Speaker Identification in Noise Mismatch Conditions based on Jump Function Kolmogorov Analysis in Wavelet Domain

Tran Huy Dat and Li Haizhou

Human Language Technology Department, Institute for Infocomm Research, ASTAR, Singapore
hdtran@i2r.a-star.edu.sg, hli@i2r.a-star.edu.sg

Abstract

We present a novel method for speaker identification in noise mismatch conditions. The proposed method is based on Jump Function Kolmogorov (JFK), a new stochastic instrument, which is (a) additive so sum of signal and noise yields the sum of their JFKs; (b) sparse so signal and noise have better separable supports in JFK's representations. The separability of signal's and noise's representations is the main advantage of JFK to make this instrument more robust in the classification then the conventional probability density function (PDF). In the approach, we develop a speaker identification system based on JFK analysis in the wavelet domain, i.e. the JFKs are estimated in each subband to match the nearest from trained templates. The experimental results show that the proposed method is comparable to the conventional method under clean condition but significantly outperformed them under noise mismatch conditions.

Index Terms: Speaker Identification, Jump Function Kolmogorov, Wavelet, Noise Mismatch Condition

1. Introduction

Speaker identification is a part of speaker recognition, a rich applicable area of biometric, information retrieval, business security, safety control, human-robot interaction and so on. The present speaker identification technology has archived quite high accuracy in ideal noiseless conditions [1]-[2] but the performance severely degrades under mismatch training and testing conditions [3]. One of reasons is that the conventional systems (i.e. GMM/HMM) are mainly driven from the probability density function (PDF) which becomes convolutive in the presence of noise and hence very sensitive by the changes in noise conditions.

To address this problem, a number of approaches have been developed in the literature [4]-[5]. The conventional methods were mainly developed on the way to reduce the noise or normalize/compensate the model mismatch between the different noise conditions while the classifiers are remained the same.

We shall discuss a different way to address the noise mismatch problem: to develop an alternative stochastic instrument which can better separate the signal's and noise's representations and consequently better eliminate the noise effects and make the classifier more robust.

We show that the wanted instrument could be the Jump Function Kolmogorov, a new stochastic instrument, which is (a) additive so sum of signal and noise yields the sum of their JFKs; (b) sparse so signal and noise have better separable supports in JFK's representations.

The basic idea of this instrument is that the logarithm of characteristic function is additive and this function can be flattened by differentiating and it will become sparse after taking the inverse-Fourier transform. The additivity and sparsity of JFK lead to the separability of signal and noise in this domain which is the key point of the robustness of the driven classifier in mismatch noise conditions.

In the application, we develop a speaker identification system based on JFK analysis in wavelet domain. The speech signal is passed through subband filters which decompose it into narrow-band wavelet representations. The JFKs are estimated in each subband and the classification is performed by comparing the wavelet JFKs from test samples to the trained templates. The noise effect in JFK is eliminated by setting confidence intervals from training phase which is carried out in clean condition. Unlike our previous system which was designed for audio stream classification [6], where the Euclidean distance of wavelet JFKs was used in the classification, here we adopt linear support vector machine (SVM) [7] to classify the wavelet JFK representations. The wavelet transform satisfies the additivity of signal and noise which is required for the JFK's separation and therefore suitable for our proposed JFK-driven method.

The proposed method is evaluated using the YOHO database [8] which was originally designed for the text-dependent speaker verification. To test the method in noise mismatch, the training is carried out in the clean condition while the test data is superimposed with different noises under different SNR conditions. The experimental results show that the proposed method is comparable to the conventional method in clean condition but significantly outperformed it in noise mismatch conditions. Particularly, the conventional method is totally failed when the test is carried out under 5dB and 0dB SNR conditions while the proposed wavelet JFK method could archive reasonable result in these conditions. The organization of the rest of the paper is as follows. Sec.2 gives introduction of JFK. Sec.3 describes the wavelet JFK system used in our speaker identification experiment. Sec. 4 reports and discuss the experimental results and Sec. 5 will conclude the paper.

2. Jump Function Kolmogorov

In this section we give a brief introduction of the Jump Function Kolmogorov which explains why this instrument can be considered to build a robust classifier. The more details of properties and estimation of JFK can be found in our previous work [6].

Let start with a fundamental problem of the distribution of noisy signal which can be presented as a sum of signal and noise

\[ Y = X + N. \]  

The PDF of the sum becomes a convolution

\[ p_Y(z) = p_X(z) * p_N(z), \]
which make it difficult to separate the noise’s and signal’s representations. However, the multiplication is satisfied by taking the Fourier transform of PDF to the characteristic function

\[ f_Y(u) = f_X(u) f_N(u). \]  

(3)

The additivity will be satisfied in the logarithm of characteristic function domain

\[ \log[f_Y(u)] = \log[f_X(u)] + \log[f_N(u)]. \]  

(4)

and further differentiation,

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

(5)

where \( \phi^{(n)}(.) \) denotes the n-order derivative.

