Revista Argentina de Clínica Psicológica
2021, Vol. XXX, N°2, 866-880
DOI: 10.24205/03276716.2020.4093

Acoustic-Visual based Accent Identification System using Deep Neural Networks

N. Radha[a], Sachin Madhavan R M[b]

Abstract

Human-machine interfaces are evolving rapidly. Identifying the accent of a speaker can improve the performance of speech recognition systems. Although foreign accent identification is extensively explored and this paper aims to build a robust accent identification for Tamil language using acoustic and visual features. The proposed system which is first automatically recognize the speaker's accent among regional Tamil accents from three different regions of Tamil Nadu. This system is built using acoustic mel cepstral features and visual optical flow motion features, which are classified as being either local by Lucas-Kanade method, and global by Horn-Schunck technique. These proposed features are trained using a sequential model in an artificial and convolution neural network, which allows for the detection and classification of accents. Second, this system uses visual color features and cepstral features to recognize accented speakers. The speaker recognition module trained with Hidden Markov Model. The Tamil accent system performance achieves 93.7%, 89.5%, and 96% acoustically, visemically, and the combined one respectively. The recognition rate of 93.1% for Nellai and Chennai accent whereas for Nellai and Kovai Accents, the accuracy was 94%. The multi features based accented speaker recognition system achieves better recognition rate of 96.7% rate compared to the individual feature-based system feature performance.

Keywords: *Automatic speech recognition, Artificial neural networks, Mel frequency cepstral coefficient, Optical flow motion*

1. Introduction

Speech patterns are becoming more and more important in criminal investigations. Voice-activated devices have become a welcome addition to many Indian households. Artificial intelligence is used by a voice assistant to recognize and respond to spoken requests. The key advantage of voice assistants is their speech interface. The speech interface is a major advantage of voice assistants. This technology is accessible to humans all over the world. However, due to its inefficiency in handling different dialects found in a language, this technology is not being used as much as it might be. Automatic speech recognition systems perform better with American speech rather than accented speech. The problem lies in the fact that there are so many different languages with various types of dialects and accents. Classifying accents is a step to develop more intelligent virtual assistants. Hence, accent recognition has found applications in a variety of fields, including crime investigation and forensic speech analysis, as the usage of deep learning to tackle real-world issues has increased. Usually, the evidence involves recorded audio clips of 3-4 seconds. If accents can be identified accurately with these short audios, the ethnicity of the criminal can be found which may benefit the investigators.

In [1] have presented an English accent identification system which consists of three stages namely accent identification, accented speaker recognition and speech recognition which is modelled using ergodic Hidden Markov Model (HMM). Identification of accent using the approach of deriving spectral envelope from glottal excitation significance was proposed. The data were collected for the four different countries and the recognition accuracy was found to be average of 80% [2]. The CSLU Foreign accented English corpus used as a database for accent identification and uses the approach of Gaussian Mixture Model (GMM)/ supervised deep neural network were proposed [3]. This system yields improvements of 15.4% by i-vector features extraction. An automatic identification of speaker accent among Swiss French accents and uses two different GMM

[a] Assistant Professor, Department of Information Technology, SSN College of Engineering, Chennai. radhan@ssn.edu.in
[b] UG Scholar, Department of Information Technology, SSN College of Engineering, Chennai

algorithms, universal background modelling followed by map adaption, i-vector modelling performed. The speaker accent recognition system shows 15.3% improvements for the PFC database [4].

The identification of English accented speakers using the Q-factor method presented. This system uses Accent archive database for recognition and determines the English accents with high probability from the four different accents [5]. The classification of English accent speakers and the spoken words were split as monosyllabic and aligned using automatic segmentation system using Naive Bayes, Soft max logistics regression and GMM [6]. The recognition of accent using GMU English archive database using neural network and genetic algorithm yields better recognition accuracy [7]. The foreign accent identification of speakers of German and Mandarin using HMM for CSLU Foreign accent English database shown improved classification accuracy of 13.26% [8]. The automatic identification of Chinese speakers of major accent such as Mandarin GMM on basis of four different accent speakers and the recognition in reduction of error rate in both male and female utterances [9].

The classification of accent was constructed by MFCC features and the distinguish made between Indian and American English speakers using supervised Learning models, and derived receiver operating characteristics curve by using the database of CSTR VCTK corpus achieves 76% recognition rate [10]. The faster accent identification and recognition using phoneme-class models for the TIMIT corpus yields the 13.5% reduction in error rate was presented [11]. The Persian accent identification using Support Vector Machine (SVM) and K-Nearest Neighbor (KNN) was proposed. In SVM based system achieves best recognition accuracy of 90% which higher compared to KNN based Persian accent identification [12]. The accent identification of Malaysian English comprises of formant and log energy feature vectors using the KNN approach for the database of Malaysian English speech corpus and achieves the recognition rate of 94.2% [13].

