ABSTRACT

This paper presents a novel method for audio event classification in overlapping conditions. The method is based on Jump Function Kolmogorov (JFK), a stochastic representation, which is (a) additive, thus the sum of signal and noise yields the sum of their JFKs; (b) sparse, therefore audio events are separable in this domain. The proposed method is an extension of our previous works for classification under noise-mismatch conditions. Similar to that approach, the robustness of the JFK feature is obtained by limiting them within confidence intervals, which can be learned in advance. However, in order to classify overlapped events, we design the classification system as a set of event detectors and develop a novel approach which maps JFKs to a specific feature for each detector. The experiment shows that the proposed method achieves promising results in very challenging overlapping conditions.

Index Terms— Jump Function Kolmogorov, Wavelet, Estimation, Classification, Robustness, Overlap, Multiple sources.

1. INTRODUCTION

Overlapping events is a major challenge of speech and audio recognition systems. The robustness of the state-of-art methods greatly degrades under such conditions [1]. Unfortunately, overlapping can not be prevented, as it happens everywhere in real life, from meeting rooms to outdoor environments.

Simultaneous recognition or classification of overlapped events is an even more difficult task and has not got enough attention from the research community. The current techniques are mostly designed for recognizing only a single source, while the other sources are considered as interference.

A closer research topic, which has been intensively investigated, is blind source separation, where prior knowledge of sources, i.e. statistical independence [2], sparse representation [3], special spatial distribution, non-negativity, etc., are used to design separating filters. However, despite there being a very large number of works that have been carried out in the last two decades, the present methods are still not robust enough to be applied in practice. Most of them are workable under only specific setups and conditions. Furthermore, as present blind source source separation methods aim to recover the sources for human perception, not for machine recognition, even in successful cases, the output signals are partially distorted and not suitable for the task of recognition or classification. It is clear that audio event classification in overlapping conditions needs special treatment.

Recently, overlapping events have received more interest from speech communities thanks to the organizers of speech separation challenges. Among the methods available in the literature, the Factorial HMM technique, based on the MIXMAX principle, is most suitable for simultaneous classification of overlapped signals [4]-[5]. This method, however, is based on the assumption that the logarithm of the power of a sum of two signals can be approximated by the max of the logarithms derived from each single signal. This assumption can be hold only if one source is strongly dominating, and is not always satisfied in practice.

To address the overlapping problem, we propose in this paper a novel approach based on Jump Function Kolmogorov (JFK) [6]-[7], a novel stochastic signal representation, what is: 1) additive, i.e. it returns a sum when the sources are overlapping; and 2) sparse, i.e. the sources are expected to be separable in this representation. Furthermore, we develop a classification system to take advantage of the JFK separability to deal with the overlapping audio events while using only a single microphone signal. This method is a direct extension from those developed in our previous works for noise-mismatch conditions.

The organization of the paper is as follows. Next, in Sec. II, we introduce the JFK instrument and its properties. In Sec. III we present the classification system for multiple overlapped sources. In Sec. IV we report and discuss the experimental results. Finally, Sec. V summarizes the work.

2. JUMP FUNCTION KOLMOGOROV FOR CHARACTERIZING STOCHASTIC SIGNALS

Sound events are signals that have distinctive tempo-spectral signatures, hence sound event recognition should rely on the temporal characteristics from the entire event duration. Given the stochastic nature of sound signals, the probabilistic distribution of the temporal-spectral representation should be used to characterize the audio event. Since one of most popular way to obtain the temporal-spectral representation of the audio signal, with high time resolution, is subband filters (wavelets), the stochastic representation of the temporal-spectral characteristics should be found in the probabilistic distributions of the subband wavelets.

2.1. Jump Function Komogorov

The most natural way to represent the stochastic information of subband wavelets is to use the probability density function (PDF). However, this instrument has a major disadvantage in realistic applications, which is its sensitivity to the noise. Let’s consider the distribution of the sum of two random variables, which can be understood as realizations of signal and noise

\[ Y = X + N. \]  

The PDF of the sum becomes a convolution

\[ p_Y (z) = p_X (z) \ast p_N (z), \]  

which would be totally distorted from the origins.
However, multiplication can be obtained in the characteristic function representation
\[ f_Y(u) = f_X(u) f_N(u). \] (3)
This becomes additive after taking logarithm
\[ \log[f_Y(u)] = \log[f_X(u)] + \log[f_N(u)], \] (4)
and further differentiation,
\[ \log[f_Y(u)]^{(n)} = \log[f_X(u)]^{(n)} + \log[f_N(u)]^{(n)}, \] (5)
where \( g^{(n)}(.) \) denotes the n-order derivative.

