SPEAKER CLASSIFICATION USING SVM

[1]SANJAY S. BADHE, [2]SURESH D. SHIRBAHADURKAR, [3]AJAY N. PAITHANE

[1]D. Y. Patil Institute of Technology, D. Y. Patil College of Engineering, Pune, India
[2,3]Zeal College of Engineering, Rajarshi Shahu College of Engineering, Pune, India
E-mail: [1]sanjubadhe@gmail.com, [2]s_shir00@yahoo.co.in, [3]ajaypaithane@gmail.com

Abstract - In this paper speakers belongs to various geographical regions are classified using Support Vector Machine (SVM). Indic TTS database is used for experimentation. Five thousand audio samples of speakers belongs to various geographical regions are used. Spectral feature of these speech audio signals are extracted and speakers are classified by using SVM. Accuracy is calculated.

Keywords - Spectral Features, SVM

I. INTRODUCTION

In India various languages are spoken by speakers. If language is native then it is easy to identify speaker's geographical region but if language is common i.e. English then it is very difficult to identify geographical region. Some researchers classify speakers using accent. The accuracy of previous research is less and they worked on less database. Here, Indic TTS database is used for experimentations. Database of various geographical regions is used for experimentations. Five thousand audio samples of speakers belongs to various geographical regions are used. Various spectral features are extracted from these samples. Statistical features are also calculated. Spectral features (Timbral features) like Spectral Roll off, Spectral Centroid, RMS energy, Zero crossing Rate, Spectral irregularity, Brightness are extracted. After extraction of features speakers are classified using SVM and accuracy is calculated. For experimentations neuro solution software is used. 60 % samples are used for training, 25 % samples are used for testing and 15 % samples are used for cross validation.

II. THEORY

A. Block Diagram

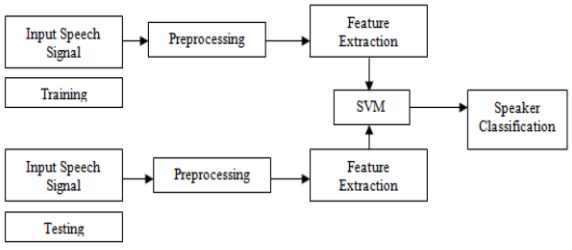


Fig.1 Block diagram

Following figure shows the block diagram of feature extraction. Input is audio signals used from database. Features are extracted from input audio signals. Extracted features are classified. Training is completed by using 60 % samples and testing is completed by using 25 % samples. Support Vector Machine (SVM) is used for classification.

III. EXPERIMENTATIONS AND RESULTS

For experimentations five thousand samples are used. Features of these samples are extracted. Mean Squared Error (MSE), Normalized Mean Squared Error (NMSE), Mean Absolute Error (MAE), Minimum absolute error, maximum absolute error and Accuracy are calculated. These are calculated using following formulae:

A. Mean Squared Error (MSE) is mean of the squares of errors. In statistical modeling the MSE can represent the difference between the actual and predicted observations by model.

$$MSE = 1/N \sum_{i=1}^{n} (Yi-\hat{Yi})^2$$

B. Normalized Mean Squared Error (NMSE) is an estimator of the overall deviations between predicted and measured values.

$$NMSE = MSE/\sigma^2$$

$$Variance\ (\sigma^2) = 1/N \sum_{i=1}^{n} (Xi-\mu)^2$$

C. Mean Absolute Error (MAE) is the average of absolute errors.

$$MAE = 1/N \sum_{i=1}^{n} | yi-xi |$$

D. Accuracy is Percentage of output classified correctly, calculated using confusion matrix.

Accuracy = (TP+TN) / (TP+TN+FP+FN)
Where TP = True Positive
TN = True Negative
FP = False Positive
FN = False Negative

