ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 3, Special Issue 7, October 2015

Prosodic Features and Decision Trees in an Inter-session Speaker Verification System

Athulya Jayakumar, Dr.Babu Anto P

Research Scholar, School of Information Science and Technology, Kannur University, Kannur, Kerala, India

Associate Professor, School of Information Science and Technology, Kannur University, Kannur, Kerala, India

ABSTRACT: Automatic Speaker Recognition system is one of the finest recognition systems in today's world. This paper presents a study carried out in inter-session speaker recognition using prosodic features and decision tree classifiers for a biometric system. The performance of pitch, energy and formants with different decision trees on a customer-adaptive database created for banking purpose is evaluated. Word corpus with a total of 2200 speech samples is obtained from two sessions. Results investigate that using pitch and J48 decision tree is a reliable and robust method for inter-session speaker recognition.

KEYWORDS: Speaker Recognition; Inter-session Speaker Recognition; Prosody; Pitch; Energy; Formants; Decision Trees; J48; Best first tree; Simple cart

1. INTRODUCTION

Biometric security system is a security system which is based on the biometric features. Identification of a person is a very traditional problem. Various tools and techniques have been used as the biometric features like finger print recognition, face recognition, signature recognition for identification of people. Speaker recognition is another such type of recognition based on the biometric features (Anupama et al. 2008). Speaker recognition technology is as such expected to build new services to make our daily lives more convenient. It is a process of identifying a person on account of speech processing. It can be more specifically described as the use of a machine to identify a person from a spoken phrase. Such a task is mostly challenging because, a person's voice can change strongly, depending on a number of factors like state of health, emotional state, familiarity with interlocutors. Many attempts have been made on focusing at such human ability for applications such as customer verification for bank transactions, access to bank accounts through telephones, control on the use of credit cards, and for security purposes in the army, navy and airforce. Speaker identity is associated with the physiological and behavioral characteristics of the speaker.

Speaker recognition uses the acoustic features of speech which is found to be different between individuals. These acoustic patterns reflect size and shape of the throat and mouth, voice pitch, and speaking style. Speaker recognition can be classified in Speaker Identification and Speaker Verification. In Speaker Identification a speech utterance from an unknown speaker is analyzed and matched up with models of known speaker. The unknown speaker is recognized as the speaker whose model go well with the input utterance. In Speaker Verification an identity claim is made by an unknown speaker and an utterance of the unknown speaker is compared with the model for the speaker whose identity is claimed. If the match is above a certain threshold, the identity claim is verified (Kajarekar et al. 2008). Speaker recognition systems can be text dependent and text independent systems. In text-dependent speaker recognition, it is assumed that the speaker is cooperative, to be recognized. This is most often used in security applications where a person may identify themselves using their voice to access sensitive information. Where as a text-independent system is typically used in surveillance and forensic as it has no constraints about what the speaker is saying (Kajarekar et al. 2008). Speaker recognition is done mainly through two process feature extraction and Classification. Majority of the research activities are focused on some conventional transform techniques like FFT, MFCC, LPC and STFT etc. As the magnitude of the short-time spectrum encodes information about vocal tract shape of the speaker most of the current speaker recognition systems rely on the spectral features derived through short-time spectral analysis of the speech signal. Therefore, spectral features are widely used for speaker modeling. This paper uses prosodic features like

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 3, Special Issue 7, October 2015

Pitch, Energy and Formants for speech feature extraction and decision tree classifiers like J48, Best first tree and Simple cart for classification.