Taking the inverse Fourier transform, we could come back to a real representation which is a non-negative, monotonously increasing function
\[ k^{(n)}_{(Y)}(z) = k^{(n)}_{(X)}(z) + k^{(n)}_{(N)}(z). \] (6)

The representation in (6) is proved to exist for \( n = 2 \) through Kolmogorov canonical representation of the characteristic function and hence we refer to it as Jump Function Kolmogorov. More details of JFK theory can be found in our previous papers [6]-[7].

2.2. Separability of JFK representation

The main advantage of JFK is its additivity, which has been proved in (1)-(6), and the sparsity. The basic idea here is that the differentiation in (5) flattens the logarithm of characteristic function and therefore transforms it into a sparse representation after the inverse Fourier transform.

Particularly, in [7], we proved that Gaussian and Poisson distributions yield delta functions in JFK representations. These results are next generalized for an important class of distributions, the generalized Poisson distribution, which can be defined as a superposition of a Poisson-distributed random number of random variables noted by
\[ \pi_g = \sum_{k=1}^{N} \alpha_k. \] (7)

The information of the distribution is encoded by both the intensity \( \lambda \) and the distribution of the impulse amplitude \( \alpha_k \) noted by its PDF function \( p_\alpha(x) \).

Remark: The JFK of the generalized Poisson distribution has a tractable analytical form expression as follows
\[ k_{\alpha_k}(x) = \lambda x^2 p_\alpha(x). \] (8)

As the impulse amplitude \( \alpha_k \) is localized, the JFK is also expected to be localized. The generalized Poisson distribution is mentioned here as this distribution family can approximate all the distributions with a finite variance. In this paper, the subband wavelets are modeled as a double-side symmetric distribution when the positive and negative parts following the generalized Poisson distribution. Later, we will show from examples that the subband-JFKs, estimated from audio signals, have the forms which can be modeled by (10).

2.3. Estimation of JFK

In previous paragraph we highlighted the theoretical fundamentals of JFK theory. Now we briefly summarize the JFK estimation from observations. In order to estimate JFK, we first estimate the empirical characteristic function
\[ \hat{f}(u) = \frac{1}{N} \sum_{k=1}^{N} e^{iu x_k}, \] (9)

where \( x_k : k = 1 : N \) are observations of a stochastic signal \( X(t) \).

In our previous works, we proved the estimation is unbiased and has consistency
\[ E\{\hat{f}(u)\} = E\left\{ \frac{1}{N} \sum_{k=1}^{N} e^{iu x_k} \right\} = \frac{1}{N} \sum_{k=1}^{N} E\{e^{iu x_k}\} = f(u) \] (10)
\[ D\{\hat{f}(u)\} = \frac{1}{N} \left| \frac{f(u)}{N} \right| \leq \frac{1}{N}, \] (11)
and therefore the estimation error is negligible when the number of samples is large enough. According to (6), JFK is the Fourier transform of the 2nd-order derivative of the logarithm of the characteristic function. Similar to the empirical characteristic function, this derivative can be estimated directly from observations.

As discussed above, the subband wavelets of audio signal has a zero-mean, double-sided symmetric form of PDF. Therefore, the characteristic function of audio subband wavelet is real and symmetric which also returns symmetric JFKs. The positive part of JFK is then estimated using FFT [6]-[7].

Fig.1 shows examples of JFKs of speech and baby cry, and their overlapping signal under car noise conditions.

3. OVERLAPPING AUDIO EVENT CLASSIFICATION

As discussed in Section II, JFK analysis in the wavelet domain allows us to separate sounds from different classes in theory. We now study how to benefit from this property in audio classification tasks, and hereafter this framework is referred to as WaveJFK. Fig.2 illustrates the processing diagram of WaveJFK for overlapping audio event classification.

3.1. Wavelet filters

The input audio signal is first decomposed into narrow-band components by a wavelet filterbank system [25]. In particular, we use
3.2. Normalization

In this step, the filtered signals in each wavelet subband are normalized to have zero-mean and unique variance. This normalization removes the power variability and puts the signal on the same scale before estimating the JFKs.

3.3. Subband-JFK estimation

After normalization, the subband-JFKs are estimated for each sample in both the training and testing databases. As the samples in our databases are manually validated, with roughly the same length and known endpoints, we do not apply any windowing. In other words, the whole sample is used in each estimation. In the case when the sample’s length varies and the endpoints of voice activity are unknown, we recommend a window of 1.0-2.0 seconds.

