Kolmogorov-Smirnov Test in Text-Dependent Automatic Speaker Identification

Sangeeta Biswas, Shamim Ahmad, M. Khademul Islam Molla, Keikichi Hirose and Mohammed Nasser

Abstract—**In this paper Kolmogorov–Smirnov test, a non-parametric statistical test, is introduced for text-dependent automatic speaker identification (ASI) with Mel-frequency cepstral coeffients (MFCCs) based speech features. In the case of closed-set ASI, the identity (Id) of the unknown speaker is assigned to the Id of that reference speaker to whom the number of MFCCs pair having same distributions is maxima with 88.26% accuracy at 8% level of significance. In open-set ASI, after determining the identity of the unknown speaker it is verified whether the unknown speaker is truly the reference speaker or not by comparing the number of matched MFCCs pair to the threshold value previously set for that reference speaker with 87.24% identification efficiency at 2% level of significance and false speaker detection rate is 99.5% at 10% level of significance.**

Index Terms—**Alpha level of significance, Kolmogorov–Smirnov test, Mel-frequency cepstral coefficient (MFCC), text-dependent speaker identification.**

I. Introduction

Automatic speaker identification (ASI) or speaker identification by machines is a behavioral biometric technique for finding out the identity of a person by using the speaker specific characteristics included in his or her speech waves. Starting at 1960[1] till now it is drawing growing research interests for its non-invasive, inexpensive practical use such as efficient banking and business transactions, controlled access of a facility or information to selected individuals and lots more.

In an ASI system both the methods used for classification and feature extraction are very important. All classification approaches employed formerly or presently in the ASI have two phases – (i) training or learning phase and (ii) testing or decision making phase. In the training phase each registered speaker has to provide samples of their speech so that the system can build or train a reference model for that speaker by using the features extracted from the speeches. During the testing phase, the features of input speech are matched with stored reference model(s) and recognition decision is made. When a speaker is bound to utter the same speech in the both phases, then the ASI is called Text-dependent; otherwise it is called text-independent. In the learning phase of all classification models, as the number of speaker increases, the number of computations increases. In this paper we have introduced Kolmogorov–Smirnov test in text-dependent ASI by which in learning phase no extra computation is done to build the reference models for the registered speakers after feature extraction. In learning phase, the proposed method is faster than any other classification methods employed formerly or presently in the ASI.

Proposed by Davis and Mermelstein in 1980 [2], the Mel-frequency cepstral coeffients (MFCCs) have consistently been shown to outperform other feature representations for clean speech. Observing the high intra-speaker similarity and low inter-speaker similarity between the MFCCs of a fixed speech, we have decided that if it is possible to prove the null hypothesis that the MFCCs of two utterances have the same distributions for a unique speaker but different distributions for two different speakers at α level of significance by using any two sample test in univariate case, then it would be possible to identify a speaker. Hence we have applied the K-S test yielding impressive results.

II. Kolmogorov-Smirnov Test

Kolmogorov–Smirnov test (K-S test) is a widely used non-parametric statistical test. It was developed in the 1930s by Andrei Nokolaevich Kolmogorov and Nikolai Vasilyevich Smirnov. Although it is mainly used as a one-sample test where it allows the comparison of the frequency distribution of a sample to some known distribution, such as a Gaussian distribution, it can also be used as a two-sample test. As a two-sample test K-S test compares the distributions of values in the two data vectors X_1 and X_2 of length n_1 and n_2, respectively. The null hypothesis for this test is that X_1 and X_2 has the same continuous distribution. The alternative hypothesis is that they have different continuous distributions. Mathematically the test statistic can be written as-

$$KS\_statistic = \max(|F_1(x) - F_2(x)|) \tag{1}$$

where $F_1(x)$ is the proportion of X_1 values less than or equal to x and $F_2(x)$ is the proportion of X_2 values less than or equal to x. When $p=P(KS\_ststistic \geq \alpha)$, that means the p-value is greater than or equal to the level of significance, α we can consider that X_1 and X_2 have the same distributions, otherwise the null hypothesis has to be rejected. The asymptotic p-value becomes very accurate for large sample size, and is believed to be reasonably accurate for sample sizes n_1 and n_2 such that $\{(n_1 \times n_2)/(n_1+n_2)\} \geq 4$. For example, let us consider three data sets, as given in Table 1. At first $F_1(x)$ and $F_2(x)$, the cumulative distributions of X_1 and X_2 respectively are

Manuscript received June 18, 2008.