The Persian accents identification using the articulatory features using GMM-Background Model and i-vector and the Fars-dat dataset used with the recognition of 75.29% [14]. The Palestinian accent recognition using GMM-UBM, GMM-SVM and I-vector framework classifiers proposed for the data collected from the four-target region around 300 speakers show the recognition accuracy of 81.5% [15]. From the survey, the accent identification system using Tamil language have not been reported in the literature. Hence this study focuses on building Robust Tamil accent identification system in a multimodal context. In this work, proposes the acoustic and visual speech information and the visual information not affected by noises which ensures the robustness.

Visual speech recognition system is an area with great potential to solve challenging problems in speech processing. Visual optical flow motion measures the spatial temporal variations present in the visemes [16]. Visual features were extracted using active shape model and active appearance model were proposed and classified using multi-classifier [17,18]. A multimodal approach using acoustic and visual features for Hindi syllabic unit recognition is proposed with highest recognition rate of 96% [19]. In [20] proposed a bi-modal approach which uses the multisource information combines the acoustic-visual features with the classification gives best recognition rate. Visual lip movement-based speech recognition using motion features [21] and combined motion and geometric information based visual system was built and achieves higher recognition accuracy [22]. In [23] deep learning technique which uses transform encoding method ResNet and Fourier transform for visual and acoustic features respectively. The motion-based features capture well define dynamic information of visual lip movement compared static features such as shape and model-based features. This study explores the use of visual motion information using motion estimation algorithms for accent identification.

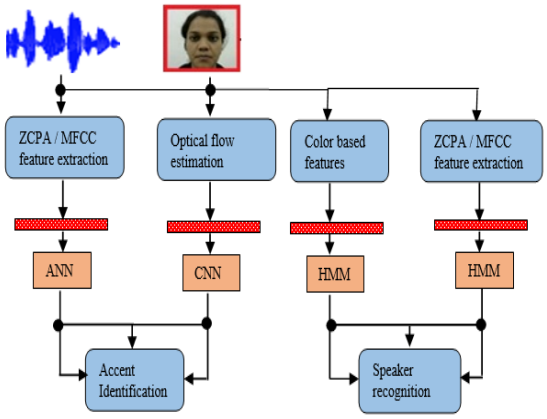


Figure 1. **Block diagram of the proposed accent identification system**

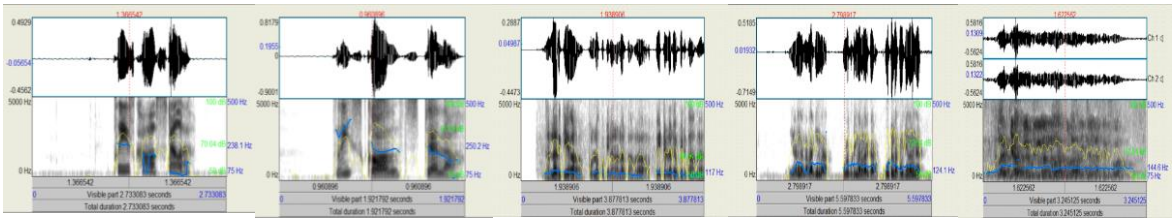
This paper aims to develop a deep learning models for accurate accent identification and speaker recognition system. The model is tested to classify different accents of two languages, English and Tamil. Figure 1 shows the proposed method involves analysis of audio and visual samples and

extracting required features for accent identification and to recognize the speaker. The features are considered and the respective frames are concatenated to form the input feature set. This input feature set is then split into train set (80%) and test set (20%). If the current data is split into training and testing sets, one of the classes would be underrepresented. To work around this system, the dataset is split into testing and training and the training set is oversampled on the underrepresented class. After oversampling, the ratio between the two outcomes in the training set is 1:1. The input set is now fed into the classifier and the result of the classifiers are then validated and compared using different accuracy metrics.

The remainder of this paper is organized as follows: Section 2 presents the analysis of acoustic measures of audio such as pitch, intensity, spectrogram and formant frequency. Section 3 describes the framework of our proposed model, the dataset details, along with the deep learning model used. Section 4 shows the results of the system. Section 5 concludes the work.

2. Acoustic-Visual Analysis of Accent Identification

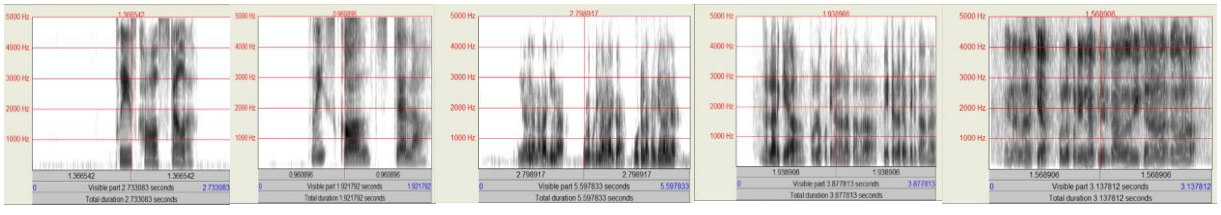
Acoustic analysis of Tamil accent samples is represented with the distinct analysis of waveform, spectrogram, intensity and formant frequency. The acoustic samples are collected from different persons of different regions and vary from 3-7 seconds. This features analysis is representing the difference between the speakers of different regions i.e., English accent (American E_A and Indian E_I) and Indian-Tamil accent (Chennai I_c, Nellai I_n and Kovai I_k) and the samples are analyzed and identified. Figure 2 shows the waveform representation of acoustic sample example "Please call Stella" for E_A, E_I and "Nalla irukiya? Pullakutti ellam nalla irukangala? Veetula ellarum epdi irukanga?" for I_c, I_n, I_k used for different speakers which captures the pitch and intensity information. The observation shown that I_c intensity variation is similar to the I_n. It is noted that I_k intensity variation is too high when compared to the I_n and I_c. The resonant frequency F2 was high for I_k which was shown as very less as in F0 and F1 with I_c and I_k respectively.



