniques. The PIFA framework was applied to the 2004 High Technology Research and Development Program (HTRDP) Automatic Speech Recognition (ASR) Evaluations [8, 9] and considered helpful by most participants.

The rest of this paper is organized as follows: In Section 2, details of the PIFA framework is presented. The application of the PIFA approach in the 2004 HTRDP Automatic Speech Recognition Evaluation is described in Section 3 and analysis results are displayed and interpreted in the same section. Finally, in Section 4, conclusions are drawn and future work is predicted.

2 Performance Influencing Factor Analysis

2.1 Definitions and Analysis Flow in PIFA

First, some terms used in PIFA are defined and interpreted as follows.

Performance influencing factor: data properties which may influence system performance and are investigated in PIFA are referred to as *performance influencing factors* (for short as *factor*). In other words, if the system performance varies much while the value of a data property changes, the property can be considered as a factor.

Levels of a factor: values or classes of a factor are called *levels*. Values of a factor may be either discrete or continuous. For example, the factor "dialectal accent" has two values "with-accent" and "without-accent", and the values of SNR are continuous. In PIFA, only discrete levels are used, so levels of factors with continuous values must be decided by dividing the value domain into several intervals. For example, the values of SNR are divided into three intervals: $(-\infty, 11\text{dB}]$, $(11\text{dB}, 14\text{dB}]$, and $(14\text{dB}, +\infty)$, and each interval is considered as a level.

The purpose of PIFA is to investigate whether the difference in performance caused by data with different levels of a factor is statistically significant for a system. The method used by most researchers for similar tasks is to collect a group of data for each level [6, 7]. For example, when investigating the factor SNR, three groups of

data may be collected: one with SNR of 10dB, another with SNR of 15dB, and the third with SNR of 20dB. This simple method has a major deficiency: if data are divided into groups only considering levels of one factor, distribution of levels of other factors maybe different across groups, resulting in difference in performance. For example, difference of performance between the data with SNR of 10dB and 15dB maybe caused not only by difference in SNR, but also by difference in accent, if most speech data in the 10dB group are with accent and data in the 15dB group are native speech. This will lead to misleading analysis results. A resolution is to make the data in different groups only differ in one factor, for example, to make the data in the three groups mentioned above all the same except for SNR. But in this way, when dealing with multiple factors, the test data will be quite large in quantity with great redundancy, making them inappropriate for use in evaluations.

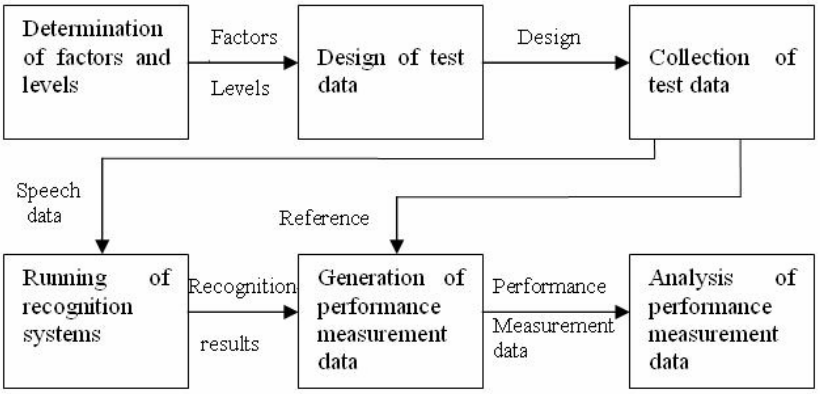

Fig. 1. Analysis flow of PIFA

To avoid this deficiency, in the PIFA framework, we adopt the idea of Design of Experiment (DOE) which is commonly used in the agriculture, medicine, and manufacture fields [10, 11]. The main idea is to divide the test data into several groups, each groups stands for a combination of levels of factors. Some experimental design and statistical analysis methods are developed to make sure effect of other factors is taken into account when investigating a certain factor, making results more reliable. Orthogonal design and analysis of variance (ANOVA) are adopted in PIFA because they are the most popular and powerful methods for design and analysis in DOE.

