Low SNR Speech Recognition using SMKLG

Qin Yuan
Suzhou University, Anhui, 234000, China

Abstract—While traditional speech recognition methods have achieved great success in a number of real word applications, their further applications to some difficult situations, such as Signal-to-Noise Ratio (SNR) signal and local languages, are still limited by their shortcomings in adaption ability. In particular, their robustness to pronunciation level noise is not satisfied enough. To overcome these limitations, in this paper, we propose a novel speech recognition approach for low signal-to-noise ratio signal. The general steps for our speech recognition approach are composed of signal preprocessing, feature extraction and recognition with simple multiple kernel learning (SMKL) method. Then the application of SMKL in speech recognition with low SNR is presented. We evaluate the proposed approach over a standard data set. The experimental results show that the performance of SMKL method for low SNR speech recognition is significantly higher than that of the method based on other popular approaches. Further, SMKL based method can be straightforwardly applied to recognition problem of large scale dataset, high dimension data, and a large amount of isomerism information.

Index Terms—SMKL; SNR; Speech Recognition

I. INTRODUCTION

Computer based automatic speech recognition started from 1950s. Given the human speech signal for certain language, the target of computer based speech recognition is to transform speech signal into the corresponding text or command through recognition and comprehension process [1, 4]. By means of computer based automatic speech recognition system, human and computer are able to communicate with each other in speech, namely, make computer accurately recognize the content of speech in a variety of situations and circumstances, even understand human language, in order to be applied by human. With the development of computer science and technique, artificial intelligence techniques, computer speech recognition technique has developed into an independent field attracting a number of researches from different fields [2, 3, 5]. Speech recognition and its related techniques can be widely used in automatic control, communication and electronic system, information processing and so on. In a word, the ultimate goal of computer based speech recognition technique is to make machine understand human language, realizing language communication between human and computer.

The first speech recognition system that recognizes isolate English and number was set up by the staff of Bell laboratory in 1952 [3, 6]. In 1960s, computer application promoted the development of speech recognition. Moreover, linear predictive coding (LPC) and digital processing (DP) techniques were developed for speech signal. 1970s have seen the development and success of LPC technique and hidden Markov model (HMM) [7,9]. In 1980s, a number of speech recognition algorithms for continuous signal were developed, where the speech recognition algorithm had an evolution from template matching technique to statistics model based technique [10, 11]. More specifically, HMM, as a breakthrough, achieved the state-of-the-art performance for continuous speech recognition. Then, parameter extraction and optimization, elaboration of model design and the system’s adaptive technique got great progress, making the speech recognition technique matured and beginning bring products to market [2]. In recent years, due to the development of computer technique and the wide application of internet, speech recognition greatly affects the industrial manufacture and daily life [12].

Usually, current speech recognition approaches are able to reach satisfied performance for signal with high SNR. However, for speech signal with low SNR, those methods are easy to be failed. It is worth noting that, in real application, speech signal mostly with low SNR which means that these methods are difficult to satisfy the real world applications [9, 11]. For this reason, how to improve speech recognition accuracy of low SNR signal to meet users’ needs has become increasingly important. Recently, research methods of recognition under low SNR mainly focus on traditional speech recognition model, such as artificially neural network (ANN) and deep learning, which are both based on statistical theory, have very high requirement of data regularity and shall be conducted under the assumption of infinite sample size. However, they show good performance when the training data is enough, and are still not robust enough to deal with the real world speech signal with low SNR, small samples.

Researches show that it cannot only get better mapping performance but also process isomerism or various data often existing in typical learning problem to combine different kernel functions. At this moment, multiple kernel method can show more flexible and effective processing ability. In addition, it can be also used as one ingenious analytical method to explain learning results, so that application problem can be understood more deeply and accurately. Moreover, synthetic kernel method is just a typical learning method in multiple kernel learning. This type of multiple kernel learning method is mainly achieved through linear combination of
various kernel functions. The structure diagram of its component is shown as Figure 2. The detailed description of multiple kernel learning can be found in Section 2.