a b c d e

Figure 2. **Waveform of (a) American (b) Indian (c) Chennai (d) Kovai (e) Nellai accent**

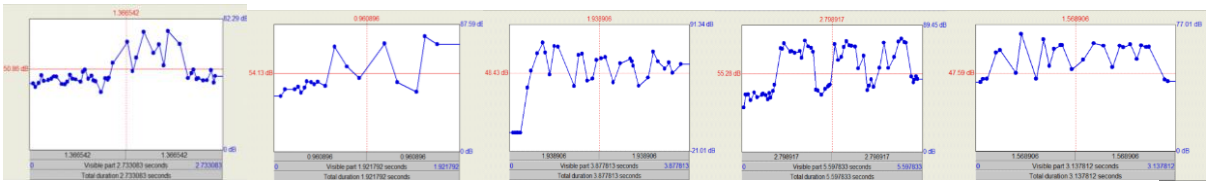
Figure 3 represents the spectrogram images of audio samples of different speakers. From the acoustic samples the E_A information captured well in the range between frequency of 1000Hz to 3500Hz and the Indian accent formants are seen lowered at F0 level when compared to E_A. An accent E_I has a long duration when compared to E_A. In the Tamil accent, I_n has a long duration of audio and I_k has some pause between words and I_c has the same long duration of words.



a b c d e

Figure 3. **Spectrogram of (a) American (b) Indian (c) Chennai (d) Kovai (e) Nellai accent**

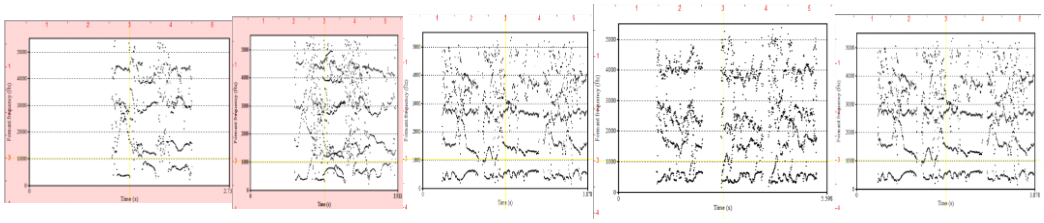
Figure 4 shows the intensity of variation of different acoustic samples of different accents. In the English accent, the variation of intensity in E_I has greater in peaks which never get ended in lower when compared to intensity of E_A. In the Tamil accent, the variation of intensity of I_k has a very short peaks and having higher variation when compared to the I_c and I_n. The I_c accent intensity peaks are high and lower in variation and this observation is similar to the I_n.



a b c d e

Figure 4. **Intensity plot of (a) American (b) Indian (c) Chennai (d) Kovai (e) Nellai accent**

The formant frequency of acoustic samples of different accents are shown in figure 5. In the English accent, both E_A, E_I accent has clear formants but there is some separation as well as the formants are seen as short in E_A when compared to E_I. In the Tamil accent, all of the three have clear formants, the resonant frequencies of I_k is depicted in a lower manner with respect to formants frequency of I_n and I_c accent. Formants representation clearly state that similarities are higher in position when comparing I_c and I_n.



a b c d e

Figure 5. **Formants plot of (a) American (b) Indian (c) Chennai (d) Kovai (e) Nellai accent**

a b c

Figure 6. **Visual lip motion of Indian Accent of (a) Chennai (b) Kovai (c) Nellai**

The optical motion flow refers to the distribution of apparent velocities of brightness pattern motion in an image. The motion features can be extracted without prior information about the order of the input data. Therefore, visual motion features might be evaluated without extracting contours and lip locations using optical flow analysis and the motion variants are presented in figure 6. This analysis gives a greater advantage over conventional lip localization method to evaluate visual features together with extracting lip locations and contours helps to build speech and accent identification and recognition.

3. Proposed Methodology

This section discusses the databases used for accent identification and speaker recognition system. Feature extraction method such as mel frequency cepstral and zero crossing peak amplitude used for acoustic information and visual optical flow motion field using sparse based technique such as Lucas-kanade method and Horn-Schunck algorithm is discussed in detail. The classification techniques used for modelling such as HMM-GMM, and neural network type convolution and artificial neural architecture is explained.