The analysis flow of PIFA is shown in Fig.1. First, after factors are chosen and their levels are decided, the test data are designed and collected as several groups according to DOE methods. Then, when recognition results are available, calculation is performed in every group to generate a series of performance measurement data satisfying requirements of ANOVA. Finally, the ANOVA procedure [11, 12] is performed to determine significance. And, if there is more than one factor, a process similar to *range analysis* [10] is also performed to decide the order of factors. As can be seen, there are three major steps which will be details in the rest of this section:

design of test data, generation of performance measurement data, and analysis of performance measurement data.

2.2 Design of Test Data

For a given task, first, several performance influencing factors must be chosen according to experience and data availability. It is well accepted that there are many factors influencing performance of speech recognition systems, such as accent, SNR, speaking rate, etc. Then, levels of each factor should be decided. For factors such as "speaker gender" or "dialectal accent", it can be easily done. For factors with continuous values, such as SNR and speaking rate, the set of values should be divided into several intervals and each interval is considered as a level. How to divide the values depends on the distribution of values of the factor.

As mentioned above, the test data are designed according to orthogonal design. Under orthogonal design, groups of data are determined according to some orthogonal tables given by statisticians, and not all combinations of levels are considered. For example, table 1 shows an orthogonal table denoted as $L_9(2^1 \times 3^3)$. It can be used in situations involving 4 factors, one with 2 levels and the others with 3 levels each. As can be seen from table 1, in data with one level of a factor, all level combinations of other factors can be found, and for two factors, all combinations of their levels can be found in the data. So analyses such as ANOVA can be performed on the 9 groups shown in table 1 instead of $2^1 \times 3^3=54$ groups, taking effect of other factors into account when a certain factor is examined.[10].

When factors and levels are decided, a suitable orthogonal table must be chosen considering the number of factors and levels [10]. Then, groups of combination of levels are determined according to the orthogonal table. And, finally, test data are collected to assure that there are enough data in each group.

Table 1. The orthogonal table $L_9(2^1 \times 3^3)$

Group	Factor 1	Factor 2	Factor3	Factor4
1	Level 1	Level 1	Level 1	Level 1
2	Level 1	Level 2	Level 2	Level 2
3	Level 1	Level 3	Level 3	Level 3
4	Level 1	Level 1	Level 2	Level 3
5	Level 1	Level 2	Level 3	Level 1
6	Level 1	Level 3	Level 1	Level 2
7	Level 2	Level 1	Level 3	Level 2
8	Level 2	Level 2	Level 1	Level 3
9	Level 2	Level 3	Level 2	Level 1

2.3 Generation of Performance Measurement Data

After systems are run with the test data, recognition results are produced for each system. In the PIFA framework, for each system, performance measurement data

must be calculated for each group of test data to provide input for further analyses. The most natural way is to calculate evaluation metrics such as *word error rate (WER)* in each group and use them as performance measurement data. So in PIFA, performance measurement data of N numbers are generated using the following algorithm.

```
randomly divide speech data in the group into N subsets
for i= 1 to N
calculate the evaluation metric (e.g. WER) on the ith subset
end for
```

The number N is a constant set arbitrarily due to data amount and other reasons. In our experience, 6-12 is appropriate for most situations.

There is one issue must be mentioned. That is, the use of ANOVA (which will be introduced in 2.3) requires two assumptions to be satisfied: (1) the performance measurement data of each group are normally distributed and (2) the variances of performance measurement data in different groups are identical. In some situations these two assumptions are not satisfied, so ANOVA cannot be used. For example, significance tests of NIST use sign test and sign-rank test [1-5] which are less powerful than ANOVA, because the performance measurement data generated violate the two assumptions.

As for our method described above, it actually simulates a procedure of a series of independent, repeated experiments, with each subset as a test set. So if data are not too few in each subset, the performance measurement data generated will be normally distributed according to the central limit theorem. And for a given system, the variances of performance measurement data in different groups are not considerably different. This is also demonstrated by experimental results given in 3.2.

