Lipreading Recognition Based on SVM and DTAK

He Jun
Jiangxi key Laboratory of Robot&Welding, Information
engineer college, NanChang University
Nan Chang, China
boxhejun@tom.com

Zhang Hua
Jiangxi key Laboratory of Robot&Welding, mechanic
electricity college, NanChang University
Nan Chang, China
Zhanghua_lab@163.com

Abstract—To enhance recognition accuracy of isolated words identification with small samples in lipreading, SVM is first introduced to act as classifier in this paper. As SVM is based on structural risk minimization, it solves the problem of pattern recognition under small samples, on the other hand, it avoids the unreasonable hypothesis in traditional classifier. To meet the requirement of fixed input feature dimensionality in SVM, several input feature dimensionality normalization methods were discussed and compared. including 3-4-3 data segmenting method, HMM based method and DTAK(Dynamic Time Alignment Kernel) based method. Two experiments were performed on the bimodal database, In the first experiment different input feature normalization algorithm were compared on SVM. Experiments showed that DTAK based normalization achieved the best result. in the second experiments SVM was compared with HMM under different number samples occasion. Experimental results showed that SVM performs better than HMM under small samples.

Keywords - lipreading ; SVM ; DTAK; HMM

I. INTRODUCTION

Lipreading is usually used to improve speech recognition accuracy in noise environment. And classification model used in speech are often introduced to lipreading directly. including DTW(Dynamic Time Warping). HMM(Hidden Markov Model), ANN(Artificial Neural Net), TDNN(Time Delay Neural Network) etc[1]. HMM is the most popular classification model in speech because of its excellent handling time-varying feature sequence capability. But there exist some unreasonable hypothesis in HMM, for example: each observed value is regarded as irrelevant, meanwhile, the training criterion which is based on the maximum likelihood is not optimal. Although ANN doesn’t have there presupposition, it’s based on Empirical Risk Minimization(ERM), the solution may be local optimum, moreover, there exists over-fitting in ANN. In comparison, Support Vector Machine(SVM) is based on Structural Risk Minimization(SRM)[2], making a comprise between model complexity and learning capability, so in theory, SVM is optimal in classification. In general, It has these merits: first, As sample number is often limited, SVM can get optimal solution with these limited samples based on SRM theory. On the contrary, traditional learning theory demand samples must be infinite. secondly, In theory, SVM is Global Optimization compared with traditional classifier; finally, with the help of building linear transformation function in high dimension space, SVM can solve non-linear separation problem in low dimension space. As experiments in lipreading are all based on database. the sample number is often limited. And inadequate for training HMM parameters. So SVM is introduced as classification model in lipreading.

In section 2 SVM theory is analyzed and several feature dimension normalization methods are discussed in section 3. section 4 perform lipreading recognition experiments, including multiclass classifier training method, And HMM is used to compare with SVM. section 5 concludes this paper with a summary and a brief discussion about existing problems and future work.

II. THEORY OF SVM

A. Rational of SVM Classification

SVM is actually a trained two class classifier, it can be used to judge the category of input feature vector $X$ between two categories. The decision function is equation (1).

$$g(x) = w^T \phi(x) + b$$  \hspace{1cm} (1)

Where $\phi(x)$ is a non-linear function, which can project $X$ to a new high dimension feature space, and realize linear separation in this new space. $W$ denotes the hyper-plane in high dimension space. $b$ is threshold value. SVM is excellent for its learning capability. As traditional classifier aims at realizing ERM. If the trained classifier based on ERM is too simple, generation will be bad, if this classifier is too complex, there exists over-fitting. The solution to this phenomenon is SRM. And SVM aims at solving this problem based on following formula:

$$\min \frac{1}{2} w^T w + C \sum_{i=1}^{N} \xi_i$$ \hspace{1cm} (2)

$$y_i (w^T \phi(x_i) + b) \geq 1 - \xi_i$$ \hspace{1cm} (3)

$$\xi_i \geq 0, i = 1,2,...N$$ \hspace{1cm} (4)

