Analysis of the Effect of Training and Test Data on the Performance of Speech Recognition Systems

Xiangdong Wang[1, 2], Feng Xie[1, 2], Shouxun Lin[1], Yueliang Qian[1], Qun Liu[1]
[1] *Key Laboratory of Intelligent Information Processing, Institute of Computing Technology, Chinese Academy of Sciences, Beijing 100080, China*
[2] *Graduate University of Chinese Academy of Sciences, Beijing 100085, China*
{xdwang, xiefeng, sxlin, ylqian, liuqun}@ict.ac.cn

Abstract

In the field of speech recognition, performance varies much when the system is trained or tested with different data. In this paper, we explore the effect of training and test data on the performance of automatic speech recognition systems. Unlike other researchers who analyze the effect of training and testing as pattern learning and recognition of vectors, the effect of data is investigated as effect of data properties, such as SNR and kind of environmental noise. For a data property, a statistical model based on ANOVA was proposed to decompose the effect on system performance into three parts----effect of training data, test data and their interaction, and each part is considered dependent on the level of data properties. Experiments were conducted on a LVCSR system for the data properties of kind of noise and SNR, and results and analysis are presented to explain how they influence the performance by training and test.

Keywords: LVCSR, speech recognition, training data, test data, ANOVA.

1. Introduction

In the field of speech recognition and other pattern recognition domains, it is well known that training and test data have strong influence on performance in addition to algorithms adopted in the system. Obviously, understanding the way training and test data affect performance can contribute to the improvement of systems. Kubala et al [1] conducted a series of experiments to investigate influence of speaker number, data amount and domain of training data on the performance of the BYBLOS LVCSR system, and conclusions are drawn by simply comparing the WERs. Marloof et al [2-4] compared the use of ANOVA and LABMRMC models in the analysis of finite-sample effects in training and testing of competing classifiers, conducted experiments by Monte Carlo simulation, and came to conclusions about the relation between variance of performance and number of samples in training and test data.

1-4244-0971-3/07/$25.00 ©2007 IEEE.

In this paper, we propose a general statistical model based on ANOVA to explore the effect of training and test data on the performance of automatic speech recognition systems. Unlike other researchers who consider training or test data as a whole, the effect of data is investigated in terms of effect of data properties, such as kind of noise and signal noise ratio (SNR). The effect of a data property is decomposed into three parts----effect of training, testing and their interaction, and each part is considered dependent on the levels of the data property. Experiments were conducted on a Chinese large vocabulary continuous speech recognition (LVCSR) system for the data properties of kind of noise and SNR, and results and analysis are presented to illustrate how they influence the performance through training and testing.

The rest of the paper is organized as follows. In Section 2, the main idea of the proposed analysis method is presented. In Section 3, we describe the experiments conducted on a Chinese LVCSR system to investigate effect of kind of noise and SNR. Experimental results and analysis are given in Section 4. Finally, conclusions are drawn in Section 5.

2. Main idea of the analysis

2.1 Data properties

In this paper, data properties are referred to as features or characteristics of data, which are decided by the speakers, recording conditions, or the text material, e. g. the gender of speaker, the signal noise ratio (SNR) or the kind of environmental noise. Values or classes of a data property are called levels of the data property. For example, the data property "speaker gender" has two levels----"male" and "female", and the levels of SNR are real numbers such as "5dB", "10dB" and "14.5dB".

It is well known that some data properties can influence the performance considerably. One system may achieve quite different performances when tested by data with different levels of data properties such as dialectal accent, SNR or speaking rate [5]. It is also demonstrated that for given test data, system trained with different levels of data properties (e.g. SNR or

kind of environmental noise) can obtain different recognition results [6]. Therefore, the levels of a data property can have effect on system performance through the process of both training and testing.

Generally, there are two types of effects related to training and test data: the effect of data amount and the effect of data content, and this paper focuses on the latter. Most researchers treat the data as a whole set and explore the effect as that of patterns or vectors. But actually, the effect of data content is very complicated and can be studied by investigating sub-effects of data properties, which is the main idea of the analysis approach proposed in this paper. For example, instead of investigating effect of data with various speakers and recording conditions, we can first explore influence of speaker genders, dialectal accent, SNR, kind of noise, etc. and further study their combination.