S. Biswas, S. Ahmad, M. K. I. Molla and M. Nasser are with the University of Rajshahi, Rajshahi, Bangladesh.

M. K. I. Molla and K. Hirose are with the Dept. of Information and Communication Engineering, The University of Tokyo, Tokyo, Japan.

Corresponding author: M. K. I. Molla; Tel: +81-35841-6767, Fax: +81-35841-6664, email: molla@gavo.t.u-tokyo.ac.jp.

calculated and then their absolute difference is found out and *p*-value is calculated. In the first experiment (experiment on first and second data sets), the test statistic (maximum absolute difference) is 0.4000 and *p*-value is 0.6974 and in the second experiment (experiment on first and third data sets), the test statistic is 1.0000 and *p*-value is 0.0038. So, at 5% level of significance in the first case we can accept the null hypothesis that X_1 and X_2 have the same distributions whereas in the second case we can reject the null hypothesis that means X_1 and X_2 have different probability distributions.

Table 1: *Example of Kolmogorov–Smirnov test*

Ex. No	X_1	X_2	$F_1(x)$	$F_2(x)$	$\|F_1(x)-F_2(x)\|$
1	3.529	4.170	0	0	0
	3.744	4.071	0	0.200	0.200
	3.004	2.850	0.200	0.200	0
	2.233	0.213	0.200	0.400	0.200
	-0.016	-1.115	0.400	0.400	0
			0.400	0.600	0.200
			0.600	0.600	0
			0.800	0.600	0.200
			1.000	0.600	0.400
			1.000	0.800	0.200
			1.000	1.000	0
2		6.967	0	0	0
		7.360	0.200	0	0.200
		7.464	0.400	0	0.400
		6.033	0.600	0	0.600
		3.950	0.800	0	0.800
			1.000	0	1.000
			1.000	0.200	0.800
			1.000	0.400	0.600
			1.000	0.600	0.400
			1.000	0.800	0.200
			1.000	1.000	0

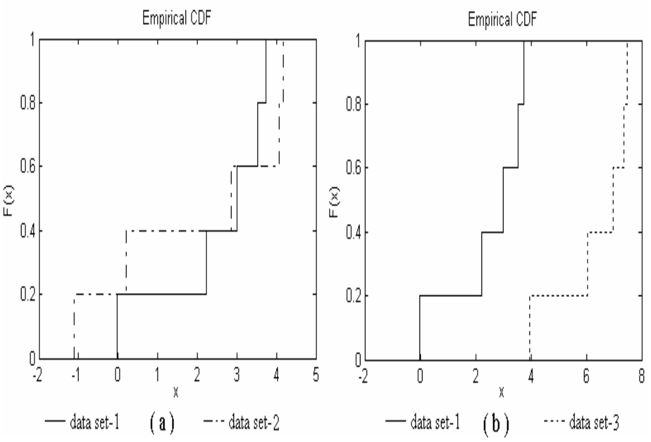


Figure 1: *Empirical CDF; (a) for data set-1 and data set-2, and (b) for data set-1 and data set-3.*

The original references for K-S test appear in [3, 4], whereas simple description is found in [5]. To visualize the difference of cumulative distributions of X_1 and X_2, the empirical cumulative distributions of the data sets are shown in Figure1. From the Figures 1(a) and 1(b), it is observed that the maximum difference occurs appears to be near $x = 3.8$ in the both case. In the first case, the maximum difference is 0.4000 and in the second case, the maximum difference is 1.000.

III. Speaker Identification with K-S Test

After plotting MFCCs extracted from the speech signals, it is found that there is a high intra-speaker similarity and low inter-speaker similarity between the MFCCs of the speech signals corresponding to the Bengali word 'Protijogeeta' as shown in Figure 2.

