Study on Method of Feature Selection in Speech Content Classification

Si An
Institute of Computer Science and Technology
Chongqing University of Posts and Telecommunications
Chongqing, China

Xinghua Fan
Institute of Computer Science and Technology
Chongqing University of Posts and Telecommunications
Chongqing, China

Abstract—**Information communication is developing rapidly now, Voice communication from a distance is more and more popular. In order to evaluate and classify the content correctly, the acoustic features is used to analyze first in this paper, Orthogonal experiment[1] method is used to find out characteristic of voice that has contribution to the speech content classification then make it and the textual characteristic together. The result of experiments shows that the feature combination of voice and content has better effect on voice content classification, the effectiveness has been improved.**

Keywords—acoustic features; orthogonal experiment; the SVM classifier; CHI statistical methods; features level fusion; LBS vector quantization algorithm

I. Introduction

As an ideal human-machine communication voice has characteristics of natural, convenient and fast. It has been the pursuit of ideals that making machine understands human speech. In the information age, internet and the telephone exchange increasingly popular, the requirements of the machine is not just to be able to understand the human speech, it also can simultaneously make the appropriate judgment of speech content. For example, to make a fair and impartial judgment of artificial service or sales calls service. Consider this, we began to analysis the voice content, and thus we propose the research of acoustic features. The acoustic features is different, which is caused by two factors, one is the physical structure of the channel itself; the other one is the different vocal habits of everyone, it can also cause the difference of acoustics features that using a different way of vocal organs. So, the voice signals are results of vocal channel structure, pronunciation habits, content of speech and environmental effects comprehensive. It determined by a variety of factors but mainly by semantic [2] content. There are many speech feature parameters existing, but there is no one related only to the voice content or the speaker, We have to choose appropriate parameters to processing or analysis, excluding other influential factors interference caused, and highlight the feature in voice signals which can expression the content to further identify the features that is discriminative to the content category. What we studied is the mostly commonly used parameters of Mel-Frequency Cepstrul Coefficients and time-domain energy and their difference combination.

We can make a judgment and distinguish directly based on text information, but for voice messages, what we first should do is to do speech recognition, in this process, we have a few things to do, such as pre-process, feature selection, the structure of acoustics model and language model. After these procedure, the received content is different from the initial meanings which will cause the miss of the result of classification. So in order to get precisely analysis and estimate, it’s necessary to composite the two kind of information. That is data fusion. We can train the classifier by using the feature vectors that combining the acoustics features and text feature. But because these two features come from different time domain and they have distinct criterion, the dimension of the space will increase if merging them simply. So we’d better find the optimum combination of features to improve the robustness of the system.

II. The Feature Analysis of Speech

A. Acoustic feature extraction

At present, the researchers find the characteristics that are closely related to the pronunciation are mainly pitch frequency, short-time energy. The parameter of formant and the spectral energy distribution is related to the sound way. In this paper, What we studied is the commonly used parameters of Mel-Frequency Cepstrum Coefficients and time-domain energy and their difference combination. Because the first-order differential cepstrum parameters reflected the changes over time that its dynamic characteristics, we think it can complete more express the original speech.

1) The optimization of the acoustic features

When analyzing the content of speech, it has to have strong interference ability of environmental noise and robustness. It can’t meet the requirements of robustness if sticking with single parameters. Selecting the number M from the given number of N feature parameters X(1),X(2),...X(N) to training the classifier, it’s called feature selection. There are some ways of feature selection in other field of research, in paper [3], there is a method named multi-objective optimization. If only these features are put together freely, the dimension of features will be quite high which not improve the performance of the system but extend the training time thus affect the real-time performance. It is not convenient to use. How to get the information that has the characteristics of the complementary role from the large number of feature parameters, it’s a problem with practical significance. In the following, we analyze and optimize the characteristic parameters by orthogonal experiment.

2) Orthogonal experimental design

Orthogonal experimental design is widely applied in agriculture, process design in the developed country. There are many examples of successful application in our country[4,5].The method of orthogonal experimental design has made good use of the table—"Orthogonal" to arrange experiment.

It can elected strong small number of experimental conditions in many experiments and inference to find the best process conditions through these number of experimental conditions[6].The factors called factor which can affect the result of the experiment in this way, the state of different factor called level. Orthogonal experiment is to find the optimum combination of the factors exists. In the process of searching the required test times is fewer than in the exhaustive method. For instance, in the experiment of this paper, there need two to the power of twenty six in exhaustive method, but it just need thirty two in orthogonal experiment.

3) Orthogonal experiment steps

a) Constructing orthogonal table: Orthogonal table is usually need solid mathematical theory, but when the factor level is two, the table is very easy to construct, reference the literature [7].