E. Training:

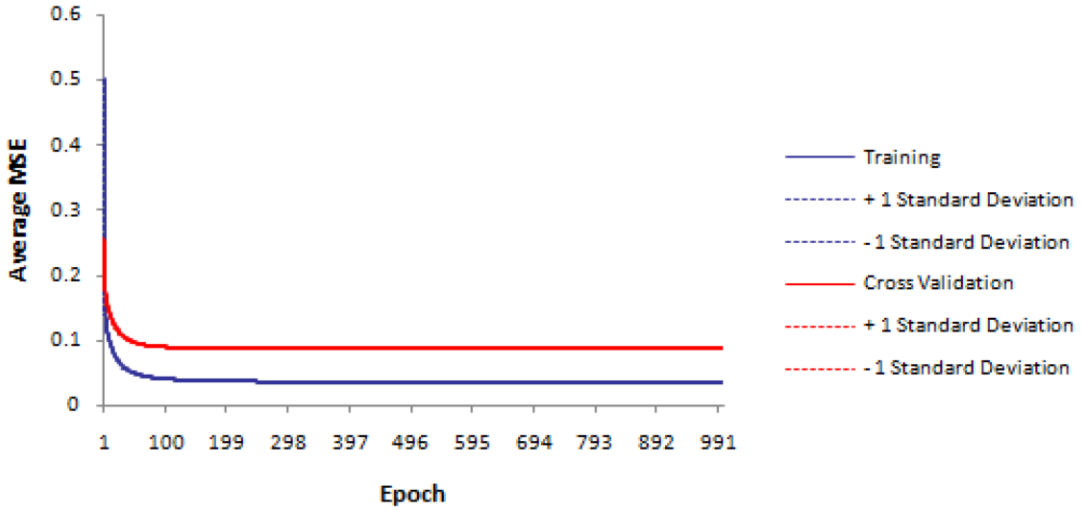


Fig. 2 Epoch Vs Average MSE

1000 epochs used for training. Average MSE is calculated. Standard deviation for training and cross validation is calculated. Figure 2 shows the graph of epoch vs average MSE. Training accuracy is calculated by using confusion matrix as following:

Output / Desired	o/p Gujarati	o/p Bengali	o/p Kannada	o/p Aasam	o/p Malayalam
o/p Guajarati	579	0	0	0	0
o/p Bengali	0	597	0	0	0
o/p kannada	0	0	604	0	0
o/p Aasam	0	0	0	603	0
o/p Malayalam	0	0	0	0	617

Table I. Confusion matrix for training accuracy

Mean Squared Error (MSE), Normalized Mean Squared Error (NMSE), Mean Absolute Error (MAE), Minimum absolute error, maximum absolute error and Accuracy are calculated for Training, Testing and Cross validation.

Performance	o/p(Gujarat)	o/p(Bengali)	o/p(Kannada)	o/p(Aasam)	o/p (Malayalam)	Average Training o/p
MSE	0.008511243	0.002945116	0.00288755	0.00222732	0.006828296	0.004679905
NMSE	0.054646475	0.018476377	0.017957597	0.013868827	0.04179705	0.029349265
MAE	0.073950625	0.046682592	0.045941956	0.043066614	0.056467411	0.05322184
Min Abs Error	9.62157E-05	7.22415E-06	0.000207188	2.93598E-06	8.78696E-05	8.02867E-05
Max Abs Error	0.478065866	0.325826066	0.327434456	0.120071739	0.471205714	0.344520768
r	0.981731466	0.995647322	0.995505683	0.997466457	0.982698781	0.990609942
% Accuracy	100	100	100	100	100	100

Table II. Training Performance Parameters

Table I shows the confusion matrix for training. Accuracy is calculated from it. Same confusion matrix is used for testing and cross validation. Table II shows training performance parameters like MSE, NMSE, MAE, Minimum absolute error, maximum absolute error, relative coefficient and percentage accuracy. All these parameters are calculated for each geographical region and in last column average of all these parameters are calculated.

Performance	o/p(Gujarat)	o/p(Bengali)	o/p(Kannada)	o/p(Aasam)	o/p (Malayalam)	Average Testing o/p
MSE	0.0424181	0.0196697	0.01968446	0.01066622	0.038184318	0.026124578
NMSE	0.2532183	0.1194014	0.12339859	0.07033213	0.244584635	0.162187036
MAE	0.1366822	0.0945888	0.09368268	0.07836112	0.116178928	0.103898768
Min Abs Error	0.0007552	6.3911E-05	0.00017693	0.00024687	0.001252545	0.000499104
Max Abs Error	1.0474864	0.9467923	0.95297119	0.46294945	1.046461848	0.89133225
r	0.8750083	0.9499960	0.94790026	0.97714072	0.876367437	0.92528258
% Accuracy	87.97	95.77	97.99	100	90.50	94.45