II. DATABASE

A Customer-Adaptive Speaker Verification database for banking purpose is created for the study. Each speech wave was recorded with Logitech clearchat comfort USB microphone using Goldwave software. For a database access application in banking, as a means of imposing a level of security for transaction, various means of customer authentication like account number, Personal Identification Number (PIN) and security questions must be used. Considering the level of security, each speaker in the database uttered the words my, voice, is, password, for, recognition followed by customer's name, account number, password and two security questions. A specific pattern was given for all speakers to choose their account numbers. It consists of nine digits like 670 264 560. The password for each customer also followed a pattern with any three alphabets and three digits. This was chosen according to customer's interest. Security questions provided was to select a weekday and color, according to customer interest. Database created consists of two sessions (DB I & II) as our aim was to access the session variation. The second session was recorded with the same speakers using the same database in the same environment. As the paper focus on short-term session variation, the second database was collected with a gap of three weeks from the first session. Each database consists of 25 (15 female and 10 male) native Malayalam language speakers. All the speakers were in the range of 20-40 age groups. Uttered speech samples were recorded at different sampling rates such as 8 KHz, 11 KHz, 16 KHz and 22 KHz for investigating the variations occurring in sampling rates. The word corpus used in this paper consist of 1100 speech samples in each session and thus, a total of 2200 speech samples combining both sessions.

III. PROSODIC FEATURES

Prosody is the study of aspects of speech that typically applied to a level above that of the individual phoneme and very often to sequences of words. Features above the level of the phoneme are referred to as suprasegmentals. Prosody is known to play an important role in human speech perception process. Prosodic features are known to be less influenced from channel distortion and noise in a speaker recognition system. Therefore, prosodic features demand more for the advancement of speech and speaker recognition technology.

A. Fundamental frequency and Formants

The voicing process is generally a contribution from the opening and closing of the vocal folds, and the frequency of this pattern is known as pitch or fundamental frequency. The F0, (fundamental frequency) of speech signal is a widely used non-linguistic speech feature which can be directly identified by human observers as it is well audible. F0 is one of the main factors which can discriminate the speaker's sex. Typical values of F0 for male speech are 110 Hz, 210 Hz for female speech and 300-500 Hz for children. Values of F0 between 20 to 70 years aged people lie between 80-170 Hz for men and 150-260 Hz for women (Leena and Yegnanarayana 2008). According to signal theory, pitch is the lowest frequency in a harmonic series representing periodic parts of a speech signal. In normal speech pitch changes constantly, providing linguistic information, as in the different intonation patterns associated with the speech data. Besides, the pitch pattern determines naturalness of utterance production. This paper makes use of autocorrelation for pitch extraction. The autocorrelation is a correlation of a variable with itself over time. The autocorrelation computation is made directly on the waveform and is a quite easy computation (smitha et al. 2013).

For a discrete time signal x(n), defined for all n, the autocorrelation function is generally defined as

$$\varphi_{x(m)} = \lim_{N\to\infty} \frac{1}{2N+1} \sum_{n=-N}^{N} x(n)x(n+m) \qquad (1)$$

Formant features are widely accepted features used in forensic acoustic-phonetic speaker verification. These features can be associated directly to the resonance cavities in the vocal tract (Timo et.al, 2008). Formants are nothing but the spectral peaks of the sound spectrum of the voice. In speech science and phonetics, formant frequencies are an acoustic resonance of the human vocal tract which is measured as an amplitude peak in the frequency spectrum of the sound. In this paper Periodogram using FFT is used for the extraction of first and second formants (F1 and F2).

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 3, Special Issue 7, October 2015

B. Energy

One of the simplest representations of a signal is its energy. Energy associated with voiced region of a speech is large compared to unvoiced region and silence region. As energy strongly depends on the sensitivity of the human auditory system to different frequencies, it is the acoustic correlation of loudness and their relation is not linear. The sensation of loudness dependent on both the frequency of the sound and on the duration, and also, pitch perception depends on the loudness (Arputha et al. 2012). A common way to calculate the energy of a speech signal is the root mean square energy (RMSE), and is used in this work. RMSE is the square root of the average sum of the squares of the amplitude of the signal samples.

$$En_{(RMS)} = \sqrt{\frac{1}{N}} \sum_{m=0}^{N-1} [w(m)x(n-m)]^2 \quad (2)$$

Where N is the number of samples in the window *w(m)*. *w(m)* generally tends to zero monotonically as *m* gets larger.

IV. CLASSIFIERS

ANN is a type of information processing network whose architecture is inspired by the structure of biological neural system. Knowledge is acquired by the network through a learning process. They are easy to understand and modify. Decision tree learning is one of the most popular technique in classification because it is fast and produces models with sensible performance. This paper compares three decision tree algorithms J4.8, BF tree and Simple cart.

