Compensations for SVM in Text-Independent Speaker Verification

Xiang-Feng Lu, Jia Liu
Department of Electronics Engineering, Tsinghua University, Beijing, 100084
luxf04@mails.tsinghua.edu.cn, liuj@tsinghua.edu.cn

Abstract. Support Vector Machines (SVMs) technique, as a kind of pattern classifier, is widely used in pattern classification including speaker verification. We study the asymmetrical character of speaker verification that uses SVM since the asymmetry between true and imposter speaker training sets degrade recognition rate. Asymmetrical costs kernel is implemented and based on it, we introduce a new method that compensate for the SVM scores according to SVM models. Experimental results are attached and analyzed, which show improvements in performance. Since compensations in score field do not entail heavy computations, the method introduced can easily be applied to the standard system.

1. Introduction

Support Vector Machines technique, as a novel classification technique, has proved to be successful in various fields including image and audio classification. Recently, it was introduced into the realm of speaker recognition, and proved to be effective in various speaker recognition tasks. In Nist 2004 and 2005 Speaker Recognition Test, some of the SVM systems have produced close results compared with the traditional classic GMM systems [1].

Generally speaking, SVM can only be effective when appropriate kernel function is used. In the case of text-independent speaker verification, several kernels such as GLDS kernel [2] and the GSV kernel [10, 11] have been proposed. In either case, the original speech samples are first divided into frames and each frame is processed separately. However, in training of SVM, due to the variability of length of speeches, as well as the asymmetry between the length of target speech samples and the background speech samples in verification tasks, the samples of SVM is not treated fairly. The variances of sample counts may cause the classification thresholds fluctuate in different conditions. This problem is especially serious in text-independent speaker-recognition, since in such problems, usually a large gap exists in two classes. Several methods are suggested to mitigate this effect.

Authors of [3] proposed a normalization tactic that tries to normalize the kernel function of SVM. Compared with normalization of features, normalization of kernel preserved the scale information of features, and proved to be an effective method in image classifications. However, when transplanted to the field of speaker verification, it performs poorly because the instability of speech signals. It has been shown that
through subtle controls of cost parameter of SVM can improve the training model of SVM and compensate for the variance of sample lengths [4]. While this technique proved to be effective, empirical cost function parameter in SVM needs subtle tuning and require practical experience.

Through investigation of the train process of SVM, we find that the number of support vectors itself can be employed as a guideline for measure of the unbalance of the training models. Through appropriate modeling of the number of support vectors parameter and parameterized solution of the model, an improved result was observed.

In the following section, firstly the standard SVM is discussed. The degeneration of performance of standard SVM and its cause is then analyzed. After that through investigation of the correlation between SVM scores and corresponding support vector number, their relation is then implied. In section 3, we construct a model that describes this relation and obtain the rational parameters through parameterized solution. Section 4 reveals the performance of the baseline system and the altered system and compares the result. In the final section, the result is concluded.

2. Degeneration of SVM in Asymmetrical Condition

SVMs are originally proposed as an extension to linear classification problems. SVM binary classifier tries to find a hyper plane that best separate two classes with minimum misclassified samples. This goal is typically achieved by minimizing the cost function which is usually influenced by misclassified samples. In normal conditions, samples of two classes are averagely distributed as is shown in Fig. 1. However, performance degenerates dramatically when samples of one class and samples of the other class are not distributed in equilibrium. Since the kernel function of SVM treats two classes unfairly, the arithmetic tries to minimize misclassified samples of all samples instead of misclassified ratios of each class.

Table 1 shows the performance degeneration in asymmetrical conditions in standard SVM. Here we refer to samples of two classes as 2 class samples. Positive samples and negative samples are both random 2 dimension vectors that satisfy Gaussian distribution with $X_p \sim N(\mu_p, \sigma^2_p)$, $X_n \sim N(\mu_n, \sigma^2_n)$. Different proportion of positive sample count against negative sample count conditions are trained for each system and corresponding models are acquired. To evaluation the performance of different system, equal proportion of 2 class samples are test against each system model and the result obtained.