This work was supported by a grant from NanChang University fund

U.S. Government work not protected by U.S. copyright
Where $x_i$ belong to training set, each $x_i$ is associated with output $y_i \in \{+1,-1\}$, $\xi_i$ is a slack variance defining allowable error. Parameter $C$ realize the compromise between ERM and SRM, $C$ is often obtained by cross validation. The solution to SVM can resort to equation (5) using lagrange.

$$Q(\alpha) = \sum_{i=1}^{n} \alpha_i - \frac{1}{2} \sum_{i,j=1}^{n} \alpha_i \alpha_j y_i y_j (x_i \cdot x_j)$$  \hspace{1cm} (5)

Subject to the constrain: $\sum_{i=1}^{n} y_i \alpha_i = 0$, $\alpha_i \geq 0$, $i=1,2,...,n$.

, the optimal separation plane is as follows:

$$w^* = \sum_{i=1}^{n} \alpha_i^* y_i x_i$$ \hspace{1cm} (6)

According to equation (6), those $x_i$ corresponding to non-zero $\alpha_i$ are called support vector, In general, $\phi(x)$ needn’t to be known in advance. In the non-linear separation cases, $\phi^T(x_i)\phi(x_j)$ is often replaced by kernel function $K(x_i, x_j)$.

Radial Basis Function (RBF) is used in this paper, so the final decision function is as follows:

$$g(x) = \text{sgn} \{ \sum_{i=1}^{n} \alpha_i^* \exp(-\frac{|x-x_i|^2}{2\sigma^2}) + b \}$$ \hspace{1cm} (7)

B. The Rational of Multi-class Classifier

The rational of SVM mentioned above is only for two class. As to multi-class classification, there are two common technique in practice:

The first is “one-versus-all”[3]. k classifier would be trained for k class problem. To identify an unknown sample, choose the class which classifies the test datum with greatest margin.

The second is “one-versus-one”[4]. In this case, k(k-1)/2 SVM have to be created for k class samples. To classify an unknown sample, choose the class that is selected by the most classifier. While this involves building k(k-1)/2 classifier, training classifier may actually takes less time compared with “one-versus-all”, since the training data set for each classifier is much smaller. so this method is more popular in large datum case.

III. FEATURE NORMALISATION

A. Feature Dimension Normalization Based on DTAK

Standard SVMs expect a fixed-length feature vector as input, but speech is dynamic and this always leads to variable-length features. Different approaches have been presented to deal with the variable time duration of the acoustic speech units. In general, there are two methods: one is normalising the feature vector time dimension to fit the SVM input; the other one is normalising kernel to adapt the SVM to variable input feature dimension. In the first case, there are many existing methods. For example, In[5] two methods of sequence resampling are assessed. One is variable window size method which makes it possible to include the whole digit utterance for a given number of windows per digit by adjusting the size of the window to the digit duration. The other is fixed window size method which maintains the window size around a fixed number of analysis instants regardless of its coverage of the word, therefore, some information is missing for those of long duration. In[6] a kind of linearly elongating or compressing the feature sequence duration method is used. In[7] input feature dimension is divided based on 3-4-3 ratio, this method neglect the paragraph information existing in different words. In[8] HMM is introduced to perform segmenting data based on Viterbi algorithm.

Those methods mentioned above are all based on the concept that speech is composed of some isolated states. Their difference lies in using different segment data methods. In contrast, We have already mentioned the notion of kernel in SVMs, in fact, SVMs are examples of the more general class of kernel methods, that is, SVMs rely on a kernel to obtain a nonlinear decision function. This kernel defines the space in which the solution is sought and therefore its choice is problem-dependent. Kernel function is the important character of SVM, decision function can be obtained by the kernel (inner product in feature space), Shimodaira first proposed to embed Dynamic Time Alignment into kernel function[9], and this kernel is often called DTAK. Subsequently, the lipreading feature of variable dimensionality can be input to SVM directly. And standard training and recognition algorithm in SVM can be used.

Traditional DTW aims at looking for a shortest path (alignment function) between the tested feature and the stored template based on elongating or shortening time axis. In contrast, DTAK manage to look for the shortest path based on the accumulated similarity between the tested feature and the template.