2.2 The analysis model

As discussed in 2.1, the levels of a data property can have effect on system performance through the process of both training and testing. For example, for a given system, different performances can be achieved on test data with different SNRs. And systems trained by data with different SNRs perform distinctly on the same test data. It is also noticed that when the levels of training and test data are the same, the performance can be much better than otherwise. In order to interpret these phenomena, the ANOVA model in statistics is introduced to depict the relation between the performance and the levels of a given data property in training and test data.

ANOVA [7] is abbreviation for analysis of variance, which is a powerful method of hypothesis test. The analysis of the effect of a data property can be viewed as a 2-way ANOVA, where training and testing are viewed as two factors. Considering a series of experiments with training and test data with different levels, the statistical model of ANOVA can be written as

$$X_{ijm} = \mu + \tau_i + \beta_j + (\tau\beta)_{ij} + \varepsilon_{ijm} \qquad (1)$$

where i and j denote the levels of a certain data property, X_{ijm} denotes the performance metric of the m^{th} experiment with training data with level i and test data with level j, τ_i denotes the effect of training with the i^{th} level, β_j denotes the effect of testing with the j^{th} level, $(\tau\beta)_{ij}$ denotes the effect of interaction between the two factors, and ε_{ijm} denotes random experimental error [7].

The purpose of ANOVA is to test whether the effects are statistically significant, for example, for training, decision is made on whether to accept or reject the following hypothesis.

$$\tau_1 = \tau_2 = \ldots = \tau_K = 0 \qquad (2)$$

With this statistical model, the effects of a data property in training, testing and their interaction are investigated respectively. Compared to other research work, this analysis method can give clearer and more detailed results of whether and how a data property influences the performance.

3. Experiments

3.1 Overview

To explore the effect of data properties on system performance through the process of training and testing and to justify the adoption of the model proposed in 2.2, a series of experiments are conducted on a Chinese LVCSR system. And there are two data properties under investigation: kind of noise and SNR, since it is well accepted that these two data properties in test data influence performance considerably but little work has been reported considering their effect of both training and testing and their interaction.

The main idea of the experiments is to train and test the system using data with different background noises and SNRs, and analyze the recognition result according to the model in Equation 1. For each data property, there are two kinds of experiments: experiments for single level and experiments for multiple levels. The former means that there is only one level of the data property in the training or test data, while in the latter case, effect of multiple levels are investigated.

The LVCSR system for experiment is constructed using the HTK toolkit developed by Cambridge University [8]. MFCC features and tri-phone HMMs are used for acoustic model, and bi-gram is adopted as linguistic model. Since the purpose is to investigate the effect of noise, no noise robust techniques such as speech enhancement and feature compensation are incorporated in the system.

A training set and a test set are chosen and noises are added to each utterance with specific SNR. Table 1 gives the details of the original training and test data, which are both reading speech without background noise, stored in 16 KHz, 16 bit PCM WAV format. To facilitate the procedure of ANOVA, the test set is further divided into 20 subsets, performances on which are used as results of repeated experiments in ANOVA.

Table 1. Details of the original data sets

	Number of speakers		Total duration	Number of utterances
	Male	Female		
Training set	100	100	About 100 hours	71639
Test set	10	10	1 hour	1200

Performance is assessed by the metric of CER (character error rate), which is the same as the widely

used WER except that the basic unit is Chinese character instead of word.

3.2 Experiments for kind of noise

These experiments aim to explore the effect of different kind of environmental noises on recognition performance. Four kinds of environmental noises are used as 4 levels of the data property: noises collected in a factory, a restaurant, a running taxi, and white noise, which are referred to as "factory", "restaurant", "taxi" and "white" and denoted by "F", "R", "T", and "W" in the rest of the paper. In all experiments concerning kind of noise, noises are linearly added to each utterance with the SNR of 10dB.

In the experiments for single level, the system is trained and tested using data set with each kind of noise, thus yielding 4×4=16 recognition results. For example, there is one result obtained by training the system using data with "factory" noise and testing the system using data with "restaurant" noise. Then ANOVA is applied to analyze the effect of the data property, treating training and testing as two factors with account of their interaction. The ANOVA procedure is performed using the statistical software SAS [7].

In the experiments for multiple levels, the original training set is divided into two parts, on which different kinds of noise are added respectively. To study the influence of proportion of levels, ratios of data amount of the two parts varies for different times. Total 9 experiments of 3 groups are carried on according to the design shown in Table 2, where proportions are calculated in terms of total duration of speech.

Correspondingly, the original test set is equally divided into 4 subsets, and the four noises are added to each subset. Performance of systems trained with the 9 train sets and tested with the 4 test sets are calculated for further analysis.