2.4 Analysis of Performance Measurement Data

A major feature of the PIFA method is that it can perform statistical analysis considering multiple factors at one time, while other methods with similar purpose only compare performance metrics on data with different levels of one factor. This statistical analysis in mainly based on ANOVA, which is very efficient in dealing multiple factors and far more powerful than sign test or sign-rank test adopted by other researchers. ANOVA (analysis of variance) is a kind of hypothesis test method which tests for significance differences between means by means of dividing the sample variance into several parts. As mentioned in 2.2, feasibility of ANOVA depends on normality and homogeneity of variances of data, which are satisfied by the data generation method proposed in this paper.

The result of ANOVA indicates whether performance on data of different factor levels are significantly different, thus indicates whether a factor influences system performance significantly. For multiple factor situations, it is also wanted that which factor has the strongest influence. So the method of *range analysis* [10] into our PIFA method is incorporated into PIFA as follows. After the procedure of ANOVA, for a factor, means of performance measurement data of each level are listed, as show in

Table 3. For a factor, the *range of means* is defined as the difference of the maximum and minimum means of all levels. And *range proportion is* defined as follows.

$$\text{range proportion} = \frac{\text{range of means for a factor}}{\text{performance metric of the system on all test data}} \quad . \qquad (1)$$

It is inferred that for a system, the factor with the largest range of means has the strongest influence on the performance. And for a factor, the system producing the largest *range proportion* is the most strongly influenced system by that factor.

3 Application of PIFA in 2004 HTRDP ASR Evaluation

3.1 Details of the Application of PIFA in 2004 HTRDP ASR Evaluation

The HTRDP Automatic Speech Recognition Evaluation, sponsored by National High Technology Research and Development Program (HTRDP) of China, is part of an ongoing series of evaluations of Chinese information processing and intelligent human-machine interface technologies. The 2004 HTRDP ASR Evaluation is consisted of three major tasks: LVCSR (large vocabulary continuous speech recognition) for reading speeches on PC, keyword recognition for telephone speech, and command recognition on PDA.

The LVCSR task focused on recognition of reading speech data collected in various noisy environments. The test set includes 200 utterances, each of which is a sentence of 7 to 15 seconds. The utterances are spoken in Chinese Mandarin, but some speakers are with slight dialectal accent. The main evaluation metric used is Character Error Rate (CER).

The PIFA method was used for analyzing the results of three systems which achieved highest performances in the 1x real-time LVCSR task. The factors chosen are dialectal accent, SNR, and speaking rate, for they are well accepted by researchers as performance influencing factors. Each utterance is considered as a piece of data. For each utterance, the value of dialectal accent of each speech data is decided subjectively by people, and speaking rate are calculated by the simple formula

$$\text{speaking rate} = \frac{\text{number of Chinese characters}}{\text{duration of the whole speech data}} \quad . \qquad (2)$$

As for levels, it is decided that dialectal accent has 2 levels: with-accent and without-accent, because most dialectal accents are slight. For SNR and speaking rate, in order to divide the values into intervals, the frequency histograms of SNR and speaking rate values are drawn, as shown in Figure 2 and Figure 3. It can be seen from the histograms that both SNR and speaking rate are approximately normally distributed, so the values are divided into 3 intervals: $(-\infty, \mu-\sigma]$, $(\mu-\sigma, \mu+\sigma]$, $(\mu+\sigma, +\infty)$, where μ and σ are mean and standard deviation of the normal distribution. This method divided the middle part of a normal distribution from the edge parts, keeping

similar data under one level. Finally, the levels of the three factors are decided as shown in Table 2.

Once factors and levels have been set, groups of data are determined under orthogonal design. Since there are three factors, one with 2 levels and the other two with 3 levels each, the orthogonal table $L_9(2^1 \times 3^3)$ shown in table 1 is chosen, with the column of factor4 not being used. So speech data were collected to make sure that there are enough data (more than 10 utterances) in each group. To make a more general test set, speech data not belonging to any group in the orthogonal table are also collected for diversity.