Fig. 2 shows this problem. According to this figure, the classifying hyper plane shifts in correspondences with increase of ratio of negative sample count against positive sample count. Especially in serious unbalanced conditions the hyper plane is serious biased. When such models are used for speaker verification, performance degeneration becomes distinct.
We observed that the Equal Error Rates drop alone with the drop of ratio of positive samples against negative samples. Also corresponding to the drop of the ratio, thresholds of Equal Error Rate (EER) also show a trend of decrease. In the problem of speaker verification, usually true speaker training samples are independent
and of varying lengths, the background training samples are shared by all models. As a result, equal error rates are reached at different thresholds for different speaker models. Thus a common threshold shared by different speaker models can hardly yield good result.

Table 1. Demo of degeneration of SVM in asymmetrical conditions

<table>
<thead>
<tr>
<th>Positive Training Sample Count</th>
<th>Negative Training Sample Count</th>
<th>Equal Error Rate</th>
<th>Threshold of EER</th>
</tr>
</thead>
<tbody>
<tr>
<td>1000</td>
<td>1000</td>
<td>22.15</td>
<td>0.0420</td>
</tr>
<tr>
<td>1000</td>
<td>2000</td>
<td>23.45</td>
<td>-0.7579</td>
</tr>
<tr>
<td>1000</td>
<td>4000</td>
<td>23.65</td>
<td>-0.8475</td>
</tr>
<tr>
<td>1000</td>
<td>6000</td>
<td>24.75</td>
<td>-0.9993</td>
</tr>
<tr>
<td>1000</td>
<td>8000</td>
<td>27.60</td>
<td>-0.9996</td>
</tr>
</tbody>
</table>

In fact, SVM can be viewed as a projection from the input vector space into the space marked by the kernel functions. In the projected space, samples of different classes are distributed sparsely and can be easily separated. However, in asymmetrical training conditions, projected vectors are not distributed as that in test conditions. As a result, training schemes that tries to minimize experience risks may not at the same time minimize true risks. In fields of speaker verification, true speaker verification test counts and imposter verification test counts are of the same scale. As a result, standard SVMs degenerate in such situations.

![Fig. 3. Trend of threshold of EER with respect of ratio of negative samples against positive samples. This figure shows that thresholds of EER go down in correspondence to the imbalance of 2-class samples.](image)

To address this issue, the training criterion of training can be altered to compensate for this unbalance, which is considered in [4]. Another method is not try to compensate in the training process, but accept this unbalanced nature, and tries to compensate for the final scores by using the information of models themselves which
is what we are trying to do. In chapter 3, we will discuss compensations in the field of scores.

3. Compensation for Asymmetrical SVM

It is clear that the asymmetry of training sample counts affects SVM scores. Note that SVM scores are calculated using models acquired from training process, some parameters of models can be selected as inputs for this compensation. Several parameters from SVM models which are calculated in the training process are chosen for evaluation to check whether they can influence the final scores and can be used as inputs for score compensation. Table 2 shows the correlation coefficients between positive support vectors count $N_p(N_a)$ and SVM test scores. This experiment wages 11000 tests on corpus collected by our Lab. in which male tests and female tests are separated. From Table 2 we can find a weak correlation between $N_p(N_a)$ and SVM test scores exists. However, experimental results show that coefficients of true speaker and imposter are not coherent. As is shown, for $N_p$, of male true speakers, correlation coefficient are of positive values, and for male imposters are of negative values. For $N_a$, opposite trend is observed. As a result, offsets on scores which only consider the effects of $N_p$ and $N_a$ is not enough. Offsets should take into count the effect of scores.

![Table 2. Correlation coefficients between $N_p$, $N_a$ and result SVM scores.](image)

After investigating the relationship of SVM model parameters and result scores, it is determined that the count of positive support vectors $N_p$ and negative support vectors $N_a$ can be used as input of the compensation part.