For this reason, on the basis of analyzing relevant research achievement, on account of limitations and shortcomings of traditional speech recognition model, a deep searching about speech recognition under low SNR is conducted in this paper. What’s more, according to the functions and features for speech recognition, the general method step for speech recognition is analyzed and the application in speech recognition under low SNR of simple multiple kernel (SMKL) method is presented to improve speech recognition accuracy, so that methods for speech recognition under low SNR can be well applied in theory. Further, SMKL method can be well applied to conditions of large-scale sample data, complicated dimension and a great deal of isomerism information.

Our proposed SMKL method provides machine learning with widespread application prospect and rich design thought in the field of feature extraction, multi-class object detection and pattern recognition. In addition, SMKL method can get more effective recognition performance through combining different kernel functions according to certain rules. Meanwhile, classical machine learning approaches usually encounter problems, such as isomerism or data with complicated kinds, so it will be more reasonable selection to consider SMKL method under complicated conditions [1-3]. The proposed approach in this paper contains three procedures. (1) Preprocess the input speech signal. (2) Extract features for the preprocessed speech, and then match the extracted speech signal with the speech model in computer. (3) Output the matching results or transform them into specific instructions. Experimental results show that: (1) the accuracy of our SMKL method recognizing speech under low SNR is obviously higher than that in the case of SVM. (2) Our SMKL method can be well applied to the conditions of large-scale sample dataset, high dimension and a large amount of isomerism information. SMKL method provides widespread application prospect and rich design thought in the field of feature extraction, multi-class object detection and pattern recognition.

The main contributions of the proposed approaches are threefold. (1) It extracts multiple features from the speech signal for speech recognition, which improve the robustness significantly. (2) It seamlessly integrates the speech recognition problem with SMKL through using multiple kernels for multiple features. (3) This approach is computationally effective. These advantages will be verified in the experiment section. The remainder part of this paper is organized as follows. We first present the SMKL based low SNR speech recognition approach in Section 2, and then conduct the experiments in Section 3. Section 4 draws a conclusion of the whole paper.

II. THE PROPOSED SCHEME

In this section, we will propose the SMKL based speech recognition approach. This approach is composed of five main procedures, signal collection and processing, feature extraction, model training, recognition and evaluation, followed by a feedback procedure. The framework of the proposed approach can be found in Figure 1. In this section, we mainly introduce the main model, which contains training model and performing recognition. The rest procedures will be introduced in Section 3.

Figure 1. The framework of the proposed speech recognition approach

Now we will propose the idea of MKL and derive its dual form. Let i and j index the samples and m index the kernel. To reduce the use of symbol, we simply set that values of i and j are both from 1 to l , and the value of m is from 1 to M . Under linear separable circumstance, SVM directly uses hyper-plane for classification. But it is nonlinear when processing most problems, then kernel function shall be used for phenotype switching, transforming primal data j from low-dimensional space to high-dimensional space, so that the linear separable goal can be obtained.

Primal problem for SMKL, and its solution. The alternating optimization algorithm presented by Grandvalet is possible to solve problem the simple form of multiple kernel learning. First, for the optimization of problem, f, b,ξ, and d are fixed. Next, weight vector d is the objective function updated to decrease problem, f, b,ξ, and d are fixed. In this section, we show that the second step can be achieved with closed form. However, this method is short of convergence guarantee and possible to cause number problem, especially some elements in d are closed to 0. Note that these number problems can be solved through importing substitute algorithm of disturbance processing version, as shown by Argyriouin 2008.

We consider the following constraint optimization problems if another optimization algorithm are used:

\[
J(d) = \min_{f,b,\xi} \frac{1}{2} \sum_{m} \frac{1}{d_m} f_m^2 + C \sum_{i} \xi_i \quad \forall i
\]

subject to:

\[
s.t. \quad \sum_{m} \sum_{i} f_m(x_i) + y_i b \geq 1 - \xi_i, \quad \xi_i \geq 0 \quad \forall i.
\]

Then we will show that how to solve Equation (1) with simple gradient method. First, we note that objective function \( J(d) \) is actually the objective value for an optimum SVM. We can discuss and calculate \( J(\ast) \) with gradient and its kernel is just this method. We have to process Equation (1), which is a nonlinear object function constraint of simplex. Once \( J(d) \) is solved out, d is updated according to decrease direction, ensuring that this equality constraint and non-negative constraint d are
met. Equality constraint for gradient treatment can be decreased according to Luenberger’s work in 1984. Let $d_\mu$ be a non-zero entrance of $d$, $J(d)$ and $V_{red}J$ whose gradients are simplified have components:

$$
[V_{red}J]_m = \frac{\partial J}{\partial d_m} - \frac{\partial J}{\partial d_\mu}, \quad \forall m \neq \mu
$$