Table 2. Design of experiments for multiple levels of kind of noise

Group	1		2		3	
Noise	R	F	R	T	R	W
Proportion	0.7	0.3	0.7	0.3	0.7	0.3
	0.5	0.5	0.5	0.5	0.5	0.5
	0.3	0.7	0.3	0.7	0.3	0.7

3.3 Experiments for SNR

Experiments for SNR are quite similar to those for kind of noise, except that the noise added to speech data is invariable and SNR varies in different experiments. There are also 4 levels used: 5dB, 10dB, 15dB, and 20dB. In all the experiments concerning SNR, only one piece of noise is used which is collected in the street.

In the experiments for single level, system is trained and tested using data sets with different levels of SNR, and the results are analyzed using ANOVA. The ANOVA procedure is performed using the statistical software SAS.

In the experiments for multiple levels, the design is slightly different from that of kind of noise. Since the levels of SNR can be continuous real numbers, a distribution simulation technique is used when adding multiple levels of SNR to speech data. For a data set, three ways of noise adding is used: (1) Noise is added to the speech at SNR of 10dB for all utterances; (2) Noise is added under a normal distribution with mean of 15dB and variance of 4; (3) Noise is added under a uniform distribution from 5dB to 20 dB. Noises are added to both the original training and test set through all three ways, yielding 3×3=9 results.

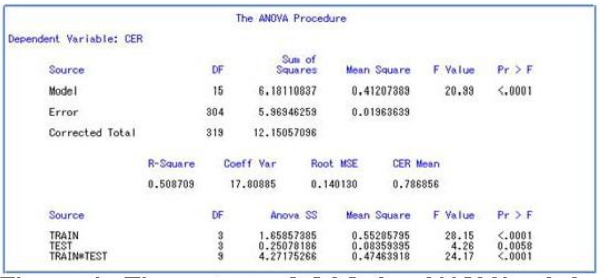
4. Experimental results and analysis

4.1. Results of experiments for kind of noise

For the experiments for single level of kind of noise, results of CERs are shown in Table 3, and the output of SAS for the ANOVA procedure in given in Fig.1.

Table 3. CERs of the experiments for single level of kind of noise

Training / Test	F	R	T	W
F	0.6700	0.7895	0.8828	0.7630
R	0.7767	0.5875	0.7260	0.8493
T	0.8435	0.6501	0.4916	0.9014
W	0.8208	1.0043	0.9696	0.7548



The ANOVA Procedure

Dependent Variable: CER

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	15	6.18110837	0.41207389	20.98	<.0001
Error	304	5.96946259	0.01963639		
Corrected Total	319	12.15057096			

R-Square	Coeff Var	Root MSE	CER Mean
0.508709	17.80885	0.140130	0.786856

Source	DF	Anova SS	Mean Square	F Value	Pr > F
TRAIN	3	1.65857385	0.55285795	28.15	<.0001
TEST	3	0.25078186	0.08359395	4.26	0.0058
TRAIN*TEST	9	4.27175266	0.47463918	24.17	<.0001

Figure 1. The output of SAS for ANOVA of the experiments for single level of kind of noise.

As mentioned in 3.1, the test set is divided to 20 subsets so that there are 20 observations for each combination of training and testing levels in ANOVA. According to theory of statistics, the feasibility of ANOVA depends on whether the CER values for each combination are normally distributed and whether the variances of CERs for all combinations are identical. Therefore, before conducting the ANOVA analyses, all the CER values are tested for normality and homogeneity of variances using the Shapiro-Wilk test and the Levene's test [7] respectively, and the results show that the conditions are both satisfied.

The result of ANOVA shown in Fig.1 indicates that the effect of training, testing and their interaction are all significant (the values of Pr>F are all less than 0.05).

And the F values and Pr values also imply that the effects of training and interaction are stronger than that of testing.

In Table 3, it is noticed that the diagonal CERs are quite less than those in the same row or column, which implies that the interaction effect is the strongest effect when levels of training and test data are the same.

So it can be concluded that kind of noise significantly influence the system performance through both training and testing, for example, systems trained by data with different noises may performance distinctly for most data whatever noise the test data are with. Moreover, the effect of interaction is very strong, which leads to that best performances are achieved when training and test data are with the same kind of noise.

For the experiments for multiple levels of kind of noise, results of CERs are shown in Table 4, where R, F, T, and W denote the four levels "restaurant", "factory", "taxi" and "white", "R0.3+T0.7" refers to the training set in which the proportions of data with noise of "restaurant" and "taxi" are 0.3 and 0.7 respectively, and the rest may be deduced by analogy.