In order to compensate for the score fluctuation caused by asymmetrical data set, a compensation model is supposed to be constructed. We introduce a general model given by

$$S_i = f(S_o, N_p, N_a)$$

(1)

where $S_o$ stand for the original score calculated from SVM test system, $S_i$ stand for the compensated score and $f(\bullet)$ stand for a general function. The target of the function is trying to find a function expression that can make result scores $S_i$ easy to differentiate and yield good performance. To specify this problem, the EER
criterion is adopted as evaluation for the performance. The problem can be reconsidered as trying to find a general function that result in the lowest EER. Let \( S_T = (S_{T1}, S_{T2}, \ldots, S_{Tn_T}) \) stand for scores of true speaker test scores, and \( S_I = (S_{I1}, S_{I2}, \ldots, S_{In_I}) \) stand for scores of imposter speaker test scores. \( N_{Tp} = (N_{T1}, N_{T2}, \ldots, N_{Tn_T}) \), \( N_{Ip} = (N_{I1}, N_{I2}, \ldots, N_{In_I}) \) stand for corresponding positive support vectors counts; \( N_{Tn} = (N_{T1}, N_{T2}, \ldots, N_{Tn_T}) \), \( N_{In} = (N_{I1}, N_{I2}, \ldots, N_{In_I}) \) stand for corresponding negative support vectors counts. \( E(\bullet) \) stand for EER function. \( E(\bullet) \) is the function of true speaker scores \( S_T \) and imposter scores \( S_I \) which is given by

\[
EER = E(S_{TT}, S_{IT}) = E[f(S_{TO}, N_{Tp}, N_{Tn}), f(S_{IO}, N_{Ip}, N_{In})]
\]  

(2)

Here

\[
f(S_{so}, N_{sp}, N_{sn}) = (f(S_{so}, N_{sp}, N_{Tn}), f(S_{so}, N_{sp}, N_{In}), \ldots, f(S_{so}, N_{sp}, N_{Tn}))
\]

(3)

means each dimension of scores is calculated independently.

The goal of optimizing general function \( f(\bullet) \) is to minimize EER. That is

\[
f = \arg \min_f [E(f(S_{TO}, N_{Tp}, N_{Tn}), f(S_{IO}, N_{Ip}, N_{In}))]
\]

(4)

It is obvious that function \( f(\bullet) \) is not specific and thus mathematically incalculable. At the mean time, due to the scale of dataset, sophisticated model may not yield good result. Note that the correlation factors of \( N_{p} \) (\( N_{s} \)) and SVM scores show different trends for true speaker and imposter set. Enlightened by the idea proposed by the author of [7], a compensation function of quadratic type is adopted as follows.

Let \( S_v = (S_{s_p}, N_{p}, N_{s_s}, S_{s_s}, N_{s_s})^T \), \( S_v \) is a 4 by 1 row vector, then

\[
S_v = f(S_{so}, N_{p}, N_{s_s}) = S_v + S_v^T \bullet V
\]

(5)

This model considers the influence of \( N_{p} \) and \( N_{s_s} \), and combines them in a quadric form. \( S_{s_s} \) is offset by parameterized expression, which considers the effect of \( N_{p} \) and \( N_{s_s} \) with respect to \( S_{so} \).

\( V \) is a 4 by 1 row vector parameter that can be tuned to optimize the model. By mean of min search techniques, we anticipate a performance gain in the final result. Minimum search arithmetic can be simple gradient descendental method or other global minimum search schemes.
4. Experimental Results

We have implemented 3 systems and performed several experiments on them based on the telephone corpus collected at our laboratory. Firstly, 3 Systems are described in detail. Then the database in which we do the experiments is mentioned. Finally the performance are given and analysed.

4.1 System description

Three systems are evaluated and compared in this paper. One is baseline system; another one, the asymmetrical system, is the system that use different cost factors for positive and negative samples as suggested by [4]; the third one is the score compensation system, revision of the second system that perform SVM compensation suggest by this paper.

The baseline system is a standard SVM system utilizing text-independent generalized linear discriminant kernel [2]. Baseline system is composed of 3 parts: feature extraction, model training and speaker verification.

The first part deals with the feature extraction task. It is composted of pre-filter unit, Mel Frequency Cepstral Coefficients (MFCC) extraction unit, the gaussianization unit and the GLDS wrapper unit.

In the pre-filter unit, firstly there is a real speech detector. Real speech is acquired through two steps. The input speech is processed by G.723.1 VAD (Voice Activity Detection) detector [5], which is employed to remove long term silence. Then the active speech is processed frame-by-frame with a frame length of 20ms and frame rate of 10ms. The bottom 20% frames ranked by energy are removed with a dynamic threshold. After that with a pass band filter from 250 to 3450Hz, a clean speech is obtained.

In MFCC extraction unit, the 19 dimensions MFCC features are extracted from each frame. Delta MFCC features are also computed.