*b) Factors:*In the orthogonal table, the amount of each level is equal to the average and between any two columns of different levels of the total number of combinations is equal to the average, so, when arranging orthogonal experiments, all sorts of factors collocation is balanced. In the table, every row said an experiment scheme, which is a combination of various factors in state; each column figures show that the corresponding factors of the state.

c) The experiment results analysis:Analysis of variance is that can distinguish the difference between experimental results and error caused by the fluctuation of differences between the experimental results, this method of math make up the deficiency of the poor analysis method in this respect, so, in this paper, we using the analysis of variance to experimental results. According to the theory of difference analysis, we can get the discriminative that the change of level caused by the difference between the experimental results. If the experimental results changed caused by the changes of factor levels within the error range or has little difference with the error, the change of this factor level can determined not cause a significant change in results; On the other hand, if factor levels' change will not cause changes than error range in the experimental results, we can sure that the factor has a significant impact on the experimental results. The purpose of this analysis is to find out the things which have a significant impact factor through the data.

*d) The selection of orthogonal table and the structure:*For 2 levels orthogonal table, Hadamard horse matrix can be used to construct: let the second-order matrix

Hardamad is the basic matrix, the rest can be done in the same manner, in this paper we use the table L_{26} (2^{26}) .Remove the first column which is full one, for simplicity, remove the back of the five columns and turn -1 to 0, the result was a matrix of 32*26.

B. Speech content textual feature extraction

In addition to the acoustic features, the textual feature should also be considered. There is some correlation between the audio data and content data, the incomplete of the audio can be added in some ways for example the text information. So the most direct way to evaluate the speech is do classification by using the text after recognizing.

SVM is a method of sample learning that has a solid theoretical foundation. It implements an efficient "transduction reasoning" from training sample to predict samples and simplifying the classification. With the support vector machine classification's better overall performance, it is used in this paper. For content, the word frequency is the characteristic of the text.

1) The method of textual feature selection

There are two factors can be observed in the text in fact, that is word frequency and document frequency, there are some feature selection algorithm based on the document frequency such as CHI statistics, Information Gain(IG), Mutual Information(MI). Many experiments show that the CHI statistics is more commonly used method. Its basic idea is to determine the theory correct by observing the deviation of the actual value and the theoretical value or not. It is a measure of the relevance between feature word t and document category c_j, assuming that meet the distribution of the first order χ^2 between t and c_j. The bigger of the value of chi-square statistic the key words belong to a category the greater the relevance between the key words and the category. The Chi-square statistic calculated which the key words t to the category c_j is defined as :

$$\chi^2(t, c_j) = \frac{N \times (AD - BC)}{(A + C) \times (B + D) \times (A + B) \times (C + D)} \quad (2)$$

In the formula (2), A stands for the number of documents which belong to class c_j and contains the key word t; B stands for the number of documents which not belong to class c_j but contains the key word t; C stands for the number of documents which belong to class c_j but not contains the key word t; D stands for the number of documents which neither belong to class c_j nor contains the key word t; N represents the total number of total text in training corpus. In this paper, we simply believe that the characteristic words are the words which has high CHI value.

III. The Structure of The Experiment

Audio content classification [7] is that to train a classifier by using the extracted features data. In order to train the classifier, we should build a data set used in the experiments, which includes voice files and the text files after recognizing. SVM is used to training the data set to get the classifier, the average value of many experiment is taken as the final result. The experiment is divided into four parts in this paper: the first part is training the audio features individually; the second part is doing orthogonal experiment to get the optimized combination; the third part is training the classifier by using the text features (we can considered it as ideal that the text file is not the result of speech recognizing); The forth is training the classifier by the fusion features combined the acoustic features with textual features.

The overall structure of experiment in this paper is showed in figure1.

