PROBABILISTIC DISTANCE SVM WITH HELLINGER-EXPONENTIAL KERNEL FOR SOUND EVENT CLASSIFICATION

Huy Dat Tran and Haizhou Li

Institute for Infocomm Research, A*STAR Singapore
1 Fusionopolis Way, #21-01 Connexis, Singapore 138632

ABSTRACT

This paper presents a novel method for sound event classification based on probabilistic distance SVM. The basic idea is to embed probabilistic distances into classical SVM to classify the sound events. The main point of this method is that the long-term characterization of sound events is better used in the classification compared to conventional method. Furthermore, taking into account the relative short time span of sound events, we develop a probabilistic distance SVM approach based on Hellinger distance from exponential modeling of temporal subband envelopes. An experiment on classifying 10 types of sound events was carried out and showed promising results of the proposed method compared to conventional methods.

Index Terms — Sound event recognition, sound characterization, subband temporal envelope, probabilistic distance, support vector machine.

1. INTRODUCTION

Sound event recognition is a relatively new research topic. It has a huge potential in different applications like media search, environment control, military, security surveillance, assistive living devices, etc. Live sound event recognition is typically more challenging as the decisions are based on short sound signals. Recently, sound event recognition has been attracting more attention from the speech and signal processing research community. In particular, the Rich Transcription benchmarking framework [1] incorporates unvoiced events such as coughing, breathing and footsteps into their speech recognition task. Acoustic event detection benchmarking [2], [3] is incorporated in the CLEAR and NIST evaluations, where the task is to detect sound events under meeting room conditions. Unfortunately, the prior work in [1]-[3] is a supportive utility for meeting room speech recognition applications, instead of a general sound event recognition framework. The proposed methods are commonly driven from well-known speech processing systems with conventional frame-based MFCC features.

The physical nature of sound events is much broader than that of speech, however their time-frequency representations are more distinctive. Therefore, the sound event recognition feature should be extracted from temporal characteristics throughout the entire event duration. In short, the popular method in speech recognition, with frame-based processing with HMM recognizer may not be the best choice for sound event classification.

The time-frequency discrimination principle has been applied in auditory-motivated feature extraction [4] as an alternative to the MFCC feature in speech recognition. Instead of using short-term FFT, the auditory-motivated methods decompose the waveform signals into subband components using auditory-motivated filters (wavelets), which give a higher resolution in the time domain. The subband temporal envelopes (STE) are normally used as features to characterize the audio signals in classification or recognition. However, previous studies on STE features for speech recognition, which rely on the same “short-term frame regime” on “deterministic” measurement, have not adequately captured the temporal information. In particular, its distribution over the entire span of a sound event.

Given the stochastic nature of sound signals, we believe that the probabilistic distribution and distance should be used to develop the classifier. Taking into account the short span of sound event, we will develop a probabilistic distance SVM based on exponential distribution of the STE and Hellinger distance. We refer the method as Hellinger-Exponential Kernel.

The organization of the paper is as follows. Sec. II introduces the fundamentals of probabilistic distance SVM. Sec. III presents the classification system, based on the exponential distribution of STE and Hellinger distance. Sec. IV reports experimental results and lastly Sec. V summarizes the work.

2. PROBABILISTIC DISTANCE SVM

In this section we will discuss both the linear and kernel SVM and then develop the theory for probabilistic distance SVM.

2.1. Linear SVM

Consider the problem of designing a separating hyperplane for m vectors \(\mathbf{x}_i \in \mathbb{R}^n\), \(i = 1, 2, ..., m\). Each point \(\mathbf{x}_i \in \mathbb{R}^n\) belongs to one of two classes, by its label \(y_i \in \{1, -1\}\), \(i = 1, 2, ..., m\).