3.1 Dataset

For English language, the speech dataset is a collection of acoustic files from VCTK-corpus with two classes of accents, American English with 1269 acoustic samples accent and Indian English accent with 1032 acoustic samples used. This dataset consists of .wav acoustic samples which are 3 – 10 seconds long, and therefore including both the classes the total number of samples taken for the English language is 2301. The Tamil speech SSN accent database consist of simultaneous recording of acoustic and visemes of data. This corpus was built by manually collecting data from 25 speakers with various Tamil dialects, including Chennai, Nellai, and Kovai, as well as 72 samples for each Tamil accent. Therefore, including all the three classes the total number of acoustic samples taken

for the Tamil language is 216*2 = 432 samples used accent identification system. For the Tamil language, the dataset is a collection of .wav audio samples which are 5 – 20 seconds long. Acoustic and visemes of speakers with Tamil accent of Chennai, Nellai and Kovai accent are recorded uttering the same content. These acoustic samples are collected with Philps acoustic headset with the sampling frequency of 16000Hz and visual data is collected using Sony Handy cam camera with frame rate of 60fps.

3.2 Acoustic and Visual feature extraction

For this proposed methodology, two different feature extraction methods are proposed for accent and speaker recognition. The first method involves testing and training using Mel-Frequency Cepstral Coefficient (MFCC) features, optical flow motion-based features, and a combination of features are used to develop an accent identification (λ_a) system. MFCC characteristics are mostly employed in the development of speech recognition systems. MFCC features are extracted from an audio $x(n)$ signal stream as follows: The spectrogram of spoken speech segments of $x(n)$ shows that the lower frequency band has more energy than the higher frequency band. The pre-emphasis of $x(n)$, which amplified high frequency energies, resulting in $x'(n)$. After that, windowing with a certain frame size (10ms) and frame shift is done (15ms). When Hamming windowing w_h is applied to $x'(n)$, the result is $w_h(n) = 0.54 - 0.46\cos(2n\pi|N-1)$. The energy coefficients of $x'(n)$ are derived from the windowed signal. For each N discrete bands, signal $z = x'(n)$ is analyzed again using Discrete Fourier Transform. The magnitude and phase value of that frequency component in the $x'(n)$ is represented by the consequent complex value $Z(k)$, which is written as

$$Z(k) = \sum_{n=0}^{N-1} ze^{-j2\frac{\pi}{N}kn} \qquad (1)$$

On the spectrum $Z(k)$, a Mel filter bank with uniform spacing is applied. Mel is the pitch unit, and the Mel scale is approximately linear below 1 kHz and logarithmic above 1 kHz. The total of all filtered spectral components is the output of each filter. The dynamic range values in the spectrum components were compressed by log, which was used to compute spectral components. A cepstrum is represented by a spectrum of the log spectrum, which requires Fourier analysis, which is performed by applying inverse DFT to DCT. The MFCC coefficients were obtained as a result of this transform, which created substantially uncorrelated features.

$$M(k) = \sum_{n=0}^{N-1} \mathrm{x(n)}\, \mathrm{e}^{-\mathrm{j}\frac{2\pi}{N}\mathrm{kn}}\, F_c$$

$$= \sum_{n=0}^{N-1} M(k)\, \mathrm{e}^{\mathrm{j}\frac{2\pi}{N}\mathrm{kn}} \qquad (2)$$

$$D_{cn} = \frac{F_c(n+1) - F_c(n-1)}{2} \quad A_{cn} = \frac{D_c(n+1) - D_c(n-1)}{2} \qquad (3)$$

MFCCs, which are 39-dimensional (F_c, D_c, A_c) are used to represent acoustic speech signals. The 39 coefficients are made up of 13 static coefficients (F_c), 12 cepstral coefficients (D_c) plus energy (reflecting the range of vocal tracts), 13 differential coefficients, and 13 acceleration coefficients A_c (indicating spectral content change during phonetic transition and maybe providing phone identification clues).

$$Z_k = \frac{1}{2W_L}\sum_{n=1}^{W_L} |s[x_k(n)] - s[x_k(n-1)]|\ Where\ s[x_k(n)] = \begin{cases} -1, & x_k(n) < 0 \\ 1, & x_k(n) \geq 0 \end{cases} \qquad (4)$$

An acoustic frame Zero-Crossing Peak Amplitude (ZCPA) is the rate at which the signal sign changes during the frame, and it's the number of times the signal value changes from positive to negative and vice versa, divided by the frame length. The ZCPA feature is a spectral approximation that is directly generated from the signal in the temporal domain and can be used as a spectral shape descriptor. In this proposed work combined acoustic features for the system $\lambda_a = F = F_c + D_c + A_c + Z_k$ of MFCC and ZCPA were used features extraction with dimension of 49.