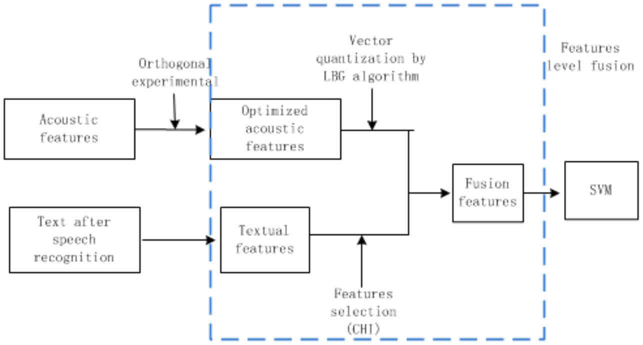

Fig. 1. The overall structure in this paper

A. *Design of experiment and result analysis*

1) Corpus sources : The source of experiment data in this experiment is: the real situations of dialogue from two men and three women who are articulate simulate the real scene. There are 2179 dialogues, the content of dialogue are two parts for transportation and legal launched two topics. There are 1045 speech data about traffic class, 1134 speech data about legal. The sampling frequency is 16 KHz, The quantitative accuracy is 16bit, and the voice of the frame length is 256 sampling points. In these experiments, because the signal noise ratio of data is low, we have something to do to improve the efficiency of endpoint detection; according the paper [9], when to set the threshold to time we set a new threshold by weighting the maximum and minimum values of the volume and averaged it. More accurate effective interception of voice is got and it provides favorable conditions to feature selection.

2) Experimental evaluation criteria: The following indicator is used to evaluate the performance of the classification results of the experiment,

Precision P that is also named accuaracy*:*

$$p = \frac{A}{B} \times 100\% \tag{3}$$

Recall R:

$$R = \frac{C}{D} \times 100\% \tag{4}$$

F_1 value :

$$F_1 = \frac{2 \times P \times R}{P + R} \times 100\% \tag{5}$$

In the formula (3), A stands for the number of text that determined correctly by classifier, B represents the number of text determined by classifier.

In the formula (4), C stands for the number of text that determined correctly by classifier; D represents the number of text in the test set. The F_1 value is a comprehensive evaluation standard.

B. *Experiment*

1) Using the audio features only to train the classifier: The classifier is trained by the audio features only, the test result is:

TABLE I. ACCURACY OF ORIGINAL ACOUSTIC FEATURES

Features	Dimensions	accuracy%	F_1 %
Acoustic	26	74.9	70.3

2) Optimize the audio feature parameters by orthogonal experiment

*a) Factor selection:*The selection of the characteristic parameters of a total is 26 in this paper, each factor has two levels which 1 stand for used and 0 is unused. Characteristic parameters including the combination of MFCC feature and energy mentioned above and their dynamic first-order difference, a total of 26 dimension:

$$mfcc_1, mfcc_2, mfcc_3, mfcc_4, mfcc_5, mfcc_6, mfcc_7, mfcc_8, mfcc_9,$$

$$mfcc_{10}, mfcc_{11}, mfcc_{12}, En, mfcc_{13}, mfcc_{14}, mfcc_{15}, mfcc_{16}, mfcc_{17},$$

$$mfcc_{18}, mfcc_{19}, mfcc_{20}, mfcc_{21}, mfcc_{22}, mfcc_{23}, mfcc_{24}, \Delta En$$

Because the first-order differential cepstrum parameters reflected the changes over time that its dynamic characteristics, we think the dynamic characteristic parameters can complete more express the original voice.

b) The experiment design

According to the design of orthogonal experimental design method, each column corresponds to characteristic parameters and each row represents a kind of combination plan, the number 1 stands for used and 0 is unused. The last column of the table is each set of features combination experiment by the end of the audio classification effect. The experimental scheme is shown in table 2.

After orthogonal experiment result is analyzed from the table, the P value in the statistical is obtained by look-up table after getting F value, every parameter was coded B1,B2,B3...:

It can be seen through the analysis of significance that the audio content for the classification result is greatly influenced by B3,B5,B6,B8,B11,B14,B17,B18,B25.

That is the parameters:

$$mfcc_3, mfcc_5, mfcc_6, mfcc_8, mfcc_{11}, mfcc_{14},$$

$$mfcc_{17}, mfcc_{18}, mfcc_{25}$$

the parameters has little effect on experimental results are:

$$mfcc_9, mfcc_{13}, mfcc_{21}, mfcc_{22}, mfcc_{23}$$

TABLE II. THE ACCURACY OF DIFFERENT FEATURE COMBINATION

No .	Combined Solutions	accur acy %
1	1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1	74.9
2	0 1 0 1 0 1 0 1 0 1 0 1 0 1 0 1 0 1 0 1 0 1 0 1 0 1	66.5
3	1 0 01 1 0 0 1 1 0 0 1 0 1 0 1 1 0 0 1 1 0 0 1 1 0	70
4	0 0 1 1 0 0 1 1 0 0 1 1 0 1 1 1 0 0 1 1 0 0 1 1 0 0	64
5	1 1 1 0 0 0 0 1 1 1 1 0 0 1 1 0 0 0 0 1 1 1 1 0 0 0	68
6	0 1 0 0 1 0 1 1 0 1 0 0 1 1 0 0 1 0 1 1 0 1 0 0 1 1	69
7	1 0 0 0 0 1 1 1 1 0 0 0 1 1 0 0 0 1 1 1 1 0 0 1 0 1	65
8	0 0 1 0 1 1 0 1 0 0 1 0 0 1 1 0 1 1 0 1 0 0 1 0 1 0	74.5
9	1 1 1 1 1 1 1 0 0 0 0 0 0 1 1 1 1 1 1 0 0 0 0 0 0 0	69
10	0 1 0 1 0 1 0 1 1 0 1 0 1 1 0 1 0 1 0 0 1 0 1 0 1 0	64
11	1 0 0 1 1 0 0 1 0 1 1 0 1 1 0 1 1 0 0 0 0 1 1 0 0 1	68
12	0 0 1 1 0 0 1 0 1 1 0 0 0 1 1 1 0 0 1 0 1 1 0 0 1 1	68.5
13	1 1 1 0 0 0 0 1 0 0 0 1 1 1 1 0 0 0 1 0 0 1 0 1 1 1	73.5
14	0 1 0 0 1 0 1 0 1 0 1 1 0 1 0 0 1 0 1 0 1 0 1 1 0 1	71
15	1 0 0 0 0 1 1 0 0 1 1 1 0 1 0 0 0 1 1 0 0 1 1 1 1 0	72.5
16	0 0 1 0 1 1 0 0 1 1 0 1 1 1 1 0 1 1 0 0 1 1 0 1 0 0	69
17	1 1 1 1 1 1 1 1 1 1 1 1 1 0 0 0 0 0 0 0 0 0 0 0 0 0	73.8
18	0 1 0 1 0 1 0 1 0 1 0 1 1 0 1 0 1 0 1 0 1 0 1 0 1 0	71.5
19	1 0 0 1 1 0 0 1 1 0 0 1 0 0 1 0 0 1 1 0 0 1 1 1 0 1	68.5
20	0 0 1 1 0 0 1 1 0 0 1 1 1 0 0 0 1 1 0 0 1 1 0 0 1 1	69.5
21	1 1 1 0 0 0 0 1 1 1 1 0 0 0 0 1 1 1 1 0 0 0 0 1 1 1	70.5
22	0 1 0 0 1 0 1 1 0 1 0 0 1 0 1 1 0 1 0 0 1 0 1 1 0 1	67.5
23	1 0 0 0 0 1 1 1 1 0 0 0 1 0 1 1 1 0 0 0 0 1 1 1 1 0	73.5
24	0 0 1 0 1 1 0 1 0 1 1 0 1 1 0 1 0 0 1 0 1 1 0 1 0 0	71.5
25	1 1 1 1 1 1 1 0 0 0 0 0 0 0 0 0 0 0 0 0 1 1 1 1 1 1	73.5
26	0 1 0 1 0 1 0 0 1 0 1 0 1 0 1 0 1 0 1 1 0 1 0 1 0 1	69.5
27	1 0 0 1 1 0 0 0 0 1 1 0 1 0 1 0 0 1 1 1 1 0 0 1 1 0	70
28	0 0 1 1 0 1 1 0 1 1 0 0 0 1 0 1 1 1 0 1 0 0 1 1 0 0	69.5
29	1 1 1 0 0 0 0 0 0 0 0 1 1 0 0 1 1 1 1 1 1 1 1 0 0 0	69.1
30	0 1 0 0 1 0 1 0 1 0 1 1 0 0 1 1 0 1 0 1 0 1 0 1 1 0	70.2
31	1 0 0 0 0 1 1 0 0 1 1 1 0 0 1 1 1 0 0 1 1 0 0 1 0 1	70.6
32	0 0 1 0 1 1 0 0 1 1 0 1 1 0 0 1 0 0 1 1 0 0 1 0 1 1	72.4

We got the characteristics of the combination that have a greater impact on experiment in theory, but whether it is effective in practice or not, in order to test whether the result is optimal, put these features into the experimental group and then trained classifier.

Finally obtained as shown in table:

TABLE III. THE CONTRAST TABLE

	dimensions	accuracy%	F_1 %
The original feature parameters	26	74.90	70.3
The combined scheme	13	79.00	77.5

It can be seen that when do experiment with 26 d characteristic at orthogonal experiment, the classification accuracy is 74.9%. After the theoretical analysis, the most influential characteristic parameters for the efficiency of experiments is found, taking them into the subsequent experiments, the accuracy is 79 %, it has been improved and the feature space is reduced and the F_1 is also improved from 70.3 % to 77.5%.

C. Training the classifier using textual features

When analyzing the audio text after speech recognizing, features selection is the first step. SVM is used to training the classifier by using the CHI statistic value, then the classifier is on the test set for testing.

