JUMP FUNCTION KOMOGOROV AND ITS APPLICATION FOR AUDIO STREAM SEGMENTATION AND CLASSIFICATION

Tran Huy Dat, Li Haizhou

Institute for Infocomm Research
21 Heng Mui Keng Terrace, Singapore 119613

ABSTRACT

This paper proposes a new similarity measurement based on Jump Function Komogorov (JFK) and presents its application for audio content analysis. This is done by means of comparing JFK, a stochastic representation which is (a) additive, so a sum of sources yields a sum of JFK's, and (b) sparse, so the signal and noise are better separated in the JFK domain. The properties of JFK make it more robust than the probability density function when comparing the signal distributions. In the application, we use the JFK in wavelet domain for the audio stream segmentation and classification. The experimental results show that the proposed method is comparable to the conventional methods under normal condition but significantly outperformed them under miss-match conditions.

Index Terms— Similarity measurement, Jump Function Komogorov, Estimation, Segmentation, Classification.

1. INTRODUCTION

Multimedia content analysis is an important research direction with a wide list of possible applications in entertainment, health care, surveillance and security. Recently, as a counterpart of visual information, audio content analysis has been paid more attention due to its semantic discrimination capacity. Audio stream segmentation and classification are two major subjects of audio content analysis and in both cases the probability density function (PDF) is the main instrument that was used in the conventional approaches [1]-[3]. Particularly, for the segmentation, the Kullback-Leiber distance (KLD) derived from Gaussian distribution is used in [2] and for the classification the GMM/HMM are the most popular classifiers [1],[3]. A serious problem of using PDF instrument is that it become convolutive in the presence of noise and therefore is very sensitive to the change of noise conditions. The conventional methods are presently not effective in the low-SNR or miss-match conditions. To address this problem, many robust feature extraction and/or noise reduction method has been proposed, many of them were successfully implemented in certain conditions. However, in general, the robustness issue is remained as a significant problem for speech/audio signal segmentation and classification.

In this paper we shall discuss a different direction in the high level of the robust signal classification: developing a robust similarity measurement. In contrast to the PDF-driven measurements such as KLD or like-lihood, here we introduce the distance based on Jump Function Komogorov (JFK): a new stochastic representation, which is (a) additive, so a sum of sources yields a sum of JFK's, and (b) sparse, so different sources will have non-overlapping supports in the JFK representation. The similarity measurement based on Jump Function Komogorov is robust under noisy conditions as the noise subspace can be removed from the JFK representations.

In the application, we implement the proposed measurement for the audio content analysis. Unlike the conventional methods where the analysis is mainly done in the MFCC domain, here we analyze the JFK in the wavelet domain by adopting a wavelet filter set.

For the segmentation, the possibility of the environment change is justified by the JFK distance (JFKD) between left and right sub-windows. The skip point can thus be defined by thresholding the measured JKFD. In the experiment, we compare the proposed method to the method based on KLD [2].

For the classification, we estimate the JFK in each wavelet subband and compare the estimated JFKs to the trained references. To eliminate effects of the noise, a confident interval of JFK is applied. This interval is estimated in the training phase using a clean database and then being used in the test for the calculation of the JFK distance. We compare the proposed classification method to the method based on GMM classifier in the MFCC domain. To test the robustness of methods, we perform the test in both clean and noisy conditions while the reference models are trained only in the clean one.

The experimental results show the significant superiority of the JFK-driven compared to the PDF-driven approaches in the miss-match noisy conditions.

The organization of the rest of the paper is as follows. In Sec.2 we will introduce the Jump Function Komogorov, its properties and the estimation method. In Sec.3 we will then describe the proposed JFK-driven method for the audio stream segmentation and classification. Sec. 4 will report and discuss the experimental results. Finally, in Sec.5 we will summarized the work.