The optical flow motion based visual features are extracted from visual lip movements. To extract the features Lukas-Kanade and Horn-Shunk algorithm are used. The visual image $I(x(t), y(t), t)$ be the brightness at $p = (x(t), y(t))$ at time t with the flow field of u, v. Features dimension of four (Displacement: horizontal-vertical components, magnitude, and velocity) with the viseme size of 32*32 is used. In Lukas-Kanade method F_{lk}, motion flow is constant for all the pixel (sparse) presented in the visemes. This method-based feature computation as follows

$$I_t(p_i) + \nabla I(p_i).\begin{bmatrix} u \\ v \end{bmatrix} = 0\ F_{lk} = I_x(p_i) I_y(p_i) \begin{bmatrix} u \\ v \end{bmatrix} = -\sum_{i=1}^{i=32} I_t(p_i) \qquad (5)$$

The flow of motion with the four-dimensional vector is calculated from pixel to pixel in the Horn-Shunk method F_{hs}, and smooth flow motion is computed as follows:

$$F_{hs} = (\hat{u}_{nk}, \hat{v}_{nk}) I_x(p_i) I_y(p_i) \quad (6)$$

$$\hat{u}_{nk} = \bar{u}_{nk} - \frac{I_x \bar{u}_{nk} + y \bar{v}_{nk} + I_t}{\lambda^{-1} + I_x^2 + I_y^2} I_x \quad \hat{v}_{nk} = \bar{v}_{nk} - \frac{I_x \bar{u}_{nk} + y \bar{v}_{nk} + I_t}{\lambda^{-1} + I_x^2 + I_y^2} I_y \quad (7)$$

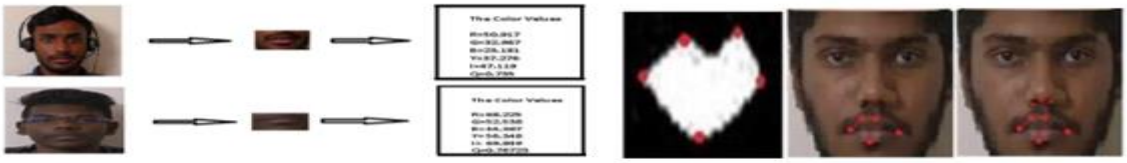


Figure 7. **Color values of the speaker**

Acoustic MFCC/ZCPA features, color-contour (F_r, F_c, F_{rc}) based feature dimension of 10 (refer figure 7), and combination features are used in the second method of feature extraction. A speaker recognition system (λ_s) built with total of 59 dimension features. The next subsection discusses in detail about accent identification system as well as speaker identification modelling.

3.3 Proposed System Modelling

This proposed system uses three modelling techniques such ANN, CNN for λ_a and HMM for λ_s system. An artificial neural network (ANN) is a network of nodes that is influenced by the simplicity of neurons in the brain. It is made up of neurons that are connected by connecting links, each of which has a weight that is multiplied by the signal transmitted in the network. The benefit of ANNs is that they can give output even if the data is incomplete after training. By commenting on comparable events, ANN learn events and make decisions.

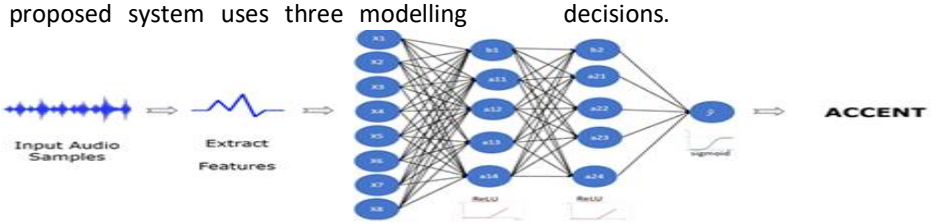


Figure 8. **Proposed Artificial Neural Network model**

An ANN has a node input layer, one or more hidden layers, and an output layer which is proposed for λ_a system having acoustic features. A multi-layer perceptron is used for binary classification in our context. Sequential model is implemented using the python library 'Keras' with Google's Tensor Flow as the backend. After each iteration, the weights are updated using the optimizer 'Adam'. The input layer consists of neurons that represent the sequential MFCC/ZCPA features. The hidden layer is a group of neurons that serve as a link between the input and output layers. In our application, the relu activation function is used to activate two hidden layers. The number of neurons in the output layer is determined by the number of classes. The output layer consists of one neuron as output which states the correct class of the accent inputted, out of the two compared accent classes shown in figure 8. In this layer, Sigmoid activation function is used. The input to the function is transformed into a value between 0.0 and 1.0. The Relu activation f_r and Sigmoid function f_s are defined as follows

$$f_r = y = \max(0, x) \quad f_s = 1/1 + e^{-x} \quad (8)$$

The proposed λ_a system using visual features are modelled using Convolution Neural Network (CNN). A CNN is a deep learning technique that can take a visual input image and assigns weights and biases to the viseme, and distinguish between them. The input, output, and polling layers are among the seven layers employed. Viseme's input layer is represented by a single feature map of 32*32 pixels. In the next subsequent process, for convolution and average polling, a feature mapping value of 6 was used with a viseme of size 28*28 and 14*14, respectively. The input visual image further convolved and pooled with the feature map 16, 120 with the size of 5*5 and 1*1 respectively. All input layers use the tanh activation function, followed by a fully linked output layer with a viseme size of ten and a soft max function used as classification.