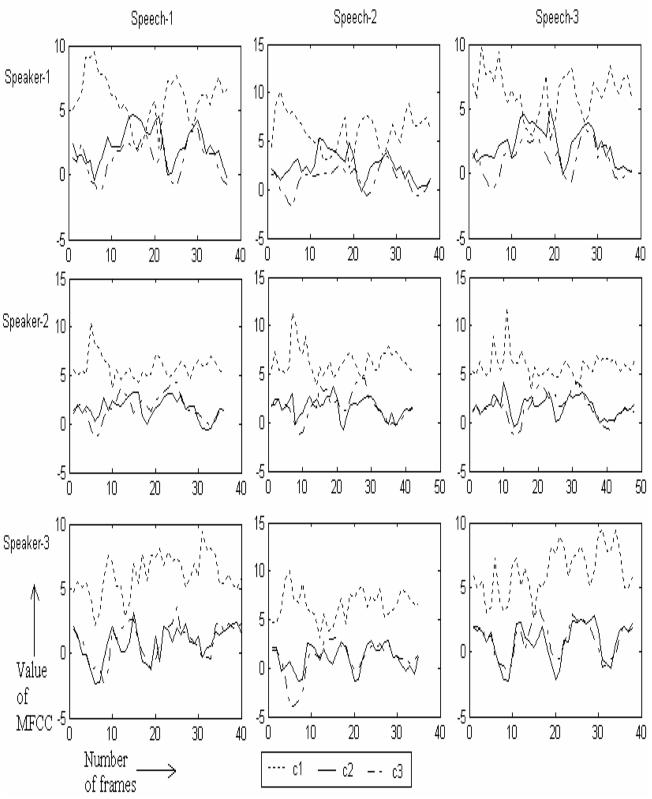


Figure 2: *Three lower MFCCs(c_1-c_3) of three utterances of sample speech ("Protijogeeta") taken from three speakers. Notice MFCCs have high intra-speaker similarity and low inter-speaker similarity.*

We can assume that the underlying probability distributions are same for the MFCCs when they are extracted from the speech signals uttered by the same speaker and different for the MFCCs when they are extracted from the speech uttered by different speakers. To prove the hypothesis we have used K-S test. The proposed algorithm for ASI using K-S Test is shown in Figure 3.

The algorithm shown in Figure 3 is applicable both for open-set and close-set ASI. In the case of close-set ASI, at first n MFCCs are calculated for the utterances of the m reference speakers and unknown speaker. Then it is determined whether the j^{th} MFCC of the unknown speaker and the i^{th} reference speaker follow the same probability distribution or not. The total number of matched MFCCs pair is counted and the Id of the unknown speaker is assigned to the Id of that reference speaker to whom the number of matched MFCCs pair is maxima. The open-set ASI is slight different from the close-set ASI. In this case, after determining the identity of the unknown speaker it is verified whether the unknown speaker is truly the reference speaker or not by comparing the number of matched MFCCs pair to the threshold value previously set for that reference speaker.

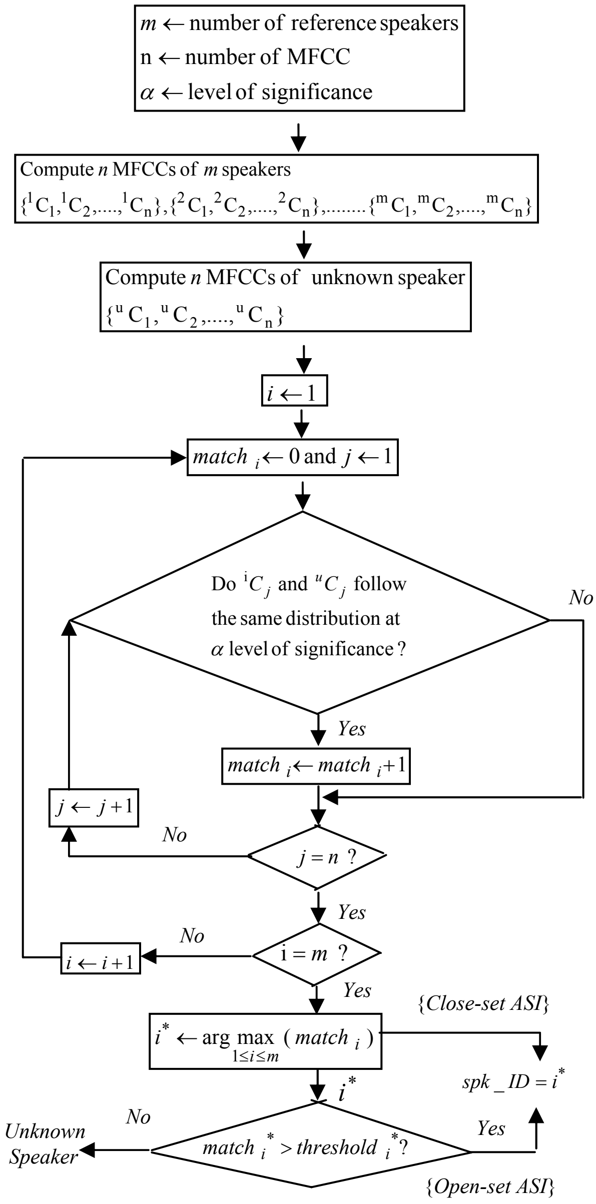
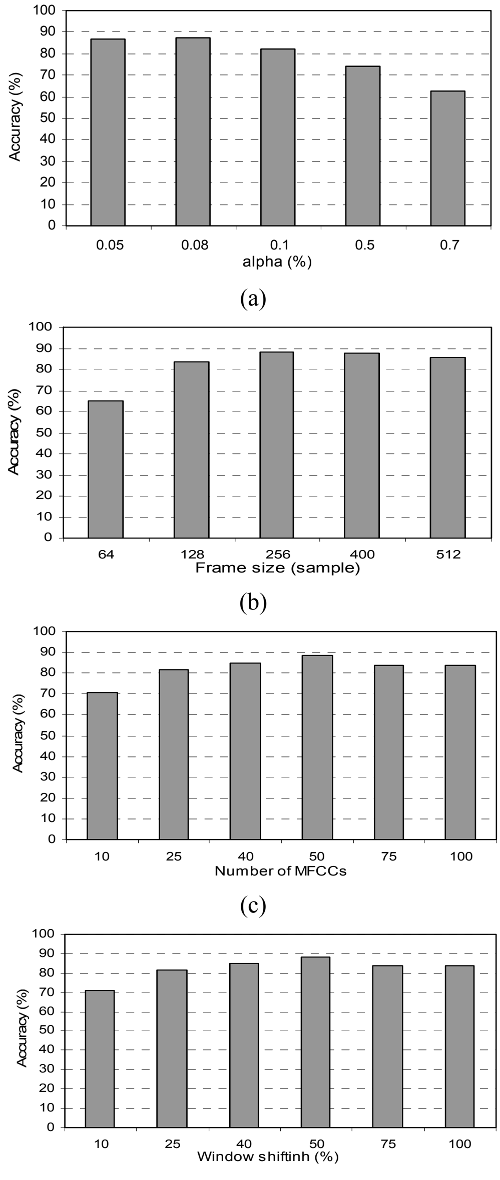