(2)

To get better numerical stability, we select $\mu$ as the index of vector $d$ of the largest component. This positive constraint is also considered at decrease direction. Because we want to reduce $J(\cdot)$, $-V_{red}J$ is a decrease direction, if there is an index $m$ making $d_m = 0$ and $[V_{red}J]_m > 0$, it will violate positive $d_m$ of the constraint to use this direction. For this reason, the component is set as 0 at decrease direction.

$$
D_m = \begin{cases} 
0 & \text{if } \left| \frac{\partial J}{\partial d_m} - \frac{\partial J}{\partial d_\mu} \right| > \sum_{s \neq m, d > 0} \left| \frac{\partial J}{\partial d_s} - \frac{\partial J}{\partial d_\mu} \right| \quad \text{for } \mu = m \\
-\frac{\partial J}{\partial d_m} + \frac{\partial J}{\partial d_\mu} & \text{if } d_m > 0 \text{ and } m \neq \mu 
\end{cases}
$$

(3)

The general update plan is $d \leftarrow d + \gamma D$, where $\gamma$ is step length. Once $D$ at one decrease direction has been solved out, we first need to look for the maximum allowable step length at that direction; then check whether the objective value is reduced. If the objective value is reduced and $d$ is updated, we can set $D_\mu = 0$ and normalize $D$ to observe equality constraint. This process will be repeated until the objective value stops reducing. At this point, we can look for the optimum step length $\gamma$, one dimensional is used to ensure the global convergence through suitable stopping criterion, such as Armijo.

The algorithm will stop when meeting stopping criterion. Our operation is based on dual gap, which will be introduced in remaining part in detail. The process for SMKL algorithm is shown as below: first, $n \leftarrow 0$; then randomly initialize $d^n$; repeat the above two steps. Let $K \leftarrow k(d^n)$, so that SVM can be selected to solve single kernel problem with kernels $K$ and $\alpha^n$, achieving

$$
d^{n+1} \leftarrow s^n \left( \frac{\partial r}{\partial \alpha^n} - \frac{1}{2} \alpha^n \frac{\partial H}{\partial \alpha^n} \right).
$$

Repeat $n \leftarrow n + 1$ until it converges.

III. EXPERIMENT

Flow diagram of experiment is shown as below: (1). Preprocess the input speech. (2). Extract features of the preprocessed speech, and then match the extracted speech signal with speech model in computer. (3). Output the matching results or transform them into specific instructions, as shown in the following figure. Preprocessing for speech signal mainly includes filtering, weighting, short-time windowing processing and terminal detection. Then extract then features of speech signal. Then save these feature data as specific feature file, being the basic data of multiple kernel learning. In this experiment, we choose two popular kernels. (1) Summation kernel $k(x, z) = \sum_{j=1}^M k_j(x, z)$. (2) Extended quadratic polynomial

$$
k(x, z) = ak^p(x, z) + (1 - a)k^q(z, x)
$$

A. Data Sources and Preprocessing

Training data in experiment of this paper is taken from 8 people’s pronunciations in different SNR (0, 5, 10, 15, 20, 25 db), and the collected speech samples are English pronunciations of words respectively consisted of 5, 10, 15, 20 English letters, in addition, each person pronounces different words for 4 times in various SNR. Under each type, there are respectively 192 (8x4x6=192) corresponding training samples. 10.75Hz is selected as sampling rate of the signal during experimental process. Also, pronunciations of 5 persons in various SNR are selected as test recognition sample.

Data preprocessing mainly includes pre-weighting of collected speech signal, in which filtering processing of

![Image](92x109 to 254x261)
data is finished with one transfer function. Then operations, such as windowing and framing, shall be conducted to data, which are finished with Hamming window. For data, feature extraction shall be conducted after preprocessing, in order to better classify data.