A. J48, BF tree and Simple cart

J48 is an open source Java implementation of the C4.5 algorithm. J48 adopt greedy approach in which decision trees are constructed in a top-down recursive divide and conquer manner. J48 can handle both continuous and discrete attributes. In order to classify, J48 first needs to create a decision tree based on the attribute values of the available training data. So, whenever J48 encounters a set of training set it identifies the attribute which could discriminate the various instances most evidently. This feature is said to have the highest information gain. Now, among the possible values of this feature, if there is any value for which there is no ambiguity, we can terminate that branch and assign the target value that we have obtained. In other case, we can choose another attribute that gives us highest information gain. This is continued until we either get a clear decision of what combination of attributes provides a particular target value, or we run out of attributes. Thus, if we cannot get an unambiguous result from the available information, we assign this branch a target value that the majority of the items under this branch possess (Padhye 2005). The impurity measures for nominal dependent variables are entropy-based definition of information gain and Gini index. J48 style using Gini index looks for the largest class in the training list and strives to isolate it from all other classes. It produces good results for a large variety of classification problems and is thus the default rule used for J48. Entropy characterizes the purity of any sample set and is calculated by,

$$Entropy\ S = -p_j log_2 p_j \quad (3)$$

Where p_j is the proportion of *S* belonging to class *j*

Gini index of diversity minimizes the risk involved when making predictions once having made the test, using [Cernak, 2010]

$$Gini\ Index = 1 - \sum_j p_j^2 \quad (4)$$

The best-first decision tree is a learning algorithm for supervised classification learning. The problem in growing best-first decision trees is how to determine which attribute to split on and how to split the data. The important objective of decision trees is to seek accurate and small models (Haijian 2006). BF tree constructs binary trees, i.e., each internal node has exactly two outgoing edges. The tree growing method attempts to maximize within-node homogeneity. The extent to which a node does not represent a homogenous subset of cases is an indication of impurity.

The term Classification and Regression Tree (CART) analysis is an umbrella term used to refer both Classification tree analysis and Regression tree analysis, first introduced by Breiman et.al, in 1984. (Yohannes and Webb, 1999). It is built in accordance with splitting rule which performs the splitting of learning sample into smaller parts. Regression trees do not have classes. Splitting in regression trees is done with squared residuals minimization algorithm which indicates that expected sum variances for two resulting nodes should be minimized (Roman 2004). The measure of

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 3, Special Issue 7, October 2015

effectiveness called the information gain is the expected reduction in entropy caused by splitting the training data. Gain (S, A) of an attribute A is defined as,

$$Gain\ (S, A) = Entropy\ S - \sum_{v \in Values(A)} \frac{|S_v|}{|S|} Entropy(S_v) \quad (5)$$

where Values *(A)* is the set of all possible values for attribute A and S_v is the subset of S for which attribute A has value υ such that $S_v = \{s \in S | A(s) = v\}$ (Cernak 2010).

V. EXPERIMENTS AND RESULTS

In order to attain the performance of prosodic features and decision trees, experiments were carried out in two sessions. In each session 70% of samples were taken for training the data and 30% to test the data. A frame length of 10-25 milliseconds is used as the window should be long enough for sufficient frequency resolution, and also, short enough to capture the local spectral properties. A feature vector of nine was obtained for each prosodic feature. The recognition rates and confusion matrix obtained by using J48, BF tree and simple cart decision trees in two sessions are given in table 1 to 3. Classification using decision trees were carried out with the extracted features using Weka software.