$m \leftarrow$ number of reference speakers
$n \leftarrow$ number of MFCC
$\alpha \leftarrow$ level of significance
Compute n MFCCs of m speakers
$\{^1C_1, ^1C_2, ..., ^1C_n\}, \{^2C_1, ^2C_2, ..., ^2C_n\}, ... \{^mC_1, ^mC_2, ..., ^mC_n\}$
Compute n MFCCs of unknown speaker
$\{^uC_1, ^uC_2, ..., ^uC_n\}$
$i \leftarrow 1$
$match_i \leftarrow 0$ and $j \leftarrow 1$
Do ^iC_j and ^uC_j follow the same distribution at α level of significance ?
No
Yes
$match_i \leftarrow match_i + 1$
$j \leftarrow j+1$
No
$j = n$?
Yes
No
$i \leftarrow i+1$
$i = m$?
Yes
{Close-set ASI}
$i^* \leftarrow \arg\max_{1 \le i \le m} (match_i)$
i^*
$spk\_ID = i^*$
No
Unknown Speaker
$match_{i^*} > threshold_{i^*}$?
Yes
{Open-set ASI}

Figure 3: *Flowchart of the proposed speaker identification algorithm using Kolmogorov–Smirnov test*

For text-dependent close-set ASI it is found that among 16 MFCCs pair above 14 MFCCs pair is matched if the unknown speaker is truly the reference speaker. For example, let us consider that among 16 MFCCs pair of unknown speaker and 3 reference speakers, the numbers of matched MFCCs pair are 13, 10, and 11 respectively. For close-set ASI, the unknown speaker Id will be the Id of first reference speaker. If the threshold value for the first reference speaker is set to 14 then in open-set ASI the unknown speaker will not get any valid speaker identity.

IV. Experimental Results

For text-dependent ASI system, a pre-decided Bengali word, 'Protijogeeta' (means Competition) is recorded with 56 speakers (41 males and 15 females). Each speaker uttered the same word 7 times in one recording session. All utterances of speakers were recorded under the normal room environment with 11025 Hz sampling rate, 16-bits quantization level using single microphone. The value of alpha, the number of MFCCs, the frame size, the value of window shifting – all of these factors affect the identification accuracy.