TABLE IV. THE ACCURACY OF TWO KIND OF TEXT

Textual features(CHI)	Original text	Recognized text
accuracy %	90.15	84.20
F_1 %	88.60	82.35

D. Features fusion

In the study of audio classification, there are two ways usually: features fusion and decisions level fusion [8]. Features fusion means that extracting features from the audio files and audio text respectively, then training the classifier by the merged features. Decisions level fusion is that training classifier by the acoustic features and text data individually, then taken together the results in some way. In this paper, we used the method of features fusion. The biggest problem in features fusion is that the level of the phonetic characteristics and the text characteristics. For example, it's hard to say the relationship between the energy and the classes. So, the primary problem in features fusion is the conversion that the characteristics of the two form to a level. For the extracted audio features, they should be mapped to the text like "audio word". Quantization algorithm is used to achieve this mapping. In order to reduce the complexity of the calculation, The algorithm of LBG-VQ[9] is used to get the codebook from the training data. Once acquired the codebook, the feature vectors are mapped to the nearest "audio word" based on the codebook. After getting the "audio word", the TF-IDF[101 weighting function is used to calculate the weight that the "audio word" in the "audio text". Under the assumption that these two forms of audio and text "key words" were independent of each other. So we spliced together them directly to complete the fusion of these two kinds of pattern characteristics. The experimental results as shown in the figure below:

TABLE V. THE RESULT CONTRAST

Features	accuracy %	F_1%
A（Textual features）	84.20	82.35
A+B(Optimized acoustic + textual features)	87.30	85.45
C+B(Optimized Acoustic + textual features)	92.25	90.40

We can see that the accuracy of classification is improved after features fusion. The fusion of the phonetic features optimized and the textual features lead to the improved accuracy of the classifier, the accuarary is improved from 84.2% to 92.25% andt the the F_1 value is improved from 82.35% to 90.40%.So it can be concluded that combining characteristics effectively can training better classifier.

IV. CONCLUSION

When analyzing the speech, we usually separate the acoustic part and its semantic part to deal, it will lead to lose their complementary part and cause mistakes. In this paper, the acoustic features are optimized firstly and combined with the semantic characteristics then, the classification results were improved and it cost less time. It proves that the method of features fusion is effective. In the process of the fusion of the two kinds of features, we string them together simply, not considering the distribution of their weights; it's what we should do next.

REFERENCES

[1] YANG Da-Li, XU Ming-Xing, WU Wen-Hu.Study of Feature Selection for Speech Rccognition[J]. JOURNAL OF COMPUTER RESEARCH AND DEVELOPMENT. 2003, 40(7): 963-939.

[2] SHAN Song-wei, FENG Shi-cong, LI Xiao-ming. A comparative study on several typical feature selection methods for Chinese web page categorization[J]. Computer Engineering and Applications, 2003(22):146-148.

[3] E. Zitzler and L. Thiele. (1999) "Multiobjective evolutionary algorithms: a comparative case study and the strength pareto approach."IEEE Transactions on Evalutionary Computation, vol. 3. pp. 257-71.

[4] Chunyuan, Zhao "Discussion on the application of multimedia teaching on classroom teaching of advanced mathematics"[J]. Journal of shenyang Institute of Engineering(Social Sciences). Jul. 2010, Vol. 6, No.3, pp. 399-400.

[5] Ho S Y, Lin H, Li W H, et al. Orthogonal particles swarm Optimization and its application to task assignment problems [J].IEEE Transactions on Systems, Man and Cybernetics, 2008,38(2) :288－298

[6] Li X L, Zhao Q, Zhang C J. Research on multiple index optimization method of the orthogonal test design[C] // IEEE International Conference on Computer Science and Information Technology.[s.l.]:[s.n.], 2010:224－226.

[7] Bai Liang; Hu Yaali; Lao Song yang; Chen Jianyun; Wu Lingda; Feature analysis and extraction for audio automatic classification. IEEE International Conference on Volume 1, 10-12 Oct.2005:767-772

[8] Z. Zeng, Y. Hu, M. Liu, Y. Fu, and T.S. Huang, Training Combination Strategy of Multi-stream Fused Hidden Markov Model for Audio-visual Affect Recognition[C].Proc.14th ACM Int'l Conf.Multimedia(Multimedia'06),2006:65-68

[9] A.Gersho and R.M.Gray.Vector quantization and signal compression [M].Norwell, MA, USA: Kluwer Academic Publishers, 1991.

[10] Kazunari Sugiyama et al. Refinement of TF-IDF Schemes for web Pages using their Hyperlinked Neighboring Pages, 14th ACM Conference on Hypertext and Hypermedia 2003, Pages 198-207