Figure 3. Flow diagram of experiment

**B. Feature Extraction**

It is mainly that extract feature parameters reflecting speech essence from primal speech signal for data feature extraction, forming proper vector sequence. The selectable speech feature parameters include time domain parameters, frequency domain parameters. Time domain parameters consist of short-time average zero-crossing rate, short-time average energy, and pitch period and so on. Short-time average zero-crossing rate and short-time average energy are usually used to detect speech terminals while pitch period is used to distinguish tone and unvoiced or voiced sound for Chinese characters. Frequency domain parameters include spectrum (MFCC and LPCC), short-time frequency spectrum (DFT and average spectrum of 10-30 channel filter banks) and the first three formants.

People’s auditory system is a specific nonlinear system and has different responds to signals with various frequencies, mostly logarithm relationship. LPCC factors are based on synthetic parameters and do not take use of hearing characteristics of people’s ears, MFCC parameters combine generation mechanism of speech with hearing perception characteristic. For this reason, MFCC with better performance is selected as feature parameters of speech. The solution of MFCC is shown in the following Figure 4.

Figure 4. Solving process of MFCC

Speech recognition can be grouped into narrow and broad speech recognition, the former means one technique extracting text content from speech signal while the latter means one technique extracting any interesting content. There are many classification methods for speech recognition, mainly including classification methods according to vocabulary, manner of articulation, speaker and recognition method; classification method according to vocabulary is applied in this paper. The vocabulary can be divided into small vocabulary, middle vocabulary and large vocabulary. Generally speaking, 10-100 entries is small vocabulary, that 100-500 entries is middle vocabulary and that more than 500 entries is large vocabulary. Usually, with the increase of vocabulary, recognition accuracy of speech recognition will decrease, so with the increase of vocabulary, the degree of difficulty for the study on speech recognition also gradually increases.

**C. Selection of Kernel Parameters**

We have to select proper kernel functions, kernel parameters and high-dimensional mapping spaces when classifying, so that we can get the separator with excellent learning and generalization ability. Error penalty factors C and σ are the key factors for SMKL [6], so these parameters have great effect on classification precision and generalization ability for SMKL.

<table>
<thead>
<tr>
<th>TABLE I. SELECTION FOR KERNEL PARAMETERS</th>
</tr>
</thead>
<tbody>
<tr>
<td>1: Select the scope of parameters, such as C and σ</td>
</tr>
<tr>
<td>2: A coarse search is conducted for the grid</td>
</tr>
<tr>
<td>3: Construct two-dimensional grid</td>
</tr>
<tr>
<td>4: Calculate mean values of accuracy rate for parameters in various group when predicting</td>
</tr>
<tr>
<td>5: A subtle search is conducted for the grid</td>
</tr>
</tbody>
</table>

Grid search method is introduced to optimize and select kernel parameters C and σ when selecting SMKL parameters. This method can be simply operated and easily understood. The procedure of kernel parameters selection is shown in the following Table 1. Note that, during step 1, C and σ are generally selected; during step 2, the step length is usually set 2 when a coarse search being conducted. The goal of step 5 is to make the obtained search result be more accurate. Generally speaking, areas with high predicting accuracy are selected, which is equivalent to decrease step length to conduct a binary search.

**D. Experimental Result and Analysis**

Features are input SMKL classifier to conduct speech recognition experiment under low SNR after sample feature being extracted. In this paper, grid search method is used to determine kernel parameters, for details, see section 3.3. SVM method and SMKL method are respectively used to conduct experiment, under each condition. The experiment is repeated for 20 times, then average recognition accuracy for each experiment as the final experimental results.

In the first experiment, SVM method and SMKL method are compared to verify the effectiveness of SMKL method in speech recognition under low SNR. In each experiment, 2/3 of each kind of data samples can be randomly selected as training sets, and remaining 1/3 can be test sets. Grid search method is used to determine kernel parameters during experimental process. The experiment is conducted under SNR of 0,5,10,15,20,25 and word number of 5,10,15,20, and experimental results are shown as Table 2. Experimental parameters are error penalty factors C and σ.