A speaker recognition system λ_s built using acoustic and visual speech stream are modelled using a left-to-right HMM technique. For given

state models $C = \{s_1, s_2, \ldots, s_5\}$, the feature set observation $V = \{f_1, f_2 \ldots . f_n\}$, the starting probabilities (SP) of model $SP = \{\pi_1, \pi_2 \ldots . . \pi_n\}$ and the $SP_n = P(f_n = C_i)$, $T_i = P(f_{i+1} = C_i | f_t C_i)$ is the transition probability. In the HMM model, each state is represented by a GMM. The sum of weighted Gaussian distributions approximates this parametric model, which is used to express multivariate probability distributions. Each acoustic and viseme units are modelled using by GMM and the associated model parameter $M = \{P_1, .. P_A, \mu_1, \ldots \mu_n, \varepsilon_1, \varepsilon_n\}$ is shown in figure 9. The probability distribution for the observed characteristics $a_i(V)$ is calculated as follows

$a_i(V) = P(f = v | f_t C_i) = \sum_{k=1}^{M_i} \propto_{jk} P\left(\frac{V}{\mu_{jk}}, \varepsilon_{jk}\right) where\ B = \{a_i(V)\}$. (9)

From this HMM-GMM model is represented as a combination of T_i, $a_i(V)$ and SP_n.

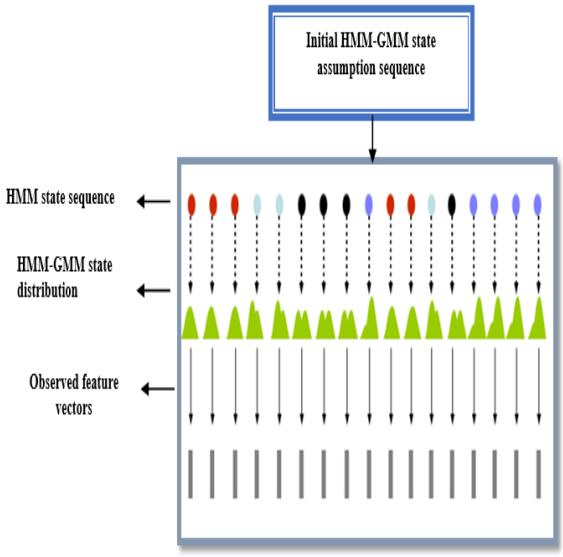


Figure 9. **A simple HMM-GMM Model**

Three HMM models are built for all the dataset which comprises of acoustic, visual and the combined feature sets. Experimental analysis is carried out for all the different models and the recognition accuracy is discussed next section.

4. Experimental Results

This section first analyzes the existing system of English language which experiments the American and Indian accent using VCTK corpus with the proposed features. In the second, this system evaluates the proposed λ_a system using acoustic characteristics (λ_{aa}) , visemes (λ_{av}) and the combined features (λ_{aav}). The comparative analysis of λ_a performed using proposed features. Third experimental evaluation carried out speaker recognition task λ_s using the proposed features and the comparative analysis is also represented. All the input acoustic samples range between 3 to 10 seconds. For instance, any two classes of a language are taken as input 80% taken for training and the remaining 20% for testing the data. The training set is then oversampled to balance the classes. Once the training is completed, the test set is used to test the trained model. The proposed system results are evaluated based on various metrics such as accuracy which is the area under Receiver Operating Characteristic (ROC) curve, precision, recall, rejection and acceptance rate which are defined using the confusion matrix.

The ratio of correct positive predictions to total projected positives is known as precision. It's also known as a positive predictive value (PPV). The ratio of correct positive predictions to total positive examples is known as recall. Sensitivity/True positive rate is another name for it. Recall is defined as the proportion of correctly predicted positive examples to total positive examples. It's also known as the sensitivity/true positive rate. Acceptance is the accuracy which is calculated as the total number of two correct predictions divided by the total number of a dataset. Table 1 shows positive predication value computed for VCTK-corpus which consists E_A, E_I. From the table acceptance rate higher in American compared to Indian accent. It is observed that precision rate is better in Indian accent and overall predication values have shown that recall is higher in rate with respect to American accent.

Table 1. **PPV representation of American and Indian Accent**

Metrics (%)	American Accent	Indian Accent
Precision	98	99.01
Recall	96.8	95.71
Rejection	99.3	99.2
Acceptance	98.5	97.61

An accent identification system recognition was tested against VCTK corpus with our proposed features with varying dimension (d) with MFCC/ZCPA features is shown in figure 10a. When d = 13 with energy coefficients gives lower in recognition rate performance for Indian accent compared with American. While increasing the dimension, there is seen more performance gain and the recognition rate was 97.5% which shows 2% in increase rate compared to Indian accent. Both E_A, $E_I$accent identification system gives the better recognition of 98% and 96.7% respectively.