3.4. Learning confidence intervals

Since different audio events map to different areas in the subband JFK representation, it is possible to separate them. Similar to previous works [6]-[7], we first learn the confidence intervals of subband JFKs for each sound class that retain the intervals of high concentration. Unlike [4]-[5], where the task is single event classification, here, for overlapping event classification, we develop a multiple detection scheme. Therefore, the confidence intervals from the sound classes are used to map a sound sample to a specific feature for a particular detector. More details of the classification system will be described in the next paragraphs. Now we describe how to learn the confidence intervals from training data. To reduce the processing time, we randomly select 10 samples from each class of signals from the clean data. The subband JFKs are then estimated over the pre-selected data using a FFT-driven method [7] within a large interval of argument \( x \). Particularly, we calculate JFK in the interval \( 0 \leq x \leq 50 \) with a step size of 0.1. In each subband, the median-averaged JFK is estimated over samples from each class and the confidence interval is set at the level where the subband JFK exceeds 70\% of its peak.

\[
\left[ s_{\min}^{(i)}, s_{\max}^{(i)} \right] = \arg \min_x \left\{ k(x) > 0.7 \max \left[ k^{(i)}(x) \right] \right\}. \tag{12}
\]

In next paragraphs, we will describe how the feature is constructed for classifying multiple overlapped events.

3.5. Classifier for multiple overlapped events

After the feature vectors are constructed from the subband-JFKs, a template classifier can be employed. Unlike the previous works, when multi-class OAO-SVM was used, here, to simultaneously classify multiple overlapped events, one-against-all (OAA) SVMs are employed. The classification system includes a set of \( N \) binary SVM detectors, giving decision for each class of events. For each testing sample, the classification output can be can be single event, multiple events or even empty, depending on the outputs from each binary SVM.

One important point of the proposed classification system is that the feature is specifically designed for each event detector using its confidence intervals described in previous paragraph.

3.6. Constructing feature vector for each binary SVM

For each binary (one-against-all) SVM in the classification system, the feature is constructed using particular confidence intervals of the detecting event which are estimated form training. It is done by stepping one into a fixed number of points (say 100) and then repeating the subband-JFK estimation for each interval over all training and testing databases. The feature vector is constructed by concatenating the subband-JFKs estimated from each sample. The dimension of the feature vector is therefore the subband number multiplied by its resolution.

4. EXPERIMENT

In this section we report some preliminary results on an overlapping audio event classification task.

4.1. Database

The database is taken from our previous audio database [cite]. In this experiment, 60-minutes of 16kHz sampled audio data of speech, baby cry, foot steps, and explosion and were used to generate 6 overlapped events each of them consisting of two single events. The overlapped signals are generated by playing back and simultaneously recording play-lists of the two events. Then the clips are obtained by cutting the recording, using the start and end points from the first and second events respectively. Note that all the clips are manually balanced in advance. For the experiment, 100 clips from each of the 6-overlapped classes generated are used for training and another 100 clips are used for testing.

4.2. Methods and evaluation measurements

As there is no state-of-the-art method for overlapping audio event classification, we compare our proposed method to two methods: 1) Baseline MFCC-GMM single classification, i.e. the testing against four single event GMM models (4-component); 2) the MIXMAX
Table 1. Overall classification accuracy and false alarm rate over each sound event in percentage [%]

<table>
<thead>
<tr>
<th>Audio event</th>
<th>Method</th>
<th>WaveJFK</th>
<th>MIXMAX</th>
<th>Baseline</th>
</tr>
</thead>
<tbody>
<tr>
<td>Speech</td>
<td>TP</td>
<td>82.5</td>
<td>76.7</td>
<td>80.3</td>
</tr>
<tr>
<td></td>
<td>FA</td>
<td>7.6</td>
<td>12.0</td>
<td>12.5</td>
</tr>
<tr>
<td>Baby cry</td>
<td>TP</td>
<td>90.7</td>
<td>2.7</td>
<td>58.4</td>
</tr>
<tr>
<td></td>
<td>FA</td>
<td>10.7</td>
<td>4.2</td>
<td>10.7</td>
</tr>
<tr>
<td>Explosion</td>
<td>TP</td>
<td>83.4</td>
<td>4.5</td>
<td>72.5</td>
</tr>
<tr>
<td></td>
<td>FA</td>
<td>72.5</td>
<td>4.9</td>
<td>63.4</td>
</tr>
<tr>
<td>Footstep</td>
<td>TP</td>
<td>69.3</td>
<td>13.4</td>
<td>57.1</td>
</tr>
<tr>
<td></td>
<td>FA</td>
<td>16.7</td>
<td>7.1</td>
<td>29.3</td>
</tr>
<tr>
<td>Averaged</td>
<td></td>
<td>81.5</td>
<td>7.0</td>
<td>64.6</td>
</tr>
</tbody>
</table>