2. JUMP FUNCTION KOMOGOROV

2.1. Definition of Jump Function Komogorov

Let consider a fundamental problem of the distribution of the sum of two random variables which can be understood as realizations of signal and noise [4]

$$Y = X + N. \tag{1}$$

The PDF of the sum becomes a convolution

$$p_Y(z) = p_X(z) * p_N(z), \tag{2}$$

which would totally changed its form from the original one. However, the multiplication is satisfied in the characteristic function representation

$$f_Y(u) = f_X(u) f_N(u). \tag{3}$$

This becomes additive after taking logarithm

$$\log[f_Y(u)] = \log[f_X(u)] + \log[f_N(u)], \tag{4}$$

and further differentiation,

$$\log[f_Y(u)]^{(n)} = \log[f_X(u)]^{(n)} + \log[f_N(u)]^{(n)}, \tag{5}$$

where $g^{(n)}(.)$ denotes the n-order derivative.

Assume that there exist the Fourier transform of the n-order derivative of the logarithm of characteristic function

$$k_{(.)}^{[n]}(x) = \int_{-\infty}^{\infty} \log\left[f_{(.)}(u)\right]^{(n)} \exp(-iux)\, du, \tag{6}$$

this transform should return the additivity in real function space noted by

$$k_{(Y)}^{[n]}(z) = k_{(X)}^{[n]}(z) + k_{(N)}^{[n]}(z). \tag{7}$$

The function in (7) could be used as stochastic representation and it would be more suitable for the additive model of signal and noise.

The question is when does the Fourier integral (6) exist (i.e. when does it return a real function)? Fortunately, A.N. Komogorov, one of greatest mathematician of 20-century , had proved an important theorem [5] as follows

Theorem: Characteristic function of any finite-variance distribution can be presented in a canonical form noted by

$$f_\xi(u) = \exp\left\{ium_\xi + \int_{-\infty}^{\infty} (e^{iux} - 1 - iux)\frac{dK_\xi(x)}{x^2}\right\}, \tag{8}$$

where $f_\xi(u) = E\left[e^{iu\xi}\right]$ is the characteristic function of ξ, m_ξ - the expectation, and $K_\xi(x)$ is the Jump Function Komogorov (JFK) which is an increasing and bounded function satisfying the following inequality

$$0 \le K_\xi(x) \le \sigma_\xi^2(t), \tag{9}$$

Noted that here $E[.]$ denotes the expectation operator.

From (8), after taking logarithm and differentiation, it can be seen that Fourier transform (6) exists for the second order derivative of the logarithm of characteristic function. We denote it as follows

$$k_\xi(x) = \int_{-\infty}^{\infty} \frac{\partial^2 \ln[f_\xi(u)]}{\partial^2 u} \exp(-iux) du, \tag{10}$$

where $k_\xi(x)$ is the density of $K_\xi(x)$. Since this function will be used more than the original JFK, for an convenience, hereafter we apply the terminology JFK for the jump density function .

2.2. Advantage of JFK: Linearity and Sparsity

From (9), it can be seen that the JFK has the same mathematical properties as the PDF: this is non-negative and its integral is bounded. However, for the additive model, the advantages of JFK over PDF are its linearity and sparsity properties what make this representation better separable under noise conditions.

2.2.1. Linearity

The proof of the linearity of JFK can be seen from (1)-(7) and can be formulated in a general form as follows: Assuming the observed signal as a superposition of independent components

$$X = \sum_{i=1}^{L} X_i, \tag{11}$$

the JFK of X is then a superposition of the JFK of each component.

$$k_X(x) = \sum_{i=1}^{L} k_{X_i}(x) \tag{12}$$

2.2.2. Sparsity

One more important property of JFK is that many distributions have sparse representations in this domain. Two below examples show how the JFK map from Gaussian and Poisson distributions to the delta-functions.

Remark 1: The JFK map from Gaussian distribution to a delta function

$$k_\omega(x) = \sigma^2 \delta(x). \tag{13}$$

It can be seen that the JFK of the Gaussian noise is located at only zero-point in the horizontal axis and therefore is easy to remove.

Remark 2: The JFK maps from Poisson distribution to another delta function

$$k_\pi(x) = \Lambda\delta(x - 1.) \tag{14}$$

In sections 4, we will show that the separability of speech and noise in the JFK domain is useful to improve the robustness of the signal classification.

2.3. Estimation of JFK

In this paragraph we turn our attention to the JFK estimation from observation. In this paper, we estimate JFK through empirical characteristic function, which is driven from observations.