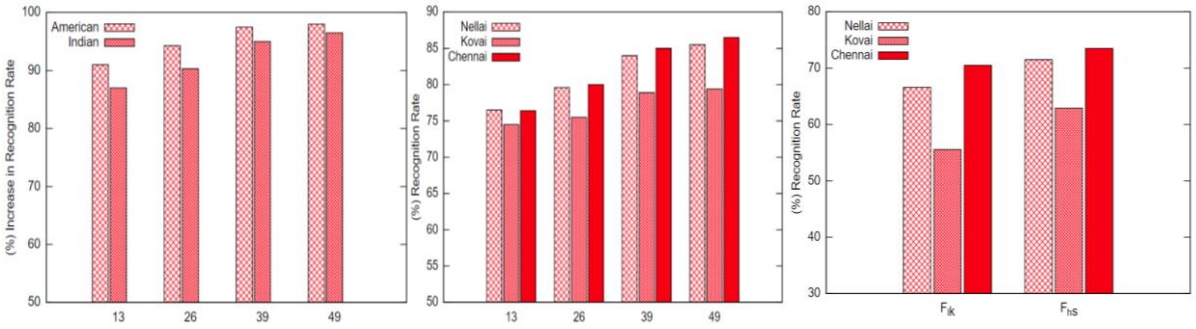
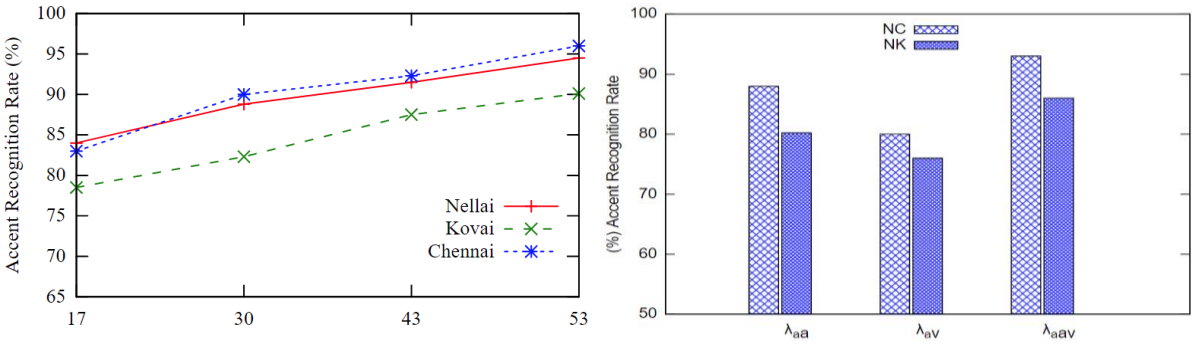


Figure 10. **Recognition rate of (a) English (E_A, E_I) (b) λ_{aa} (c) λ_{av} Accent identification system**

Table 2. **PPV representation of proposed λ_a system**

Metrics (%)	λ_{aa}	λ_{av}	λ_{aav}
Precision	91.00	78.6	92.86
Recall	89.50	75.4	92.56
Rejection	88.55	68.99	93.33
Acceptance	90.50	80	94.50

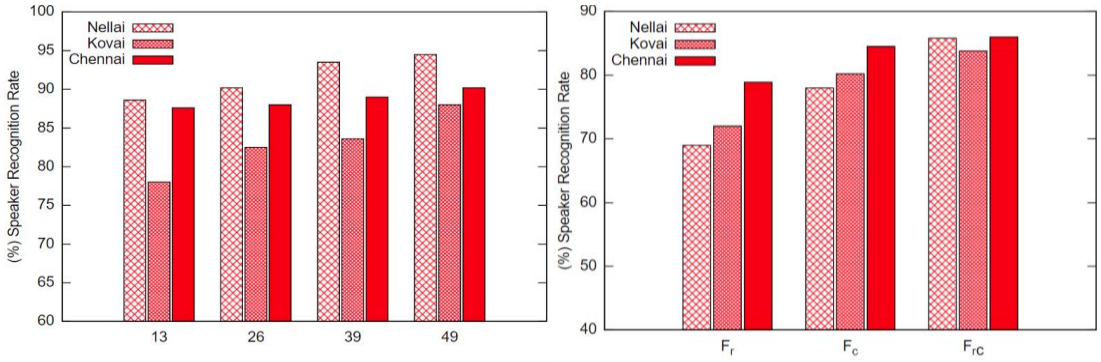
The proposed λ_a system trained and tested with SSN corpus collected manually. The figure 10b shows the performance of the acoustic system λ_{aa} with respective varying feature dimension for the different classes of sub system within each of the I_n, I_k, and I_c groups. In this system, the proposed features with the combined one gives highest recognition accuracy of 94.5% for Nellai compared with normal MFCC features. It is clearly stated that I_c and I_k based subsystem of λ_{aa} shown improvement in their recognition accuracy compared to I_k subsystem. The recognition rate of λ_{av} system using Lukas method I_k sound recognition is better in rate of 70% compared to I_n and I_k based subsystems depicted in figure 10c. In Horn method, Kovai based λ_{av} system shows the very lesser recognition performance compared to other subsystems. Since, F_{hs} feature is chosen because of the better performance in feature wise representation compared to F_{lk} and hence F_{hs} used further as visemes for combined system. Table 2 presented the positive prediction values of the acoustic based accent identification system performance with various metrics.



a b

Figure 11. **Recognition rate of λ_{aav} system compared with different categories**