Table III. Testing Performance Parameters

Performance	o/p(Gujarat)	o/p(Bengali)	o/p(Kannada)	o/p(Aasam)	o/p (Malayalam)	Average Testing o/p
MSE	0.047325759	0.025302713	0.025039702	0.013361057	0.043160998	0.030838046
NMSE	0.288649924	0.163970186	0.158897487	0.078202724	0.282733714	0.194490807
MAE	0.1477614	0.104192335	0.102889835	0.084446101	0.127572652	0.113372465
Min Abs Error	0.000232609	0.000437764	0.000477077	0.000448478	0.000135748	0.000346335
Max Abs Error	1.012872735	0.994705213	0.977325682	0.527760155	1.006554957	0.903843748
r	0.853432416	0.926432808	0.928644669	0.972150572	0.858689608	0.907870015
% Accuracy	82.58	93.00	97.28	100	92.91	93.15

Table IV. Cross Validation Performance Parameters

Table III shows testing performance parameters MSE, NMSE, MAE, Minimum absolute error, maximum absolute error, relative coefficient and percentage accuracy. All these parameters are calculated for each geographical region and in last column average of all these parameters are calculated. Table IV shows cross validation performance parameters MSE, NMSE, MAE, Minimum absolute error, maximum absolute error, relative coefficient and percentage accuracy. All these parameters are calculated for each geographical region and in last column average of all these parameters are calculated.

IV. CONCLUSION

It is observed from table III that the accuracy of Aasam is 100 %, Gujarat is 82.58 %, Bengal 97.28 %, and Kannada 97.99 % and average accuracy is 94.45 %. It is observed from table VI that the accuracy of Aasam is 100 %, Gujarat is 87.97 %, Bengal 93 %, Kannada 97.928 % and average accuracy is 93.15 %.

REFERENCES

[1] Sadaoki,"50 Years of Progress in Speech and Speaker Recognition Research", ECTI Transaction on Computer and Information Technology, 2005.

[2] K. H. Davis, et. al. "Automatic recognition of spoken digits," J.A.S.A., 24 (6), pp. 637-642, 1952.

[3] Rekha Hibare & Anup Vibhute "Speaker Identification System for Hindi and Marathi Languages using Wavelet and Support Vector Machine" International Journal on Recent and Innovation Trends in Computing and Communication, April 2015.

[4] Yeshwant K. Muthusamy, Neena Join and Ronald A. Cole "Perceptual Benchmarks for Automatic Language Identification" International Conference on Acoustics, Speech and Signal Processing, Apr. 1994.

[5] Mayur R Gamit, Prof. Kinnal Dhameliya, Dr. Ninad S. Bhatt "Classification Techniques for Speech Recognition: A Review" International Journal of Emerging Technology and Advanced Engineering, February 2015.

[6] Shi-Xiong Zhang and Mark J. F. Gales "Structured SVMs for Automatic Speech Recognition" IEEE Transactions, March 2013.

[7] Chofu, Japan, Yamazaki,Yoshiura, H. , " Speaker verification method for operation system of consumer electronic devices" IEEE Transaction, Feb 2015.

[8] Ossama Abdel-Hamid, Abdel-rahman Mohamed, Hui Jiang, Li Deng, Gerald Penn, and Dong Yu "Convolutional Neural Networks for Speech Recognition" IEEE/ACM Transactions, October 2014.

[9] Khaled T. Assalehand Richard J. Mammone, "New LP-Derived Features for Speaker Identification", IEEE Transactions on Speech And Audio Processing, Vol. 2, No. 4, October 1994

[10] LIU Wai Kat and Pascale FUNG," Fast Accent Identification And Accented Speech Recognition", 1999 IEEE.

[11] Khalid Saeed, Mohammad Kheir Nammous, "A Speech-and-Speaker Identification System: Feature Extraction, Description, and Classification of Speech-Signal Image", IEEE Transactions on Industrial Electronics, Vol. 54, No. 2, April 2007.

[12] Marco Grimaldi and Fred Cummins, "Speaker Identification Using Instantaneous Frequencies", IEEE Transactions on Audio, Speech and Language Processing, Vol. 16, No. 6, August 2008.

[13] Mohamed Kamal Omar and Jason Pelecanos,"A Novel Approach To Detecting Non-native Speakers and Their Native Languag", 978-1-4244-4296-6/10/$25.00 ©2010 IEEE.