After recognition results of all systems are produced, these texts are compared to the reference to calculate CERs in each group. For each system, as described in 2.2, a series of performance measurement data are generated for each group by randomly selecting subsets. In our experiments, N = 6, which means six CERs are calculated for each group. To avoid extreme occasions, subsets containing less than 5 utterances are not used.

Finally, for each system, statistical analyses described in 2.3 are performed. The ANOVA procedure is performed using the statistical software SAS, and the procedure similar to range analysis is done manually. During the ANOVA procedure, while the significance of the influence of a factor is decided, means of different levels of a factor are also compared to each other using the Student-Newman-Keuls method [12] to determine whether the two levels are significantly different.

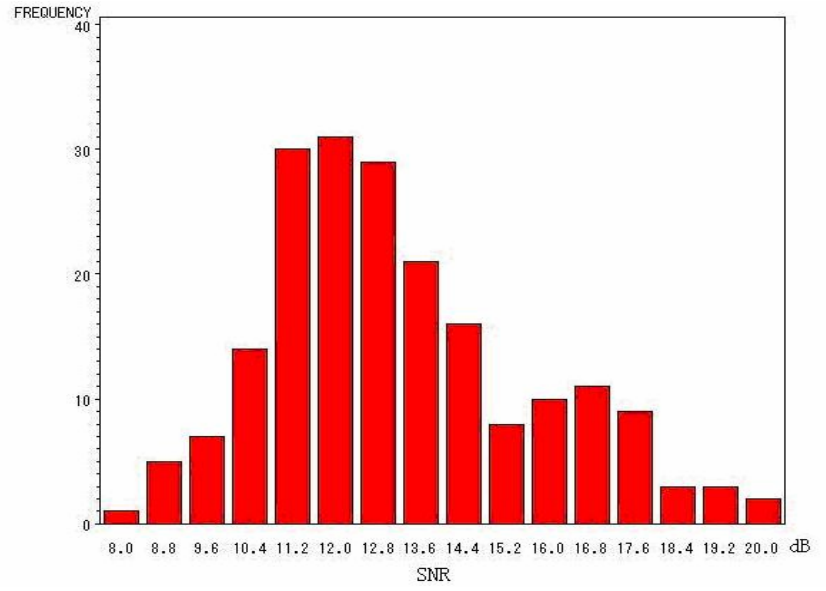

Fig.2. Frequency histogram of SNRs of the LVCSR test data.

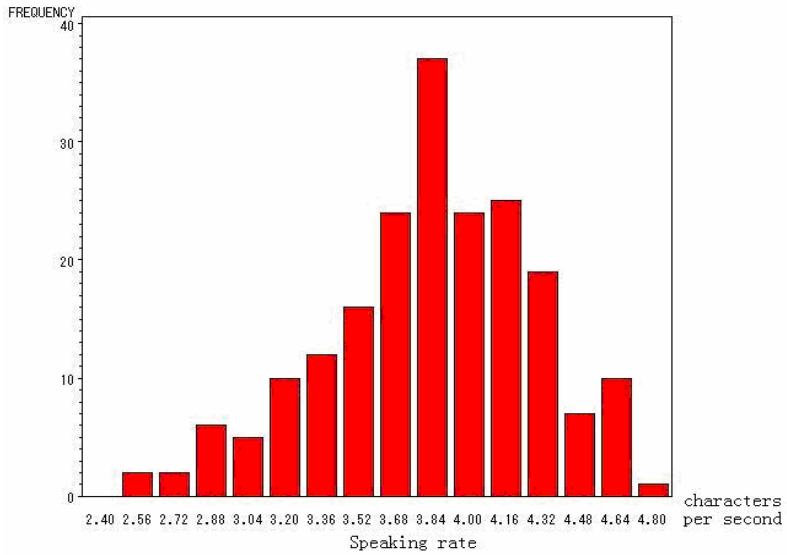

Fig. 3. Frequency histogram of speaking rates of the LVCSR test data.