The important point of the method is that the log-characteristic function are flattening by differentiating and therefore its inverse-Fourier transform, if existed, will map the distribution to a sparse representation which at the same time carries full information of the distribution. Denoting the Fourier transform of the n-order derivative of the logarithm of characteristic function

\[ k^{(n)}_\xi(x) = \int_{-\infty}^{\infty} \log[f_{\xi}(u)]^{(n)} \exp(-iux) \, du, \]  

(6)

this transform yields the representations which are additive and sparse and therefore could be used for the wanted robust classifier

\[ k^{(n)}_{\xi Y}(z) = k^{(n)}_{\xi X}(z) + k^{(n)}_{\xi N}(z). \]  

(7)

The question is when does the inverse Fourier integral (6) exist (i.e. when does it return a real function)? Here we should thank to great Russian mathematician A.N. Kolmogorov for his canonical representation of characteristic function [9]:

**Theorem:** Characteristic function of any finite-variance distribution can be presented in a canonical form as follows

\[ f_{\xi}(u) = \exp \left\{ iu m_{\xi} + \int_{-\infty}^{\infty} \left( e^{iux} - 1 - iux \right) \frac{dK_{\xi}(x)}{x} \right\}, \]  

(8)

where \( f_{\xi}(u) = E[e^{iux}] \) is the characteristic function of \( \xi \), \( m_{\xi} \) - the expectation, and \( K_{\xi}(x) \) - the an increasing and bounded function satisfying the following inequality

\[ 0 \leq K_{\xi}(x) \leq \sigma^2_{\xi}(t), \]  

(9)

Noted that here \( E[.] \) denotes the expectation operator.

From (8), it can be proved that Fourier transform (6) exists for the second order derivative of the logarithm of characteristic function. We denote the transform as

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

(10)

where \( k_{\xi}(x) \) is the density of \( K_{\xi}(x) \). Hereafter we call this function Jump Function Kolmogorov.

The details of the properties of JFK is presented in [6]. Here we summarize the main points as below:

1. The JFK is non-negative and its integral is bounded by the distribution’s variance (similar to the PDF);
2. The JFK is additive so sum of sources yields sum of JFK;
3. The JFK has sparse representation. Particularly this maps the Gaussian distribution and Poisson distributions to delta functions located at zeros and one, respectively.

The details of the JFK estimation can also be found in [6]. Here we summarize the main points of the estimation method as follows:

1. The JFK is estimated through empirical characteristic function implementing (10) by FFT.
2. The empirical characteristic function is estimated directly from data samples (not from the histogram)

\[ \hat{f}(u) = \frac{1}{N} \sum_{k=1}^{N} e^{iux_k}, \]  

(11)

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

3. The estimation (11) is unbiased and consistent, i.e. the mean of the estimation errors is 0 and its variance is proportional to \( 1/N \).

4. If the distribution is symmetric then second order derivative in JFK estimation (10) can be simplified by

\[ \frac{\partial^2 \log(f)}{\partial u^2} = -\sum_{k=1}^{N} x_k^2 \cos(u x_k) - \sum_{k=1}^{N} \frac{\sum_{k=1}^{N} x_k \sin(u x_k)}{\sum_{k=1}^{N} \cos(u x_k)} \]  

(12)

Fig.1 shows an example of the JFK estimations of clean speech, babble noise and the superimposed noisy speech, respectively. The estimation was carried out in a wavelet domain, particularly the Mel-filter with Gabor-shape located at 2kHz [7].

3. Wavelet-JFK based Speaker identification

In this section we introduce our speaker identification system whose block digram is illustrated in Fig.3. The speech signal is filtered by a set of subband filters. The output wavelets are normalized before the JFKs are being estimated in each subband.
The confidence intervals are set in clean training phase to protect the subband JFKs from the noise. In the testing phase, the estimated subband JFK is projected to each confidence interval. Finally, we adopt SVM to classify the JFKs’ feature vector which is formed by concatenating the projections of subband JFKs. Below we describe each block in more details.

3.1. Wavelet filters
Similar to our previous system [6], we adopt a non-orthogonal wavelet filter system of 24 zero-phase Mel-Gabor filters which are implemented in FFT domain. The center of each filter is located at the same Mel-frequencies as for the MFCC but the filter shape is Gabor’s function. The Gabor-shape filters are better localized in frequency domain and yield better statistical independence of the filtered signal and therefore is more suitable in our JFK-driven approach.