Table 2. Absolute accuracy of overlapping events in percentage [%]

<table>
<thead>
<tr>
<th>Audio event</th>
<th>Method</th>
<th>WaveJFK</th>
<th>MIXMAX</th>
</tr>
</thead>
<tbody>
<tr>
<td>Speech and Baby cry</td>
<td>TP</td>
<td>80</td>
<td>58</td>
</tr>
<tr>
<td>Speech and Explosion</td>
<td>TP</td>
<td>75</td>
<td>70</td>
</tr>
<tr>
<td>Speech and Footstep</td>
<td>TP</td>
<td>67</td>
<td>57</td>
</tr>
<tr>
<td>Baby cry and Explosion</td>
<td>TP</td>
<td>80</td>
<td>58</td>
</tr>
<tr>
<td>Baby cry and Footstep</td>
<td>TP</td>
<td>66</td>
<td>58</td>
</tr>
<tr>
<td>Footstep and Explosion</td>
<td>TP</td>
<td>65</td>
<td>56</td>
</tr>
</tbody>
</table>

method, one of very few methods investigated for speech recognition in overlapping conditions [5]. However, as our task is classification rather than speech recognition, in our experiment we simplify the HMM by a GMM classification. The basic idea here is the MIXMAX principle applied for Mel-Frequency Spectral Coefficients

\[
\log |Z(f)| = \max \{ \log |X(f)|, \log |Y(f)| \},
\]

where \( Z(f) = X(f) + Y(f) \) is the overlapped signal. From here, the PDF of the overlapped MFSC can be decomposed from PDFs and CDFs of clean signals

\[
p_Z(\alpha) = p_X(\alpha) c_Y(\alpha) + p_Y(\alpha) c_X(\alpha).
\]

In our experiment, the 4-component GMMs are first trained using signals recorded from 4 original events. Then the 6 overlapped models are created using (14). The final classification is based on 10 GMM models (4 single and 6 overlapped). We refer this method as MIXMAX.

As evaluation measurement we calculate the classification accuracy and the false alarm over each single sound class (i.e. speech, baby cry, foot steps and explosion) and then averaged over the class. The classification accuracy over each class is calculated as the ratio between number of overlapped audio clips which correctly output the given class, to the total number of clips containing the given class (i.e. 450). Analogously, the false alarm is the ratio to the total number of clips which do not contain the given class (i.e 450), to those from them which wrongly output the given class in the classification.

4.3. Results and discussions

Table-1 reports the results of classification in terms of classification accuracy and false alarm over each class. We can see that the MIXMAX-GMM improves the classification accuracy compared to the baseline method. However the increment is not significant taking into account that the false alarm rate has also raised from 5% to 11%, respectively. Meantime, the proposed method yields a significant increment of the classification accuracy up to 16 % of absolute averaged classification accuracy when maintains the reasonable false alarm rate. We note that the low FA rate of baseline is due to that the number of output events is just half of the actual one. Therefore the accuracy is always low. While the baby cry yields the highest accuracy rate with the proposed ovWave-JFK, the explosion is more robust with the baseline single GMM classification. The results from proposed method are very close to those applicable in real-life applications. Among the investigated sound events, footstep is surprisingly found to be the worst performer. But in all cases, the proposed method is superior to the baseline methods. Table-2 confirms this superiority in term of absolute accuracy of overlapped samples. In this evaluation, the sample must give the correct output to both presenting events. A disadvantage of the proposed method, compared to MIXMAX-GMM, is the possibility of getting more then the actual number of sources. But it seems that this number is not significant in our experiment.

5. CONCLUSIONS

We propose a novel method for simultaneous classification of overlapped audio events. This method transforms the overlapped signal into separable JKF representations. Consequently, a one-against-all SVM system is developed to detect when more than one source is present in the signal. The preliminary experiment shows the good potential of the proposed method. The future work shall investigate a larger scale of audio events to mark the specific sound events which are suitable/unsuitable with the proposed method.

6. REFERENCES