The goal of linear support vector machines is to find the optimal separating hyperplane \(f(x) = \mathbf{w}^T \mathbf{x} + b\), which maximizes the margin, i.e. \(\frac{2}{||w||}\), or equivalently minimizes \(||w||\). For the solution, the non-negative variable \(\xi_i\) is introduced so that the soft margin can be found by quadratic programming:

\[
\min_{\mathbf{w}, b, \xi} \left( \frac{1}{2} \mathbf{w}^T \mathbf{w} + C \sum_{i=1}^{n} \xi_i \right)
\text{s.t. } y_i (\mathbf{w}^T \mathbf{x}_i + b) \geq 1 - \xi_i ; \ i = 1, 2, ..., n
\]

\[
\xi_i \geq 0,
\]

where the term \(\sum_{i=1}^{n} \xi_i\) denotes the upper bound of the misclassification from the training samples and \(C\) is a coefficient that regulates between the misclassification and the robustness of the classification (width of margin).

There are several ways to solve the optimization problem (1) which would return the solution in following form:
\[ w^T = \sum_{i=1}^{m} \alpha_i x_i^T; \]  
\[ b = \sum_{i=1}^{m} \alpha_i y_i, \]  
where \( \alpha : \alpha_i \) is called support vector.

### 2.2. Kernel SVM

The separating hyperplane can be denoted now in terms of the inner product of vectors \( x_i \)

\[ f(x) = \sum_{i=1}^{m} \alpha_i x_i^T x + b \]  

The linear SVM can be generalized into non-linear by replacing the inner product in (4) by a kernel function

\[ x_i^T x \rightarrow K(x, x_i). \]  

The separating hyperplane can be denoted by

\[ f(x) = \sum_{i=1}^{m} \alpha_i K(x, x_i) + b \]  

It can be proved that the SVM optimization solution can be found for any continuous, positive, semi-defined and symmetric kernel.

### 2.3. Probabilistic distance SVM

In conventional linear and deterministic kernel SVMs, the feature \( x \) must be a vector, hence in practice the feature should be averaged over the time frames. Such methods therefore can not take into account the temporal discrimination of sound events. The idea of this paper is to utilize a new class of kernel based on probabilistic distance between STE distributions, i.e

\[ K(x, x_i) = \langle D(x, x_i) \rangle, \]  

where \( D(x, x_i) \) is a probabilistic distance between the distributions of samples \( x \) and \( x_i \). With this novel class of SVM, the feature can be either vector, or matrix with different time span.

One of the main problems of kernel SVM is that the non-linear separating surface, in the general case, is extremely computationally intensive. This problem can be solved if the kernel can be factorized as a inner-product in a high-dimensional space

\[ K(x, x_i) = \varphi(x_i)^T \varphi(x) \]  

In this way, the separating surface can be linearized and denoted by

\[ f(x) = \sum_{i=1}^{m} \alpha_i \varphi(x_i)^T \varphi(x) + b = w^T \varphi(x) + b. \]  

### 2.4. Probabilistic Distance SVM with Hellinger distance

Recently, empirical probabilistic distance kernel has been studied in the context of texture classification [5], web document classification [6]. In speaker verification, the probabilistic distance kernel derived from single Gaussian distribution assumption [7] or conducted through GMM supervectors [8] have been proposed.

The main idea of the proposed method is to use a specific parametric distribution of the sound characterization to conduct the kernel classification.

The f-divergence [9] is used as probabilistic distance between samples’ distributions. Let \( f(t) \) be a convex function defined for \( t > 0 \), with \( f(1) = 0 \). The f-divergence is an average, weighted by the function \( f \), of the odds ratio given by probability distributions \( P \) and \( Q \).

\[ D_f(P||Q) = \int_x f \left( \frac{p(x)}{q(x)} \right) q(x) \, dx. \]  

As described in Section 2.2, the kernel SVM must satisfy three conditions 1) positive; 2) semi-defined; 3) symmetric. Hence, in this case, we can choose Hellinger distance, a special case of f-divergence, which satisfies all required conditions.

\[ f(t) = \frac{1}{2} (\sqrt{t} - 1)^2 \Rightarrow D_f(P||Q) = 1 - \int_x \sqrt{p(x)q(x)} \, dx \]  

### 3. SOUND CHARACTERIZATION WITH EXPONENTIAL MODEL OF STE

In this section, we will discuss ways to characterize sounds with STE distributions and measure the subband probabilistic distance (SPD) using SVM kernels, also denoted as the SPDSVM framework.