2.3.1. Estimation of characteristic function

The empirical characteristic function is estimated as follows,

$$\overline{f}(u) = \frac{1}{N}\sum_{k=1}^{N} e^{iux_k}, \tag{15}$$

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

Here after we prove the unbias and consistency of this estimation.

Unbias: The expectation of the estimation (15) yields the true characteristic function.

$$E\{\overline{f}(u)\} = E\left\{\frac{1}{N}\sum_{k=1}^{N} e^{iux_k}\right\} = \frac{1}{N}\sum_{k=1}^{N} E\left\{e^{iux_k}\right\} = f(u) \tag{16}$$

Consistency: Now we investigate the variance of the estimation

$$D\left\{\overline{f}(u)\right\} = E\left\{\overline{f}(u)\overline{f}(-u)\right\} - |f(u)|^2. \tag{17}$$

The first component in the right side can be expressed by

$$E\left\{\overline{f}(u)\overline{f}(-u)\right\} = \tfrac{1}{N^2} + \tfrac{1}{N^2} \sum_{\substack{k,s=1 \\ k \ne s}}^{N} E\left\{e^{iux_k} e^{iux_s}\right\}. \tag{18}$$

Assuming the independence of observations yields

$$M\left\{\overline{f}(u)\overline{f}(-u)\right\} = \frac{1}{N^2} + \frac{N^2 - N}{N^2}|f(u)|^2. \tag{19}$$

Substituting (19) into (17) yields a constrain for the variance of empirical characteristic function as follows

$$D\left\{\overline{f}(u)\right\} = \frac{1 - |f(u)|^2}{N} \le \frac{1}{N} \tag{20}$$

From (20) it can be seen that the estimation error is negligible when the number of samples is large enough.

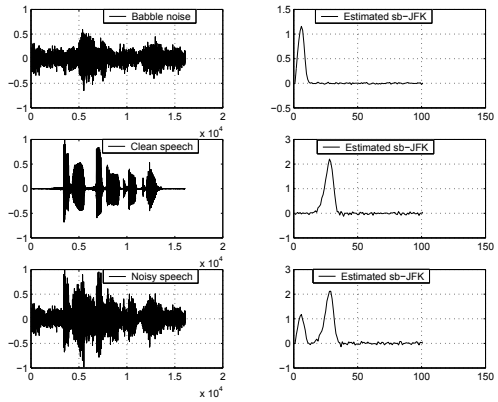

Fig. 1. Example of estimated JFK in a wavelet subband of 1) babble noise; 2) clean speech; 3)5-dB noisy speech

2.3.2. JFK estimation

According to (10), the JFK is Fourier transform of the 2rd-order derivative of the logarithm of characteristic function. Similar to the empirical characteristic function, this derivative can be estimated directly from observations. For our approach, where the JFK is estimated for the audio signal in the wavelet domain, it is reasonable to suppose that the observations have zero-mean and symmetric PDF and therefore the imagination part of the characteristic function can be ignored. The empirical characteristic function in this case can be denoted as

$$\hat{f}(u) = \frac{1}{N}\sum_{k=1}^{N}\cos(ux_k). \tag{21}$$

The second derivative of logarithm of characteristic function is thus estimated by

$$\frac{\partial^2[\log(f)]}{\partial^2 u} = -\frac{\sum_{k=1}^{N} x_k^2\cos(ux_k)}{\sum_{k=1}^{N}\cos(ux_k)} - \left[\frac{\sum_{k=1}^{N} x_k\sin(ux_k)}{\sum_{k=1}^{N}\cos(ux_k)}\right]^2. \tag{22}$$

The JFK is estimated by inverse Fourier transform of (22). In this work, the method based on FFT was implemented [6]. Note that as the JFK should be a real function the imaginary part of the estimation is ignored.

3. AUDIO CONTENT ANALYSIS BASED ON JUMP FUNCTION KOMOGOROV

In this section we describe how the JFK can be applied for the audio audio stream segmentation and classification. The flowchart of the processing is as follows. In the first level, the continuously recorded audio stream is segmented into audio clips according to the change of some statistical properties. Then each audio clip is classified into classes by adopting some classifier in feature domain. The reference models should be trained priorly.