The average recognition accuracy of various word numbers under the same SNR. Seen from Table3, recognition rates of methods proposed in this paper under different SNR are obviously higher than that of SVM method. The recognition accuracy of the proposed methods in this paper are between 84.72%~88.30%,
mostly around 86%, while recognition rates of SVM method are mostly around 81%. Reasons for the above results include the following three aspects: First, SMKL exploits the ability of the nonlinear feature mapping through learning the implicit feature space. By means of learning nonlinear feature mapping, the classifier could adapt to data distribution well. The adaptability of SMKL comes from the flexibility and the adaptability of the parameters of kernels of SMKL. Second, the selection method for kernel parameters can adapt to dataset much better, in comparison with empirical parameter selection. Third, SMKL can be applied to the condition that sample data is large-scale, high dimension data, and data containing a large number of heterogeneous information.

The second experiment is to test the effectiveness of the SMKL algorithm. This research divided the data set into 10 groups by using the 10 groups cross validation ways. Every group took turns as the test samples, and the others are as the training samples. Then we can get 10 groups different training samples and test samples. The experiment tool ten times, then we can get ten groups classification accuracy. At last we can get the average classification accuracy of the ten experiments. Experimental parameters are error penalty factors \( C \) and \( \sigma \). During the experiment it put the recognition accuracy and AUC as the evaluation standards. The two indicators have monotonic relations. Althoughvariability of AUC is greater than that of accuracy, AUC is more reliable than accuracy; the main reason is that accuracy is just the performance pointer when \( TH = 0.5 \) while AUC is the average performance pointer of all possible \( TH \).

From the experiment results in Table 3 and Figure 4, we can see the recognition accuracy of each experiment is lower than that of AUC. The average recognition accuracy of the ten experiments is 87%, but in AUC it can be as high as 92%. Reasons are mainly from the following two aspects: (1) compared with traditional machine learning methods, SMKL can be applied to the condition that the dataset is high dimension, containing a large number of heterogeneous information. (2) The adaptability of SMKL comes from the flexibility and the adaptability of the parameters of kernels of SMKL, which can make get the more effective mapping performance. However, for the classical machine learning problems, it can often encounter the heterogeneous or the complex data, so it showed the high AUC and little standard.

In the third experiment, we evaluate the kernel parameter selection approaches. Grid search method is used to determine kernel parameters in this paper when recognizing with SMKL method. In above experiments, kernel parameters are obtained by experience and determined based on grid search method, so that the effectiveness of grid search method determining the kernel parameters can be verified. The fixed parameter determination method and grid search method can be used to determine kernel parameters for conducting experiment, in each experiment, 2/3 of images can be randomly selected as training sets, and remaining 1/3 can be test sets. The average training times in Table 4 are the error of mean square and average training time under grid search method. From the experimental results, average relative error, error of mean square and average training time are used as criteria evaluating experimental results.

### Table II: Low SNR Speech Recognition Results

<table>
<thead>
<tr>
<th>SNR/dB</th>
<th>Method</th>
<th>Letter number of words</th>
<th>Average recognition (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>SVM</td>
<td>81.23</td>
<td>83.21</td>
</tr>
<tr>
<td></td>
<td>SMKL</td>
<td>83.26</td>
<td>84.36</td>
</tr>
<tr>
<td>5</td>
<td>SVM</td>
<td>83.45</td>
<td>81.65</td>
</tr>
<tr>
<td></td>
<td>SMKL</td>
<td>86.14</td>
<td>82.12</td>
</tr>
<tr>
<td>10</td>
<td>SVM</td>
<td>84.12</td>
<td>83.45</td>
</tr>
<tr>
<td></td>
<td>SMKL</td>
<td>85.05</td>
<td>86.45</td>
</tr>
<tr>
<td>15</td>
<td>SVM</td>
<td>81.36</td>
<td>82.13</td>
</tr>
<tr>
<td></td>
<td>SMKL</td>
<td>83.23</td>
<td>85.56</td>
</tr>
<tr>
<td>20</td>
<td>SVM</td>
<td>81.96</td>
<td>82.16</td>
</tr>
<tr>
<td></td>
<td>SMKL</td>
<td>82.01</td>
<td>86.23</td>
</tr>
<tr>
<td>25</td>
<td>SVM</td>
<td>84.32</td>
<td>83.56</td>
</tr>
<tr>
<td></td>
<td>SMKL</td>
<td>85.023</td>
<td>86.21</td>
</tr>
</tbody>
</table>