Short term gaussianization approach [6] is applied to ensure the 38 dimensions vectors close to that of Gaussian distribution. This parser also mitigates the negative effect of different channels, which in turn contribute to the improvement of recognition performance.

The features sequences are then sent to the GLDS parser. The GLDS kernel calculates monomials of up to degree 3. Each monomial is the product of every 3 dimension feature group of the total 38 dimensions MFCC vector. Each 38 dimensions feature vector will produce one frame of 10659 dimensions super-vector. Certain count of such vectors is averaged to produce one SVM training vector or test vector [8, 9].
The second part deals with training task which calculates a model that marks the target speaker. The target speaker samples and the background speaker samples are send to feature extraction in parallel order. Then the result vector sequences are used for training of speaker model. SVMTorch is employed for training [12]. Finally a model marked by the target speaker is acquired.

The final part deals with the verification task which verifies a certain speech sample is said by a target speaker. Vectors of unknown speaker samples are extracted and send to the SVM test core. Such vectors are then evaluated against designated model to verify whether the speaker matches the true speaker. This also employs part of SVMTorch.

In the second system, we implement a kernel that treats positive and negative samples with different cost. The asymmetrical cost factor system alters the model training process by providing different cost factors for positive samples and negative samples.

The final system, that is Score Compensation system, offset the final score based on $N_p$ and $N_n$ by Eq. (4). Minimum search arithmetic used is a simple gradient descendent method starting at $V = 0$. 
4.2 Usage of Database

We performed several experiments based on the telephone corpus collected at our laboratory.

Data set is composed of training set and test set. Training set contains telephone speech of about 4 minutes from 600 speakers (280 for male and 320 for female). And test set contains 1000 true speaker verification tests (420 for male and 580 for female) and 10000 impostor speaker verification tests (4200 for male and 5800 for female). Male and female speaker are separate and no cross-sex tests is performed.

Since the dataset has separated male and female labels, we experiment individually. Final results are integrated. As for the test set, tests are divided into 2 subsets. One is for adapting the compensation models; the other is for test purpose.

For baseline system and asymmetrical costs system, only test subset is used.

4.3 Results and Analysis

Table 3 gives the EERs for different systems. From this table we can find that the asymmetrical costs system performs better than baseline system and the Score Compensation system performs best of the three.

<table>
<thead>
<tr>
<th>Table 3. EERs for different system</th>
</tr>
</thead>
<tbody>
<tr>
<td>Male EER%</td>
</tr>
<tr>
<td>Baseline System</td>
</tr>
<tr>
<td>Asymmetrical Costs System</td>
</tr>
<tr>
<td>Score Compensation System</td>
</tr>
</tbody>
</table>

![Fig. 7. 3 DET curves for male speakers.](image)
Fig. 7, Fig.8 and Fig.9 show the corresponding Detection Error Tradeoff (DET) curves for the 3 systems. Experiment results show that compared with the baseline system, EERs drop by about 0.45% and 0.85% for male and female speaker respectively. EERs also drop by about 0.37% and 0.56% compared with the asymmetrical costs system for gender separate cases. For all speaker set, 0.66% decrease of EER is observed in Score Compensation system compared with baseline system. The score compensation technique contributes to about 0.53% decrease.

Fig. 8. 3 DET curves for female speakers.
5. Conclusions

In this paper we introduced a compensation technique that employs SVM model parameters to normalize verification scores in speaker verification. This technique minimizes the EER by a minimum search method that finds the fittest parameter for the compensation model.

Experiments results have demonstrated that compensation of score according to support vector count is an effective way in improving system performance. When used in combination with asymmetrical costs, it can result in a performance gain by drops of EER by about 0.6%. It standalone can cause drops of EER by about 0.5%. Meanwhile since it is computationally economical, it can be added in standard system freely without heavy costs.

This paper only investigate into $N_P$ and $N_u$ two parameters deriving from the training model, which may not be representative, researches on model parameters which affect test score fluctuations will be future works. Note that parameters derived from different aspects can be fused to yield better performance. This fusion can be done by traditional fusion schemes like Artificial Neutral Network (ANN) or new ways.
Acknowledgement

This project is supported by the National Natural Science Foundation of China (NSFC) (60572083)

References