(a)

(b)

(c)

(d)

Figure 4: *Speaker identification accuracy in Close-set text-dependent case when (a) 16 MFCCs were extracted from 256 samples or about 23 ms sized frame and widow is shifted 50%, (b) MFCCs were extracted from 256 samples sized frame shifting window 50% and α =0.08; (c) 18 MFCCs were extracted shifting window 50% and using α =0.08, and (d) 18 MFCCs were extracted from 256 samples sized frame and α =0.08*

In close-set ASI, when 16 MFCCs were extracted from 256 samples or about 23 ms sized frame and widow is shifted 50%, then at 1% level of significance the accurate speaker identification rate is 82.14% and as the value of alpha increased, the identification accuracy increased till 8% level of significance and after that identification accuracy is decreased as shown in Figure 4(a).

In this case the highest speaker identification accuracy is 87.5%. When MFCCs are extracted with frame size 256 samples by shifting the window 50%, with 8% level of significance the identification accuracy is increased till the number of MFCCs is 18. The accuracy is decreased as number of MFCCs is increased as shown in Fig.4 (b). In this case the highest identification accuracy is 88.26%. As shown in Figure 4(c), the identification efficiency is above 80% when the frame size is above 10ms. At 8% level of significance, the highest speaker identification is 88.26% when 18 MFCCs are extracted by using the frame of 256 samples length with 50% shifting. Keeping frame size fixed at 256 samples and the number of MFCCs at 18, the highest speaker identification accuracy is found when window is shifted 50% as shown in Figure 4(d).

In open-set ASI, a false acceptance occurs when the system incorrectly identifies an unregistered individual as an enrolled one. When one registered individual is mistaken for another and a false rejection occurs when the system incorrectly refuses to identify an individual who is registered with the system. The false acceptance ratio (FAR) can be reduced by setting a strict (low) threshold and the value of α. The false rejection ratio (FRR) can be minimized by setting the threshold to a reliably high value as well as the α.

The requirements for low FAR and FRR are observed to be conflicting as shown in Figure 5. The both parameters cannot be simultaneously lowered. The highest true speaker identification rate is 87.24% at 2% level of significance and false speaker detection rate is 99.5% at 10% level of significance. Also 15 speakers are selected from TIMIT database to test the performance of the proposed method and 78% identification accuracy is obtained.

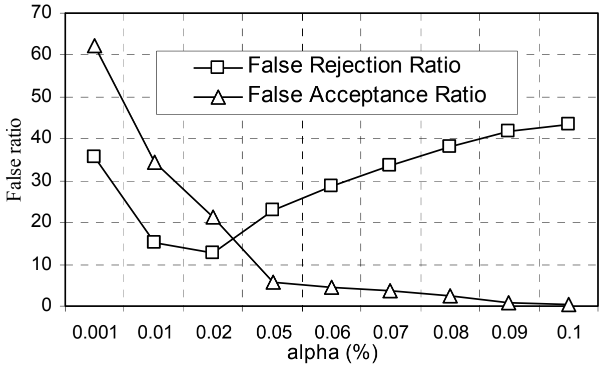


Figure 5: *FRR and FAR with Open-set text-dependent case. The lowest FRR is 12.75% for α =0.02 and the lowest FAR is 0.5% when α =0.1.*

V. Conclusions

In this paper, we have introduced for the first time a widely used non-parametric statistical test named Kolmogorov-Smirnov test in text-dependent automatic speaker identification. Using MFCC as speech feature we have found above 87% identification accuracy in text-dependent close-set and open-set ASI. In the proposed method we have neglected multivariate relationship between MFCC of individual speaker. It is evident that the two sample test in multivariate case improves the performance of the algorithm. The proposed algorithm is also applicable for the ASI when real cepstral coefficient (RCC), linear predictive cepstral coefficient (LPCC) and other cepstral cofficients are used as the features of speech. The future plan with this method is to increase the noise robustness for the use in real world applications.

References

[1] Furui, S., "50 Years of Progress in Speech and Speaker Recognition Research Identification", In ECTI Transformations on Computer and Information Technology, vol. 1, no. 2, 2003

[2] Davis, S. B. and Mermelstein, P., "Comparison of Parametric Representations for Mono-Syllabic Word Recognition in Continuously Spoken Utterances", IEEE Trans. Acoustics, Speech and Signal Processing, vol. 28, No. 4, pp. 357-366, 1980

[3] Kolmogorov, A. N., "On the Empirical Determination of a Distribution Function", vol.4, pp.83-91, 1933.

[4] Smirnov, N. V., "On the Estimation of the Discrepency between Empirical Curves of Distribution for Two Independent Samples", Bulletin of Moscow, vol.2, pp.3-16, 1939.

[5] Rohatgi, V. K., "An Introduction to Probability Theory and Mathematical Statistics", Willey Eastern Limited, New Delhi, Reprint 5, pp.557-559, 1976.