From Table 4, it can be seen that CERs are much lower when the level of test data is also included in the training data than otherwise, which shows the effect of interaction in cases of multiple levels. The effect of training and testing is also obvious; for example, test data with "white" noise obtain better performance on system trained by data with noise "R+F" than "R+T".

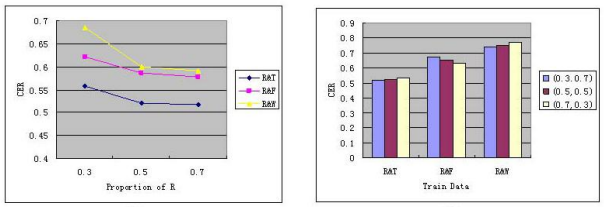
Figure 2(a) shows the relation between the proportions of data with "restaurant" noise in training data and the performances on test data with the "restaurant" noise. The figure indicates that when proportion of a level increases in training data, the performance on test data with the same level increases, which is another illustration of the interaction effect.

In Fig.2(b), the performances are calculated on the two data sets with the same kinds of noise as those in training data, for example, for the "R&F" case, the training set are with levels "R" and "F", so overall performance is calculated on the "restaurant" and "factory" test subsets, in which the proportions of both levels are 0.5. The graph in Fig.2(b) indicates that the overall performance varies little while the proportions of levels changes in training set. The reason may be that while performance on test data with one level goes up due to proportion increase of that level in training data, performance on test data with the other level goes down due to corresponding proportion decrease in training set, which is also shown in Table 4 clearly.

So it can be concluded that the effects of training, testing and their interaction still exist in cases of multiple levels. When proportion of a level increases in training data, the performance on test data with that level will increase, too. But for cases when both training and test data involves multiple levels, there is trade-off on performance and it is hard to predict the result.

Table 4. CERs of the experiments for multiple levels of kind of noise

Training data	Test data			
	F	R	T	W
R0.3+T0.7	0.7965	0.5576	0.4840	0.8452
R0.5+T0.5	0.7823	0.5206	0.5266	0.8616
R0.7+T0.3	0.7752	0.5177	0.5555	0.8698
R0.3+F0.7	0.6899	0.6216	0.7884	0.7753
R0.5+F0.5	0.7240	0.5860	0.7427	0.7876
R0.7+F0.3	0.7297	0.5773	0.7366	0.8246
R0.3+W0.7	0.7482	0.6856	0.8203	0.7342
R0.5+W0.5	0.7354	0.5988	0.7427	0.7712
R0.7+W0.3	0.7496	0.5917	0.7488	0.7931



(a) (b)

Figure 2. Relations between the performance and the proportion of levels in train and test data: (a) Performance on test data with only one level; (b) Overall performance on test data with two levels.

4.2 Results of experiments for SNR

For the experiments for single level of SNR, results of CERs are given in Table 5, and the output of SAS for the ANOVA procedure is shown in Fig.3.

Similar to the experiments for kind of noise, the test set is divided to 20 subsets so that there are 20 observations for each combination of training and test levels in ANOVA. Shapiro-Wilk test and the Levene's test are also conducted to guarantee the feasibility of ANOVA.

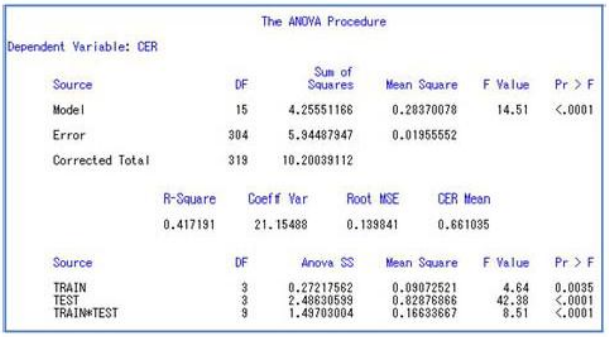
The result of ANOVA shown in Fig.3 indicates that the effect of training, testing and their interaction are all significant (the values of Pr>F are all less than 0.05). And the F values and Pr values also imply that the effects of testing are stronger than that of training and interaction.