Up to present, the main instrument used in both the segmentation and classification are mainly driven from the probability density function (PDF). For the segmentation, the KL distance or the histogram distance are more frequently used. For the classification, the likelihood measurement with GMM/HMM modeling is the most popular method. Unlike the conventional methods, our approach is based on the Jump Function Komogorov. Below we will describe the processing in details.

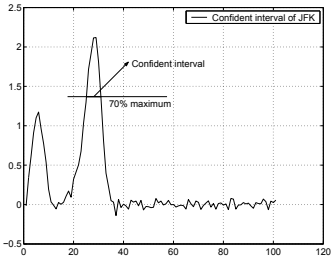

Fig. 2. Example of setting confident interval of subband JFK

3.1. JFK analysis in the wavelet domain

In contrast to the conventional method, where the cepstral domain (MFCC) is used, here we apply the JFK analysis in the wavelet domain. The audio signal is filtered by a set of wavelet filters. Particularly, a system based on Gabor filters located in Mel-frequency centers were used in our experiment. The JFK is estimated just from the filtered waveform in each subband. Given the visual property of the sound signals, this is reasonable to assume that the distribution of waveform in each subband wavelet domain follows is zero-mean and symmetric, so that the characteristic function and JFK can be estimated by the estimations described in Sec.2. Fig.1 shows an example of estimated subband JFK of clean speech, babble noise and simulated noisy speech at 5dB SNR. The sampling frequency in this example is 8kHz and the band is 4th from the 24-band Gabor-Mel band. It can be seen that the distribution of signal and noise are different and their JFKs are separable in JFK representation.

3.2. Segmentation based on JFK

For the segmentation, we processes stream by applying sliding windows, consisting of two-side sub-windows. For each sliding window, we estimate JFKs of two sub-window and the Euclidean distance between estimated JFks is used as a similarity measurement to justify the possibility of environment change in the particular window. The skip point is thus detected by thresholding the average JFK distance calculated over wavelet subbands.

3.3. Classification based on JFK

For the classification, for each class we first train the wavelet subband JFKs using a training database. The training is carried out in clean condition. In the test, the Euclidean distance between the JFKs estimated from the testing audio clip to the reference models are use to classify the clip. The simple nearest-neighbour principle is used to make the decision.

To improve the robustness of the classification, a confident interval of JFK is set from the training phase in order to avoid the invasion of the noise representation in the test. In this work, we set this interval as horizontal projections of 70 % of the maximum value to both left and right sides. An example of JFK of noisy signal and its confident interval is shown in Fig.2. It can be seen that the noise invasion in JFK can be reduced using this confident interval.

Table 1. Overall classification accuracy in percentage [%]

Methods	Clean	10-dB SNR	5-dB SNR
MFCC-GMM	92.22	52.27	22.35
Wavelet-JFK	91.19	86.67	80.56

Table 2. Segmentation accuracy in percentage [%]

Class	Clean	10-dB SNR	5-dB SNR
MFCC-KLD	95.50	80.34	65.87
Wavelet-JFK	92.35	82.54	74.32

4. EXPERIMENT

In this section we report the results evaluated from an audio stream segmentation/classification experiment. This task is a part of our project of CCTV for surveillance in public environment.

4.1. Data collection and reference methods

In the first stage of the project we are interested in real time detection of possibly aggressive sounds. The segmentation and classification of 8 classes of audio sounds including of normal speech, breaking, explosion, cry, baby cry, knock, laugh and scream are investigated in this work. We record a 3-hours audio stream consisting of above sound events. The data was collected in a clean condition. Each non-speech "audio clip" has approximate 2 second length and the overlapping of sound events will not be discussed in this work. We used one hour audio stream for training and another two hours were used for test. The sampling frequency is 44100Hz. The test audio stream was manually split into about 3600 clips. The training is carried out using clean signals. To explore the effect of noisy miss-match conditions, we artificially added the exhibition noise to the testing audio stream. Three test conditions of clean (no miss-match), 10dB-SNR and 5dB-SNR were investigated. The proposed segmentation and classification methods are implemented by the algorithm described in Sec.3 with a 64-Gabor-filter set. We refer this method as wavelet-JFK. As a reference method for the segmentation, the conventional method based on Kullback-Leiber distance of Gaussian distributions is implemented [2]. For classification, we implemented the GMM classifier as the reference. Both methods use the conventional MFCC features. We refer the reference methods as MFCC-KL and MFCC-GMM, respectively. Note that the JFK method can not be applied in MFCC domain as the additivity is not satisfied in this domain.