[14] Bozhao Tan, Qi (Peter) Li, Robert Foresta, "An Automatic Non-Native Speaker Recognition System", 2010 IEEE.
[15] Qi Li, Senior and Yan Huang, "An Auditory-Based Feature Extraction Algorithm for Robust Speaker Identification Under Mismatched Conditions", IEEE Transactions On Audio, Speech, And Language Processing, Vol. 19, No. 6, August 2011.
[16] Suma Swamy , K.V Ramakrishnan, "An Efficient Speech Recognition System", Computer Science & Engineering: An International Journal (CSEIJ), Vol. 3, No. 4, August 2013.
[17] Saurabh Bhardwaj, Smriti Srivastava, Madasu Hanmandlu and J. R. P. Gupta," GFM-Based Methods for Speaker Identification", IEEE Transactions On Cybernetics, Vol. 43, No. 3, June 2013.
[18] Sriram Ganapathy,Sri Harish Mallidi,and Hynek Hermansky, "Robust Feature Extraction Using ModulationFiltering of Autoregressive Models", IEEE/ACM Transactions on Audio, Speech, And Language Processing, Vol. 22, No. 8, August 2014.
[19] Kalinli, O., Seltzer, M.L., Droppo, J. and Acero, A., "Noise Adaptive Training for Robust Automatic Speech Recognition", IEEE Transactions on Audio, Speech, and Language Processing, 2014.
[20] Weibin Zhang and Fung, P., "Discriminatively Trained Sparse Inverse Covariance Matrices for Speech Recognition", IEEE/ACM Transactions on Audio, Speech, and Language Processing, 2014.
[21] Shilin Liu and Khe Chai Sim, "Temporally Varying Weight Regression: A Semi-Parametric Trajectory Model for Automatic Speech Recognition", IEEE/ACM Transactions on Audio, Speech, and Language Processing,2014.
[22] Rahim Saeidi, Paavo Alku, Tom Bäckström, "Feature Extraction Using Power-Law Adjusted Linear Prediction With Application to Speaker Recognition Under Severe Vocal Effort Mismatch", 2015.
[23] Yow-Bang Wang and Lin-shan Lee, "Supervised Detection and Unsupervised Discovery of Pronunciation Error Patterns for Computer-Assisted Language Learning", IEEE/ACM Transactions on Audio, Speech, and Language Processing (Volume: 23, Issue: 3, March 2015).
[24] J. Sirisha Devi, Dr. Siva Prasad Nandyala and Dr. P. Vijaya Bhaskar Reddy, "Speaker Verification with Optimized Feature Subset using MOBA",978-1-5090-1623-5/16/$31.00 ©2016 IEEE.
[25] Huy Phan, Lars Hertel, Marco Maass, Radoslaw Mazur Alfred Mertins, "Learning Representations for Nonspeech Audio Events Through Their Similarities to Speech Patterns", IEEE/ACM Transactions on Audio, Speech, and Language Processing (Volume:24, Issue: 4, April 2016).
[26] Giorgio Biagetti, Paolo Crippa, Laura Falaschetti, Simone Orcioni and Claudio Turchetti, "An Investigation on the Accuracy of Truncated DKLT Representation for Speaker Identification With Short Sequences of Speech Frames", IEEE Transactions on Cybernetics, 2168-2267 _C 2016 Ieee, 2016.
[27] Suman Deb, Samarendra Dandapat, Jarek Krajewski, "Analysis and Classification of Cold Speech using Variational Mode Decomposition", IEEE Transactions on Affective Computing (Volume: PP, Issue: 99), October 2017.
[28] Igor Bisio, Chiara Garibotto, Aldo Grattarola, Fabio Lavagetto and Andrea Sciarrone, "Smart and Robust Speaker Recognition for Context-Aware In-Vehicle Applications", IEEE Transactions on Vehicular Technology, 2018 (Accepted).
[29] C. R. Salamea , L. F. D'Haro, R. Córdoba, " Language Recognition Using Neural Phone Embeddings and RNNLMs", IEEE Latin America Transactions, Vol. 16, No. 7, July 2018.
[30] Qian Zhang and John H. L. Hansen, "Language/Dialect Recognition Based on Unsupervised Deep Learning", IEEE/ACM Transactions on Audio, Speech, and Language Processing, Vol. 26, No. 5, May 2018.
[31] Ma Jin, Yan Song, Ian McLoughlin and Li-Rong Dai, "LID-Senones and Their Statistics for Language Identification", IEEE/ACM Transactions on Audio, Speech, and Language Processing, Vol. 26, No. 1, January 2018.
[32] Zhiyuan Tang, Dong Wang, Yixiang Chen, Lantian Li, and Andrew Abel, "Phonetic Temporal Neural Model for Language Identification", IEEE/ACM Transactions on Audio, Speech, and Language Processing, Vol. 26, No. 1, January 2018.

★★★