#### 3.1. Complex Mel-Gabor filters

As sounds are time-frequency localized signals, Gabor representations are known to be superior to Fourier representations, [10]. In this paper, we implement the Gabor representation through the complex Morlet wavelet [11], which can be called from the Matlab Wavelet Toolbox. The mother complex Morlet is an asymptotic wavelet with \( \omega_c \geq 5 \) [11], given by

\[ g(t) = \frac{1}{\sqrt{\pi f_b}} \exp \left( \frac{\omega_c t^2}{2} \right). \]  

Consequently, the frequency response of the complex Morlet wavelet can be obtained as

\[ G(\omega) = e^{-\frac{f_b (\omega - \omega_c)^2}{4}}. \]  

We see that the wavelet transform is a Gaussian band-pass filter, centralized at \( \omega_c \) and with a bandwidth of \( f_b \). Setting the shifting parameter \( b = 0 \) and using the scaling property of Fourier transform

\[ \frac{1}{a} \Psi \left( \frac{\tau}{a} \right) \leftrightarrow G(a\omega) = e^{-\frac{f_b (a\omega-c)^2}{4}}, \]  

we are able to find the relationship between the scale parameters and central frequencies of wavelets as

\[ f_s = \frac{f_b f_c}{a}. \]
3.2. Exponential modeling of STE

One important property of our online processing system, is that the sound sample has a relatively short duration, ranging from 0.5 to 2.0 seconds. This property motivates us to look into a model of a single parametric distribution.

Denote the output of the complex Mel-Gabor wavelet filters of a sound sample as

$$S = S_R + jS_N,$$  \hspace{1cm} (16)

where $S_R, S_N$ are the real and imaginary parts, respectively.

Starting with the conventional zero-mean Gaussian assumption of complex wavelet components

$$S_R, S_N \sim N \left( 0, \frac{\sigma^2}{2} \right),$$  \hspace{1cm} (17)

where $\sigma^2$ is the variance of the distribution in the specific subband.

It is easy to prove that the above assumption yields the exponential distribution of the STE square, denoted by

$$|S|^2 \sim \lambda \exp(-\lambda x),$$  \hspace{1cm} (18)

where $\lambda = \frac{1}{\sigma^2}$. Fig.1 shows examples of distribution modeling of STE calculated from examples of explosion and knocking sounds. We can see that the exponential and gamma distributions fit the STE histogram well. The comparison of STE models will be an interesting topic and we aim to investigate it in a near future work. In this paper, we will focus on exponential model leading to fast algorithm in the implementation.

Given $|S_1|^2, |S_2|^2, ..., |S_T|^2$ are STE of a sound sample in a subband. The maximum likelihood estimation of exponential distribution can be obtained in very simple form expression

$$\hat{\lambda} = \frac{N}{\sum_{i=1}^{N} |S_i|^2}. \hspace{1cm} (19)$$

In next paragraph, we will show that the squared Hellinger distance from exponential distributions also has a tractable form expression for the probabilistic distance kernel.

3.3. Construct kernel for sound event classification

Given two exponential distributions, with parameters $\lambda_1$ and $\lambda_2$, the squared Hellinger distance can be obtained in following form expression

$$D_f (\lambda_1 || \lambda_2) = 1 - \left( \frac{\lambda_1}{2\lambda_2} + \frac{\lambda_2}{2\lambda_1} \right). \hspace{1cm} (20)$$

As discussed in paragraph 2.3, in order to have a realistic implementation in practice, the probabilistic distance kernel (20) should be able to be factorized into an inner product form

$$D_f (\lambda_1 || \lambda_2) = \varphi^T (\lambda_2) \varphi (\lambda_1). \hspace{1cm} (21)$$