Assuming $X,Y$ are two to be compared feature of different dimensionality, $\psi_j(k)$ and $\psi_j(k)$ is time-alignment function at instant k on two time axis respectively. Inner product or kernel function can be used to act as a kind of measurement between $X$ and $Y$.

$$X \circ Y = \max_{\psi_j,k} \frac{1}{M_\psi} \sum_{k=1}^{k} m(k) x_{\psi_j(k)} x_{\psi_j(k)} \hspace{1cm} (8)$$

Subject to the constrain:

$$1 \leq \psi_j(k) \leq \psi_j(k) + 1 \leq |X|, \quad \psi_j(k) + 1 - \psi_j(k) \leq Q$$

$$1 \leq \psi_j(k) \leq \psi_j(k) + 1 \leq |Y|, \quad \psi_j(k) + 1 - \psi_j(k) \leq Q$$

Where $M_\psi$ is a normalized factor: $M_\psi = |X| + |Y|$, $m(k)$ is a non-negative scale factor. It stands for the weight value of different path, $Q$ is a constant for limiting succession.
The solution to this optimizing problem can use dynamic programming algorithm, recursion formula is as follows:

\[
D(i, j) = \max \left\{ D(i-1, j) + x_i^T \cdot y, D(i-1, j-1) + 2x_i^T \cdot y, D(i, j-1) + x_i^T \cdot y \right\}
\]

(9)

Then inner product can be expressed as equation (10):

\[
X \cdot Y = D(|X|, |Y|) / (|X| + |Y|)
\]

(10)

The above inner product analysis is performed on feature with different dimensionality, if inner product is replaced by non-linear function, such as RBF kernel and embed into SVM, DTAK function is realized as follows:

\[
K_{DTA}(X, Y) = \phi(X) \cdot \phi(Y) = \max_{\phi \neq \psi} \frac{1}{M} \sum_{k=1}^{M} m(k)K_{RBF}(x_{\phi}(k), y_{\psi}(k))
\]

(11)

and the decision function can be expressed as equation(12):

\[
f(x) = \text{sgn}\left\{ \sum_{i=1}^{N} \alpha_i^* y_i K_{DTA}(x_i \cdot x) + b^* \right\}
\]

(12)

B. Feature Value Scaling

LIBSVM tools is used in this paper[4], because too small value would be submerged under large scope data occasion. According to equation (13), assuming \( x_n^i \) is ith dimensionality feature of the nth class. Feature value will be scaled to [0,1].

\[
x_n^i = \frac{x_n^i - \min x_n^i}{\max x_n^i - \min x_n^i}
\]

(13)

IV. EXPERIMENTAL RESULTS AND ANALYSIS

To test SVM in lipreading, we made a bimodal database which contain 10 Chinese words, each word repeated 40 times. in our experiments, using “leave-one-out”training method, that is, only one samples of each word was tested, the remains were used for training.

![Figure 1. Examples of recovered lip image based on inverse DCT. 3a is the original lip image, 3b,3c,3d,3e are recovered images based on 27,54,72,126 DCT coefficients respectively.](image)

Input feature vector is derived from the DCT(Discrete cosine transform) coefficients of ROI (Region Of Interesting)[10]. In most documents, the ROI is regarded as a square containing the image pixels of the speaker’s mouth region. Discrete cosine transform (DCT) is widely used in image compression. Image information can be saved in a few larger DCT coefficients, so feature dimensionality can be greatly reduced. In Fig1, there are some examples of recovered lip image based on inverse DCT. 3a is the original lip image, 3b,3c,3d,3e are recovered images based on 27,54,72,126 DCT coefficients respectively. First, ROI is divided into 9 blocks containing 8×8 pixels. And then each block is performed DCT operation. The DCT coefficients are regarded as feature vector of ROI.