Table 2. Levels of the three factors

Level	dialectal accent	SNR(dB)	speaking rate (characters per second)
1	without-accent	(-∞, 11]	(0, 3.40]
2	with-accent	(11, 14]	(3.40, 4.35]
3	-----	(14, +∞)	(4.35, +∞)

3.2 Results of PIFA for the 2004 HTRDP ASR Evaluation

As mentioned in 2.2 and 2.3, the feasibility of ANOVA depends on whether the performance measurement data generated for each group are normally distributed and whether the variances of data in different groups are identical. Therefore, before conducting the ANOVA analyses, all the performance measurement data are tested for normality and homogeneity of variances using the Shapiro-Wilk test and the Levene's test respectively. The results are shown in Table 3, where sys1, sys2, sys3 denoted the three participant systems. It can be seen from the table that most test probabilities are greater than 0.05, and that the few numbers that are less than 0.05 are all greater than 0.01. This means the data can be considered as satisfying the assumption of normal distribution and homogeneity of variances.

Table 3. Results of tests for normal distribution and homogeneity of variances

Level (s)			normality (Pr < W)			Homogeneity of variance (Pr>F)		
Dialectal accent	SNR	Speaking rate	Sys1	Sys2	Sys3	Sys1	Sys2	Sys3
1	1	1	0.5217	0.7188	0.2490			
1	2	2	0.8517	0.0520	0.5084			
1	3	3	0.5787	0.8160	0.5442			
1	1	2	0.1127	*0.0418*	0.1324			
1	2	3	0.7900	0.1250	0.2002	0.0998	0.1024	0.1670
1	3	1	0.3407	0.8346	0.5568			
2	1	3	0.3820	0.7235	*0.0227*			
2	2	1	0.6357	0.0522	0.1162			
2	3	2	0.4293	0.4299	*0.0355*			

Results of the analyses detailed in 3.1 are shown in Table 4. In the table, for the "comparison of levels" column, levels with the same letter are not significantly different. For example, for system 2, for the speaking rate factor, the letters for level 1 is "AB", which means it is not significantly different from neither level 2 (labeled with “A”) nor level 3 (labeled with B).

According to results of Table 4, the following conclusions can be made.

1. The factor dialectal accent has significant influence on sys2, but don't have significant influence on sys1 and sys3. So it might imply that sys1 and sys3 are more robust to dialectal accent than sys2.
2. The factor SNR has significant influence on all three systems, and the range proportions are similar, which implies the degrees of influences of SNR are similar for the three systems. And it can also be inferred from the results of comparison of levels that difference between level 1 ((-∞, 11] dB) and level 2 ((11, 14] dB) are not very significant while these two levels are highly different from level 3 ((14, +∞) dB). That means the systems perform much better on data with higher SNR (greater than 14 dB), and when SNR falls below 14dB, it doesn't make much difference whether the SNR is higher than 11dB.
3. The factor speaking rate has significant influence on all three systems, but for sys3, the range proportion is higher than the other two systems, means it has stronger influence on sys3 than others. It can be inferred from the comparison of levels that higher speaking rate brings significantly higher performance. The reason may lie in that when speaking rate is high, noise between words is less than speech with lower speaking rate.
4. For all systems, SNR has the largest range of means ratio, which means SNR is the most important performance influencing factor.

It can be seen that the conclusions are in accordance with common sense in the speech recognition field. And participants can learn from these conclusions whether their systems performs better or poorer than others for the performance influencing factors, thus boosting understanding and communication among researchers.