4.2. Results of classification with manual segmentation

In order to compare the methods in the classification, we first evaluate the classification using manual segmentation. Tab.1 shows the overall classification accuracy which is defined as the ratio of number of correctly classified clips to the total number of clips. It can be seen that the proposed wavelet-JFK is comparable to the MFCC-GMM in the clean condition but significantly outperformed it in noisy conditions. Under 5dB SNR, the MFCC-GMM totally failed while the proposed method still yields relatively high accuracy.

4.3. Results of Segmentation

Next, we evaluate the proposed wavelet-JFK segmentation method. The segmentation accuracy was calculated by alighting the segmented clips to the manual ones. The segment is considered to be wrongly segmented if 1) it does not contain start point and end point of an event from manual clips 2) contain more than start point and end point of events. Tab.2 shows the segmentation accuracy for the proposed and the conventional MFCC-KLD method. Although the superiority of the wavelet-JFK over the conventional method under noisy conditions is not as significant as for the classification, we got nearly 9% accuracy improvement under 5-dB SNR condition. The wavelet-JFK is also comparable to the conventional one in clean condition. The fact that the proposed method is more superior in the classification than the segmentation explains the effectiveness of applying the confident interval which seems to be able to eliminate the noise from the JFK representations when the test is carried out under noisy condition. Table 3 shows the final performance when combining both segmentation and classification.

Table 3. Classification accuracy with automatic segmentation: wavelet-JFK vs. MFCC-KLD/GMM [%]

Methods	Clean	10-dB SNR	5-dB SNR
Speech	84.44 vs 88.89	62.22 vs 41.56	54.22 vs 21.56
Cry	71.56 vs 72.00	51.56 vs 27.56	47.11 vs 12.00
Baby cry	69.11 vs 83.11	44.67 vs 35.56	40.22 vs 04.67
Breaking	95.56 vs 96.67	74.22 vs 56.89	60.89 vs 26.89
Explosion	100.0 vs 100.0	88.89 vs 61.78	69.33 vs 23.11
Knock	89.11 vs 89.11	78.00 vs 36.44	72.00 vs 13.78
Laugh	71.56 vs 68.89	55.56 vs 31.11	54.22 vs 12.67
Scream	94.89 vs 92.67	84.44 vs 31.33	69.33 vs 10.89
Overall	84.53 vs 86.42	67.44 vs 41.56	58.42 vs 15.69

5. CONCLUSIONS

We proposed a new similarity measurement based on Jump Function Komogorov and its application for audio stream segmentation and classification. The main advantage of the proposed method is the robustness under noisy and miss-match conditions.

6. REFERENCES

[1] Hain, T., et al., Segment Generation and Clustering in the HTK Broadcast News Transcription System, *in Proc. DARPA Broadcast News Transcription and Understanding Workshop,* 1998.

[2] H. Meinedo and J. Neto, Audio segmentation, classification and clustering in a Broadcast News task, *in Proceedings ICASSP 2003*, Hong Kong, China, 2003.

[3] J. Foote, Content-based retrieval of music and audio, *Multimed. Storage Archiv. Syst. II*,1997, pp. 138-147.

[4] M. Fisz , Infinitely divisible distributions: recent results and applications , *A*nn. of Math. Statist. , 1962, pp. 68-84.

[5] Gnedenko, B.V., and Kolmogorov, A.N. Limit distributions for sums of independent random variables, *Addison-Wesley (Translated from Russian)*, 1954.

[6] William H. Press, Saul A. Teukolsky, William T. Vetterling and Brian P. Flannery, Numerical recipes, *Published Cambridge University Press*, 2007.