A. Comparison of feature dimensionality normalization methods

In the first experiment, 3 normalization methods which based on HMM, 3-4-3 ratio and DTAK were tested. The input feature vector is composed of 27 static feature and 27 dynamic feature plus 27 dynamic feature plus 27 dynamic feature[], the HMM comprise 6 states and each state comprise 2 Gaussian mixtures. RBF kernel was used in DTAK, the experimental result is shown in table 1.

| Classifier | Normalization method | Recognition accuracy |
|------------|----------------------|----------------------|
| SVM        | 3-4-3 ratio         | 82.5%                |
|            | HMM                  | 84.6%                |
|            | DTAK                 | 85.5%                |

Experimental results show that DTAK normalization method achieve the best recognition accuracy, in contrast, 3-4-3 method is inferior to other two methods. Apparently, segmenting data based on 3-4-3 ratio may be unreasonable. That is, the obtained feature maybe does not reflect the real paragraph information exactly existing in speech.

B. Comparison between SVM and HMM

To compare the classifying accuracy of SVM+DTAK with traditional classifier, HMM is introduced to compare with SVM. the HMM comprise 6 states and each state comprise 1 Gaussian mixtures. The tested results is depicted In figure 2, it shows that when trainned samples is little, for example only 5 samples of each word, SVM recognition accuracy is 84%. In contrast, HMM recognition accuracy is only 75%. And this reflect the learning and generalizing capability of SVM under small samples occasions. But when trainned samples were increased to 150, HMM is becoming more better than SVM.

![Figure 2. Recognition accuracy comparison between SVM and HMM](image)
And with the increase of trained samples, HMM and SVM both approach 90% accuracy. The experiments showed that HMM require large enough samples to train its parameters, and SVM performs better than HMM under small samples circumstance.

V. CONCLUSION

In this paper, SVM is first introduced to act as classifier in lipreading, to meet the fixed feature dimensionality requirement, DTAK based method was discussed, meanwhile, several traditional feature dimensionality normalization methods were also compared, DTAK method achieved the best accuracy. To test the generalizing capability of SVM, HMM was used as comparison. The experiment show that SVM performs better than HMM under small samples.

REFERENCES

[1] G. Potamianos, C. Neti, G. Gravier, A. Garg, and A. Senior, “Recent Advances in the Automatic Recognition of Audio-Visual Speech”, In Proc. IEEE, 2003.

[2] Vladimir N. Vapnik. “The Nature of Statistical Learning Theory”. 1995

[3] R. Solera-Ureña, D. Martín-Iglesias, A. Gallardo-Antonio, C. Peláez-Moreno, F. Díaz-de-Maria, Robust ASR using Support Vector Machines[J]. Speech Communication, Volume 49, Issue 4, April 2007, P 253-267

[4] Chih-Wei Hsu, Chih-Chung Chang, and Chih-Jen Lin, A Practical Guide to Support Vector Classification. (2008-3-12)http://www.csie.ntu.edu.tw/~cjlin/libsvm/index.html.

[5] García-Cabellos, J., Peláez-Moreno, C., Gallardo-Antonio, A., Pérez-Cruz, F., Díaz-de-Maria, A., 2004. SVM classifiers for ASR: a discussion about parameterization. In: Proc. EU SI-PCO 2004, Wien, Austria, pp. 2067–2070.

[6] Sekhar, C., Lee, W., Takeda, K., Itakura, F., 2003. Acoustic modeling of subword units using Support Vector Machines. In: Workshop on spoken language processing, Mumbai, India.

[7] Clarkson, P., Moreno, P., 1999. On the use of Support Vector Machines for phonetic classification. IEEE Internat Conf. Acoust. Speech Signal Process. 2, 585–588.

[8] Ganapathiraju, A., Hanaker, J., Picone, J., 2004. Applications of support vector machines to speech recognition. IEEE Trans. Signal Process. 52,2348–2355.

[9] Shimodaira, H., Noma, K., Nakai, M., Sagayama, S., 2001. Support vector machine with Dynamic Time-Alignment Kernel for speech recognition. Proc. Eurospeech. Aalborg, Denmark, pp. 1841–1844.

[10] HeJun, ZhangHua. “Research on Visual Speech Feature Extraction”. 2009 International Conference on Computer Engineering and Technology (ICCET 2009). Volume 2, p 499-502.