Table.1. Prosodic feature and J48

		8	11	16	22	Average
session I	F1	60	63	58	57	59.5
	F2	68.66	61.33	64	63.33	64.33
	Pitch	**76**	72	69	74	72.75
	Energy	68	70	67	62	66.75
session II	F1	61	60	60	62	60.75
	F2	70	68	68.66	68.44	68.78
	Pitch	**73**	67	**73**	72	71.25
	Energy	62	61	70	70	65.75

Table.2. Confusion matrix obtained for 76% recognition

```
=== Confusion Matrix ===

 a  b  c  d  e  f  g  h  i  j  k  l  m  n  o  p  q  r  s  t  u  v  w  x  y   <-- classified as
 4  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0 |  a = 1
 0  3  0  0  0  0  0  0  0  0  0  0  0  1  0  0  0  0  0  0  0  0  0  0  0 |  b = 2
 0  0  4  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0 |  c = 3
 0  0  0  4  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0 |  d = 4
 0  0  0  0  3  0  0  0  0  1  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0 |  e = 5
 0  0  0  0  0  3  0  0  0  0  0  0  0  0  0  0  0  0  0  1  0  0  0  0  0 |  f = 6
 0  0  0  0  0  0  4  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0 |  g = 7
 0  1  0  0  0  0  0  2  0  0  0  0  0  0  0  0  0  0  1  0  0  0  0  0  0 |  h = 8
 0  0  0  0  1  0  0  0  3  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0 |  i = 9
 0  0  0  0  0  0  0  0  0  4  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0 |  j = 10
 0  0  0  0  0  0  1  0  0  0  1  0  0  0  0  1  0  0  0  0  0  0  0  0  1 |  k = 11
 0  0  0  0  0  0  0  0  0  0  0  4  0  0  0  0  0  0  0  0  0  0  0  0  0 |  l = 12
 0  0  0  0  0  0  0  0  1  0  0  0  2  0  1  0  0  0  0  0  0  0  0  0  0 |  m = 13
 0  0  0  0  0  0  0  0  0  0  0  0  0  4  0  0  0  0  0  0  0  0  0  0  0 |  n = 14
 0  0  0  0  0  0  0  0  1  0  0  0  0  0  3  0  0  0  0  0  0  0  0  0  0 |  o = 15
 0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  2  0  0  1  0  0  1  0  0  0 |  p = 16
 0  0  0  0  1  0  0  1  0  0  0  0  0  0  0  0  2  0  0  0  0  0  0  0  0 |  q = 17
 0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  4  0  0  0  0  0  0  0 |  r = 18
 1  0  0  0  0  0  0  0  0  0  0  0  0  1  0  0  0  0  2  0  0  0  0  0  0 |  s = 19
 0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  4  0  0  0  0  0 |  t = 20
 0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  4  0  0  0  0 |  u = 21
 0  0  0  0  0  0  0  1  0  0  0  0  0  1  0  0  0  0  0  0  1  0  1  0  0 |  v = 22
 1  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  3  0  0  0 |  w = 23
 0  0  0  1  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  1  2  0  0 |  x = 24
 0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  4  0 |  y = 25
```

Table.3. Prosodic feature with Simple Cart Tree and Best First tree

	Prosodic feature and Simple Cart Tree							Prosodic feature and Best First Tree					
Session I		8	11	16	22	Average	Session I		8	11	16	22	Average
	F1	66	57	54	57	58.5		F1	68	56	52	58	58.5
	F2	66.66	61.33	54	51.33	58.33		F2	66	60	65.33	60.66	63
	Pitch	68	54	59	**70**	62.75		Pitch	**71**	69	59	70	67.25

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 3, Special Issue 7, October 2015

	Energy	60	62	58	61	60.25		Energy	61.6	65	58	59	60.9
Session II	F1	52	66	48	44	52.5	Session II	F1	56	62	52	60	57.5
	F2	56.66	63.33	56.66	64.66	60.33		F2	66	64	58.66	65.33	63.5
	Pitch	64	51	61	**68**	61		Pitch	**70**	61	**70**	67	67
	Energy	64	62	66	60	65.5		Energy	59	64	65	62	62.5

The obtained results suggest pitch feature is more effective and reliable for speaker recognition. It is also noted that J48 performed better among the three decision trees. Pitch and J48 at session I provided maximum recognition f 76%.

Inter-session Speaker Recognition using Prosody features and decision trees As this paper focus on inter-session speaker verification, experiment were carried out by taking session I as enrollment phase and session II as verification phase. The system was trained by using speech samples in session I as training and samples in session II for testing. Results obtained are given in table 4.