### Table III: Performance of Low SNR Speech Recognition Using SMKL

<table>
<thead>
<tr>
<th>Experiment</th>
<th>Accuracy</th>
<th>Area under curve (AUC)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Run 1</td>
<td>86.24%</td>
<td>0.9184</td>
</tr>
<tr>
<td>Run 2</td>
<td>85.21%</td>
<td>0.9270</td>
</tr>
<tr>
<td>Run 3</td>
<td>86.12%</td>
<td>0.9459</td>
</tr>
<tr>
<td>Run 4</td>
<td>81.32%</td>
<td>0.9672</td>
</tr>
<tr>
<td>Run 5</td>
<td>88.69%</td>
<td>0.9187</td>
</tr>
<tr>
<td>Run 6</td>
<td>94.69%</td>
<td>0.9542</td>
</tr>
<tr>
<td>Run 7</td>
<td>89.17%</td>
<td>0.9564</td>
</tr>
<tr>
<td>Run 8</td>
<td>90.85%</td>
<td>0.9275</td>
</tr>
<tr>
<td>Run 9</td>
<td>88.14%</td>
<td>0.9433</td>
</tr>
<tr>
<td>Run 10</td>
<td>86.79%</td>
<td>0.9455</td>
</tr>
<tr>
<td>Average</td>
<td>87.72%</td>
<td>0.9395</td>
</tr>
<tr>
<td>Std</td>
<td>1.59%</td>
<td>0.0460</td>
</tr>
</tbody>
</table>
TABLE IV. CLASSIFICATION PERFORMANCE COMPARISON FOR DIFFERENT KERNEL PARAMETERS

<table>
<thead>
<tr>
<th>Parameters selection</th>
<th>Experience based kernel parameters selection</th>
<th>Determine kernel parameters with grid search method</th>
</tr>
</thead>
<tbody>
<tr>
<td>MAPE</td>
<td>9.3875</td>
<td>4.1902</td>
</tr>
<tr>
<td>MSE</td>
<td>429.12</td>
<td>386.91</td>
</tr>
<tr>
<td>Training time (second)</td>
<td>371.12</td>
<td>296.45</td>
</tr>
</tbody>
</table>

The classification performance comparison for different kernel parameters. The grid search method is more accurate than the experience-based selection method.

search method are respectively 4.1902 seconds, 386.91 seconds and 296.45 seconds, and the corresponding values under the fixed kernel parameter method are respectively 9.3875 seconds, 429.12 seconds and 371.12 seconds. So we can see that it expresses huge superiority in aspects of average relative error, mean square error and average training time to conduct experiment based on determining kernel parameters with grid search method. Reasons for the above results are twofold. (1) The kernel parameters for SVM are selected according to distributed information of input data, which can make kernel parameters for SVM have better adaptability than default kernel parameters. (2) Determine parameters of kernel function with grid search method, which can better map features from low-dimensional space to high-dimensional space, being propitious to improve recognition accuracy.

IV. CONCLUSION

On the basis of analyzing relevant research works, on account of limitations and shortcomings of traditional speech recognition model, a deep searching about speech recognition under low SNR is conducted in this paper. What’s more, according to the functions and features for speech recognition, the general method step for speech recognition is analyzed and the SMKL method is presented to improve speech recognition accuracy for low SNR speech recognition, so that methods for speech recognition under low SNR can be enriched in theory. SMKL method can be well applied to conditions of large-scale sample data, complicated dimension and a great deal of isomerism information. SMKL can be widely applied to the field of feature extraction, object detection and recognition. In addition, SMKL method can archive higher recognition performance by combining different kernels according to certain regulations. Meanwhile, classical approaches can usually encounter problems, such as isomerism or data with complicated kinds, so it will be more reasonable selection to consider MKL method under complicated conditions [1-3].

Experimental results show that: (1) the accuracy of SMKL method under low SNR is obviously higher than that in the case of SVM. (2) SMKL method can be well applied to the conditions of large-scale sample data, complicated dimension and a large amount of isomerism information. SMKL method provides machine learning with widespread application prospect and rich design thought in the field of feature extraction, multi-class object detection and pattern recognition. Moreover, it will take more calculated amount and long training time, so it’s the next research direction of this paper to improve the model’s training efficiency and the detection performance of algorithm under the circumstance of small training sample.

REFERENCES


© 2014 ACADEMY PUBLISHER