Hereafter, we refer to $\varphi (\cdot)$ as the kernel reproducing transform (KRT). We can see that KRT of Hellinger-Exponential kernel can be obtained in a closed form expression

$$\varphi (\lambda) = \left[ -\sqrt{\lambda}, 1, \frac{1}{2\sqrt{\lambda}} \right]. \hspace{1cm} (22)$$

The processing for probabilistic distance SVM with Hellinger-Exponential kernel is illustrated in Fig.2. We note that the over-band KRT is obtained by concatenating the subband KRTs (22). The proximal SVM [12] method is employed in our system.
4. EXPERIMENTS

4.1. Database

We validate the proposed framework using a sound event recognition experiment. The task is to classify unknown sound samples into one of the ten sound classes: normal speech, cry, scream, laugh, knock, breaking, door slamming, phone ring, explosion and clapping. The database consists of about 2 hours of audio taken from [13]. The total number of samples for testing, training and calibrating are 2794, 2782, and 1475, respectively. The audio samples are processed so that they are approximately two seconds in length. To validate the robustness of the proposed techniques, we played back and recorded the audio in two environments: in office with air-condition noise with SNRs of 12-18dB, and in a university canteen with SNRs of 5-10 dB. To evaluate the performance, we evaluate the classification accuracy in 10 cross-validations. In each run, one hour of training data and one hour testing data are randomly selected from the 2-hour training/testing database.

4.2. Evaluation methods

To validate our proposal, we would like to compare with conventional techniques.

1. MFCC-GMM [2]: 26-dimensional frame-averaged MFCC feature (12 MFCCs and deltas, plus log-energy and its delta) with 8-component diagonal GMM models.
2. STE-SVM: 36-dimensional averaged STE feature
3. STE-polySVM: STE feature with polynomial kernel SVM [14] \( K(x_i, x_j) = (x_i \cdot x_j)^2 \).
4. STE-SPDSVM: proposed probabilistic distance SVM with Hellinger-Exponential kernel.

Note that all the experiments are carried out in matched conditions.

4.3. Results and discussions

Table 1 reports the baseline performance of our findings. The results of MFCC-SVM was excluded because it is slightly worse than MFCC-GMM. We note that the given configuration of MFCCs was the best in the experiments. It can be seen that the STE feature with conventional SVM/GMM classifier only slightly outperforms the MFCC feature. Also, the polynomial SVM does not show any significant improvement over the linear SVM. However, the proposed STE-SPDSVM achieves significant improvement compared to all the other conventional methods. The p-test shows very good statistical significance of STE-SPD with Hellinger-Exponential kernel compared to baseline methods. The p-value comparing proposed method to MFCC-GMM over two conditions is just 0.04, which is considered to be very significant. Beside that, the computational cost of SPDSVM with Hellinger-Exponential kernel is lower than that of MFCC-SVM and MFCC-GMM. This makes the proposed method very promising for real time applications.

5. CONCLUSIONS

We proposed a novel method for sound event classification based on probabilistic distance SVM. We explored the sound characteristics through its exponential distribution modeling of subband temporal envelopes and build the kernel based on Hellinger distance. The proposed method extracts the long-term information of the time-frequency representation and outperformed the conventional frame-based MFCC methods in our tests. Furthermore, the Hellinger-Exponential kernel provides a fast signal processing algorithm which is promising for real time implementation. In the future, we intend to compare different distribution modeling of subband temporal envelopes to deliver the highest accuracy for SPDSVM.

6. REFERENCES


<table>
<thead>
<tr>
<th>Conditions</th>
<th>MFCC-GMM</th>
<th>STE-SVM</th>
<th>STE-polySVM</th>
<th>STE-SPDSVM</th>
</tr>
</thead>
<tbody>
<tr>
<td>office</td>
<td>93.1 ± 2.9</td>
<td>92.4 ± 2.1</td>
<td>94.1 ± 2.3</td>
<td>95.0 ± 2.1</td>
</tr>
<tr>
<td>canteen</td>
<td>58.1 ± 3.9</td>
<td>59.2 ± 3.5</td>
<td>58.1 ± 3.2</td>
<td>68.1 ± 2.8</td>
</tr>
</tbody>
</table>