3.2. Mean-Variance Normalization
The filtered signals in each wavelet subband is normalized to have zero-mean and unit variance. This normalization means to remove the power variability caused by recording and to put the JFKs in the same scale in order to correctly classify them.

3.3. Subband JFK estimation
The subband JFK is estimated in each subband after normalization. This reasonable to assume that the wavelet has zero-mean symmetric distribution in each subband and therefore the JFK can be estimated following algorithm described in section 3 (formulas (12) and (10)).

3.4. Confidence interval
These intervals are defined as horizontal projections of the 0.7 of the maximum level of the estimated subband JFKs.

\[
[x_{\text{min}}^{(i)}, x_{\text{max}}^{(i)}] = \arg \left\{ k(x) > 0.7 \max_{x} k^{(i)}(x) \right\}.
\]  

(13)

The interval \([x_{\text{min}}, x_{\text{max}}]\) is set over classes as follows

\[
x_{\text{min}} = \min_{i} \left\{ x_{\text{min}}^{(i)} \right\},
\]

(14)

\[
x_{\text{max}} = \max_{i} \left\{ x_{\text{max}}^{(i)} \right\}.
\]

(15)

In each subband, the interval is stepped by 100 points to form the feature vector.

3.5. Project into SVM feature
For each sample, the subband JFKs are projected into confidence intervals and then concatenated to form a JFK-driven feature vector. The dimension of the feature vector is the multiplication of the number of subbands and the resolution in each subband, in our case this is 2400 (24x100). A feature vector is extracted from each speech utterance. In our experiment, there will be 96 utterance in training and 40 utterance in testing.

3.6. SVM classifier
The multi-class linear support vector machine (SVM) is adopted in our system. The one-against-one decomposition is used in the SVM classification [7].

4. Experiments
4.1. Database
We test the method using YOHO database [10]. YOHO speech database contains telephone handset records from 138 speakers under noiseless office environment and sampled at 8kHz. We use all the phrases in the “ENROLL” folder to train the system and all the phrases in the “VERIFY” folder to test. The speech amount is about 6 minutes (96 utterances) for training and 40 utterances of 2.4 seconds for test. To test the method under mismatch noise conditions, the test data is superimposed with noises under different SNR conditions of 10dB, 5dB and
speech-unlike noises have better non-overlapping supports in down from stationary (car) to more non-stationary (train) to improvements are seen in all the noise types but it slightly slows 13 conditions. The absolute improvments are about 5 conditions but significantly outpeformed it in the miss-match noise 14 much underpeformed than the MFCC-GMM in ideal clean con-
ditions. The wavelet-JFK is not of methods are shown in Fig.5-7 for the car, train and exhibition Figure 6: Identification under miss-match train noise environ-
ment

0dB segmental SNRs. The noises are the car noise, the train noise and the exhibition (babble) noise, taken from NOISEX92 database.

We note that in this work we are just focusing in the noise-mismatch problem ans leave for the future the discussion on the possibility of applying JFK for the channel variability and speaker verification on frame of NIST evaluation.

4.2. Reference methods

As a reference we implement the conventional MFCC-GMM method recommended in [2] (evaluated in a different database). Particularly, the 36-dimensional MFCC and 64-mixture GMM models (trained with k-mean initialization and EM algorithm) are adopted in our evaluation. To better compare the PDF to the JFK we also implement the wavelet-GMM method (i.e. GMM classifier with the same wavelets). Finally, to compare the present system to our previous system presented in [6], where the simple Euclidean distance of subband JFKs were used in classification, we also implement this version named wavelet-JFK (0). We note that, the comparision between MFCC-GMM and MFCC-JFK is impossible as the MFCC does not sastify the additivity which is required for the JFK-driven approaches.

4.3. Results and discussion

To evaluate the effect of noise mismatch conditions, the training is carried out in clean conditions while the test is perormed in different noises and SNR conditions. The identification rates of methods are shown in Fig.5-7 for the car, train and exhibition noises, respectively. It can be seen that, the wavelet-JFK is not much underperformed than the MFCC-GMM in ideal clean conditions but significantly outperformed it in the miss-match noise conditions. The absolute improvements are about 5% at 10dB, 13% at 5dB and 22% at 0dB SNR conditions, respectively. The improvements are seen in all the noise types but it slightly slows down from stationary (car) to more non-stationary (train) to more speech noise (exhibition). It can be explained that the speech-unlike noises have better non-overlapping supports in

5. Conclusion

We develop a novel speaker identification method based on Jump Function Kolmogorov analysis in the wavalet domain. This stochastic instrument can better separate the signal and noise in its representations and the driven speaker identification system has shown to be more robust than the conventional one in noise-mismatch conditions.

6. References