In Table 5, for each row, the performance of the diagonal cell is better than other cells, which means given test data with a certain SNR, the best performance is achieved when the training data is with that SNR too. But that doesn't hold true for columns. For system trained by data with a certain SNR, the best performance is obtained when SNRs are high (15dB and 20dB). This can be explained by what ANOVA

indicates: the effect of interaction is stronger than effect of training, while the effect of testing is the strongest.

Table 5. CERs of experiments for Single level of SNR

Training / Test	5db	10db	15db	20db
5db	0.7287	0.7505	0.8321	0.9426
10db	0.6629	0.6244	0.6415	0.6941
15db	0.6729	0.5921	0.5654	0.5676
20db	0.7255	0.6053	0.5485	0.5367



The ANOVA Procedure

Dependent Variable: CER

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	15	4.25551166	0.28370078	14.51	<.0001
Error	304	5.94487947	0.01955552		
Corrected Total	319	10.20039112			

R-Square	Coeff Var	Root MSE	CER Mean
0.417191	21.15488	0.139841	0.661035

Source	DF	Anova SS	Mean Square	F Value	Pr > F
TRAIN	3	0.27217562	0.09072521	4.64	0.0035
TEST	3	2.48630599	0.82876866	42.38	<.0001
TRAIN*TEST	9	1.49703004	0.16633667	8.51	<.0001

Figure 3. The output of SAS for ANOVA of the experiments for single level of SNR.

For the experiments for multiple levels of SNR, results of CERs are given in Table 6. It doesn't hold true that best performances are gained when the distributions are same in the training and test set. There may be two reasons: (1) as revealed above, the effect of interaction is weaker than effect of testing; (2) there are trades-off on performance similar to that in the experiments for kind of noise. The mechanism is complicated and is to be studied in future work.

Table 6. CERs of experiments for multiple levels of SNR

Training / Test	10dB	Normal	Uniform
dB	0.5887	0.5630	0.5974
Normal	0.6017	0.5372	0.5701
Uniform	0.5697	0.5262	0.5464

5. Conclusions

In this paper, a statistical model based on ANOVA is proposed to explore the effect of training and test data on the performance of automatic speech recognition systems. The effect of data is investigated in terms of effect of data properties, and is decomposed into three parts----effect of training, testing and their interaction.

Effects of the data properties of kind of noise and SNR on a LVCSR system are investigated according to the model. A series of experiments were conducted and results indicated that both kind of noise and SNR can influence system performance through training, testing and their interaction in both cases of single level and multiple levels. For multiple levels, the three effects interact with each other while influencing the performance and more effort is still needed to explore the mechanism.

The effort started in paper to investigate effect of data properties under the proposed model aims to reveal how training and test data influence performance of speech recognition systems in a novel way. And the experiments, results, and analysis for the two data properties of kind of noise and SNR can be useful to researchers, for training data can be designed to obtain better performance according to the effects disclosed.

References

[1] Kubala, F., Anastasakos, A., Makhoul, J., et al., "Comparative Experiments on Large Vocabulary Speech Recognition", *Proc. ICASSP 1994*, vol.1, pp. 561-564.

[2] Maloof, M. A., "On Machine Learning, ROC Analysis, and Statistical Tests of Significance", *Proceedings of 16th International Conference on Pattern Recognition*, vol.2, pp. 204-207.

[3] Beiden, S. V., Marloof, M. A., and Wagner, R. F, "A General Model for Finite-Sample Effects in Training and Testing of Competing Classifiers", *IEEE Trans. on Pattern Analysis and Machine Intelligence*, vol.25, pp. 1561-1569.

[4] Marloof, M. A., Beiden, S. V., Wagnet, R. F., "Analysis of Competing Classifiers using Components of Variance of ROC Accuracy Measures", *Technical Report CS-02-01, Department of Computer Science, Georgetown University, Washington, DC*, http://www.cs.georgetown.edu/~maloof/pubs/cstr-02-01.pdf.

[5] Xiangdong Wang, Feng Xie, et al., "DOE and ANOVA ba sed Performance Influencing Factor Analysis for Evaluation o f Speech Recognition System", *Proceedings of ISCSLP 2006, (companion volume)*, pp. 431-442.

[6] Pearce, D., and Hirsch, H., "The Aurora Experimental Framework for the Performance Evaluation of Speech Recognition Systems under Noisy Conditions", *Proceedings of 6th International Conference on Spoken Language Processing*, pp.29-32.

[7] Qijun Shen, *SAS Statistical Analysis*, Higher Education Press, Beijing, China, pp. 84-107.

[8] HTK home page, *http://htk.eng.cam.ac.uk/*