Table 4. Results of PIFA for the 2004 HTRDP ASR Evaluation

System	Factor	Significance	Level	Mean	Comparison of levels	Range of means	Range proportion
Sys1	Dialectal accent	N	1	0.33222	A	0.03243	0.085
			2	0.36465	A		
	SNR	Y	1	0.42037	A	0.17446	0.458
			2	0.36281	A		
			3	0.24591	B		
	Speaking rate	Y	1	0.37708	A	0.09691	0.254
			2	0.37184	A		
			3	0.28017	B		
Sys2	Dialectal accent	Y	1	0.33766	A	0.06714	0.218
			2	0.40480	B		
	SNR	Y	1	0.44418	A	0.16994	0.552
			2	0.36170	B		
			3	0.27424	C		
	Speaking rate	Y	1	0.36709	AB	0.08869	0.288
			2	0.40086	A		
			3	0.31217	B		
Sys3	Dialectal accent	N	1	0.26047	A	0.03223	0.113
			2	0.22824	B		
	SNR	Y	1	0.29819	A	0.13668	0.478
			2	0.28947	A		
			3	0.16151	B		
	Speaking rate	Y	1	0.29569	A	0.11017	0.385
			2	0.26798	A		
			3	0.18552	B		

4 Conclusions

In this paper, we introduced the PIFA framework for evaluation of speech recognition systems. Within the PIFA framework, the test corpus is designed under orthogonal design, resulting in several groups of test data. For each group, performance measurement data satisfying the two basic assumption of ANOVA are generated and analyzed using ANOVA and range analysis. For each system, the analysis results of PIFA shows whether and how much the system performance is influenced by a certain data property, which is eagerly wanted by participants but not provided by current evaluations. Through the PIFA framework, DOE methods and statistical analyses such as ANOVA and range analysis are introduced to evaluation, make analyses more efficient and reliable.

The PIFA method has been applied into the 2004 HTRDP ASR Evaluation, providing conclusions of robustness to dialectal accent, SNR, and speaking rate for each system. Participants discussed virtues and deficiencies of each system on basis of analysis results from PIFA and improved system performances by adopting or developing better techniques on the poor part of their systems. Feedbacks say that

great progress has been made and it is also demonstrated by the 2005 HTRDP ASR Evaluation results.

So far, PIFA was only applied to the LVCSR task, so future work includes using this method in other tasks such as keyword spotting (KWS) in speech recognition evaluation. In fact, this approach can also be easily applied to evaluations of other pattern recognition or machine learning techniques to explore performance influencing factors.

References

1. David S. Pallett, Jonathan G. Fiscus, and William M. Fisher, et al.: 1994 Benchmark Tests for the ARPA Spoken Language Program. Proceedings of the Human Language Technology Workshop (pp. 5-36). San Francisco: Morgan Kaufmann Publishers, Inc.
2. http://www.nist.gov/speech/tests/sigtests/sigtests.htm
3. L. Gillick and S. Cox: Some Statistical Issues in the Comparison of Speech Recognition Algorithms. ICASSP (1989), 532-535
4. D. Pallett, J. Fiscus, and J. Garofolo: Resource Management Corpus: September 1992 Test Set Benchmark Test Results. Proceedings of ARPA Microelectronics Technology Office Continuous Speech Recognition Workshop, Stanford, CA, September 21-22, 1992
5. http://www.nist.gov/speech/tests/sigtests/wilcoxon.htm
6. N. Deshmukh, R. Duncan, A. Ganapathiraju and J. Picone, Benchmarking Human Performance for Continuous Speech Recognition, Proceedings of the Fourth International Conference on Spoken Language Processing, pp. SuP1P1.10, Philadelphia, Pennsylvania, USA, October 1996.
7. F. Kubala, A. Anastasakos, J. Makhoul, L. Nguyen, R. Schwartz, and G. Zavaliagkos: Comparative Experiments on Large Vocabulary Speech Recognition, ICASSP 1994, Adelaide, Australia.
8. http://www.863data.org.cn/english/2004syllabus_en.php
9. http://www.863data.org.cn/english/2005sttinfo_en.php
10. Luquan Ren: Optimum Design and Analysis of Experiments (Second Edition), Higher Education Press, Beijing, 2003.
11. http://www.statsoft.com/textbook/stanman.html
12. Qijun Shen: SAS Statistical Analysis, Higher Education Press, Beijing, 2005.