Table.4. Prosody features with J48 and BF tree on inter-session

	Prosody features with J48					Prosody features with BF Tree					Prosody features with Simple CART				
	8	11	16	22	Average	8	11	16	22	Average	8	11	16	22	Average
F1	57.5	57.5	51	53.5	54.88	51	56	50.5	48	51.36	41	53	51.5	32.5	44.5
F2	58	59.5	64	59	60.12	56	55	52.5	55	54.63	56.5	53	51.5	56	54.25
Pitch	66.5	67	64.5	**68.5**	66.63	59	58.5	58	**63.5**	59.75	**65.5**	42	40	54	50.38
Energy	61.5	62.5	62.5	59	61.38	56	54.5	55	59.5	56.25	50	48.5	41.5	40	45

Inter-session speaker verification resulted with a maximum of 68.5% using pitch and J48 decision tree. On analyzing the results it clear that maximum session variation was of 3%.

VI. CONCLUSION

In this paper different prosodic features and decision trees are used in a customer adaptive speaker verification database. Results suggest that an accurate and robust estimation of pitch plays a central role in speaker and inter-session speaker recognition. Experiments also put forward the efficiency of J48 decision tree classifiers.

REFERENCES

1. Anupama Shukla, Ritu Tiware, Hemanth Kumar, Meena and Rahul Kala, "Speaker Identification Using Wavelet Analysis and Artificial Neural Networks", *Proceedings of the National Symposium on Acoustics (NSA)* ,2008
2. Arputha X Rathina, K. M. Mehata, M. Ponnavaikko , "A study of prosodic features of emotional speech", *Advances in Intelligent and Soft Computing* Vol.166, pp 41-49, 2012.
3. Cernak Milos, "A Comparison Of Decision Tree Classifiers For Automatic Diagnosis Of Speech Recognition Errors", *Computing and Informatics*, Vol. 29, pp.489–501, 2010.
4. Haijian Shi, "Best-first Decision Tree Learning", 2006
5. S. S. Kajarekar, L. Ferrer, A. Stolcke, & E.Shriberg"Voice-Based Speaker Recognition Combining Acoustic and Stylistic Features," N. K.Ratha & V. Govindaraju (eds.), Advances in Biometrics: Sensors, Algorithms and Systems, pp. 183-201, Springer,London. (2008),
6. Leena Mary a,*, B. Yegnanarayana, "Extraction and representation of prosodic features for language and speaker recognition", *Speech Communication* 50, 782–796, 2008.
7. Padhye http://www.d.umn.edu/~padhy005/Chapter5.html
8. Smita S. Hande, Dr. Milind S. Shah, "Pitch Estimation" , *International Conference on Computational & Network Technologies* , 2013.
9. Roman Timofeev, "Classification and Regression Trees (CART) Theory and Applications", *Master Thesis*, 2004.
10. Timo Becker, Michael Jessen, Catalin Grigoras, "Forensic Speaker Verification Using Formant Features and Gaussian Mixture Models", *Interspeech*, *Forensic speaker recognition-Traditional and automatic approaches*, 2008.
11. Yohannes Yisehac, Patrick Webb, "Classification and Regression Trees, CART: A User Manual for Identifying Indicators of Vulnerability to Famine and Chronic Food Insecurity", *Intl Food Policy Res Inst*, pp. 4, 1999

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 3, Special Issue 7, October 2015

BIOGRAPHY

Athulya Jayakumar a research Scholar works in the area of Speech and Signal Processing at School of Information Science and Technology, Kannur University, Kerala State, India under the Guidance of Dr. Babu Anto P. She has received MSc. Degree in Physics from Kerala University, Kerala, India. Her research interest includes Artificial Intelligence, Signal Processing, and Pattern recognition. She has authored more than ten research papers in International Conferences and journals

Babu Anto P is working as Associate Professor, in School of Information Science and Technology, Kannur University, India**.** He has received his Master of Science degree from Cochin University of Science and Technology, India in 1982 and he has awarded with his Doctoral Degree in 1992 by Cochin University. He has a number of international journals and conference papers in his credit. He is guiding doctoral students for the past years. His main research interest lies in Speech processing, Pattern Recognition, Data mining and visual Cryptography. He has published more than 65 journals and conference publications.