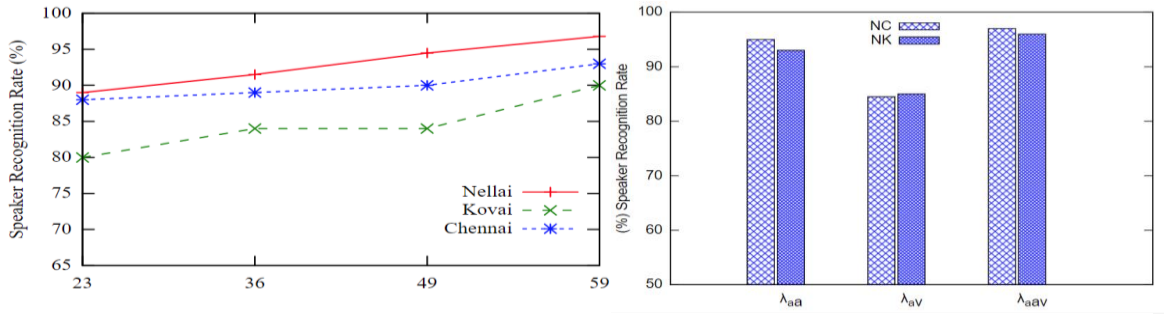
The combine accent recognition system λ_{aav} rate of Indian accent for SSN corpus trained and tested is shown in figure 11a. For both the Nellai and Chennai accent subsystems, the combined system's recognition performance is similar (96%). Figure 11b shows the proposed accent identification system performance compared with Nellai-Chennai (NC) and Nellai-Kovai subsystems. Among all these proposed Indian accent systems, the combined acoustic and visual based λ_{aav} based system shows the higher performance with the improvement in the recognition rate of 3%, 3.25% and 4%. Figure 12a shows the recognition rate of λ_{sa}, λ_{sv} Speaker recognition system with various feature combination. The Indian-Nellai accent recognition achieves better recognition of 94.5% with d=49. The Indian-Chennai gives the higher recognition rate of 90% when d=49. When there is increase in feature dimension also leads to increase in performance of λ_{sa} system. The visual feature combination with F_r, F_c, F_{cr} features such as RGB-YIQ, Active contour, and combined



a b

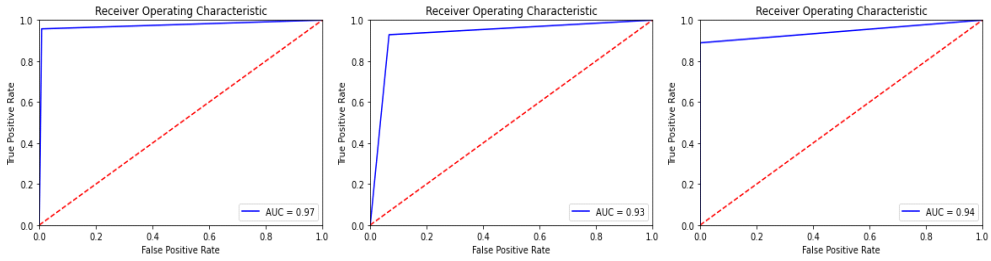
features-based speaker recognition rate is shown in figure 12b. The combined feature combination proven that the recognition is high compared with individual system feature performance.

Figure 12 **Recognition rate of λ_{sa} , λ_{sv} Speaker recognition system**



a b

Figure 13. **Recognition rate of λ_{sav} system compared with different categories**



a b c

Figure 14. **ROC of λ_{aav} system compared with different categories**

The comparative recognition performance rate of λ_{sav} system is shown in figure 13a. The combined feature with MFCC/ZCPA with the F_{cr} features dimension d=23 provides lower recognition rate. When d=36 and d=49 the λ_{aav} system with the subsystem of Nellai based achieves higher recognition rate. Figure 13b shows the proposed Indian accent system using with the three proposed λ_{sa} , λ_{sv} , and λ_{sav} subsystem. The combined λ_{sav} subsystem performance is better when combining acoustic and visual information (96%, 96.7%) respectively. Figure 14 shows the ROC of λ_a system compared with different categories λ_{aa} , λ_{av}, and λ_{aav} respectively.

5. Conclusion

Accent identification is a pre-processing step to speech recognition as it can help in fine-tuning speech recognition systems to detect accented speech better. This paper proposed new feature combination for accent identification and speaker recognition system. This work initially tested with VCKR-corpus for English (American, Indian) and American accent gives better recognition rate of

98% with our proposed feature combination. In next, Indian accent identification system with

accent subsystem of Nellai, Kovai and Chennai of the Tamil language proposed with acoustical and visual combined features. The motion visual information helps strongly for the development of accent identification system and not affected by any type of noise. The proposed combined (acoustic and visual) features give the better recognition rate compared to individual one. Among all the proposed subsystem, Chennai and Nellai sub system performance is better compared to the Kovai system. The proposed speaker identification system using same acoustic features of accented system. The multi features such color-contour with the acoustic combination-based information gives the better recognition rate for speaker recognition task (96.7%). As a result, the suggested accent recognition method retains its robustness as natural ambient conditions change.

References

[1] Carlos Teixeira, Isabel Trancoso and António Serralheiro, "Accent Identification", National Conference on Electronics, Signals and Communication, 2018.

[2] Aditya Siddhanty, Preethi Jyothix, Sriram Ganapathyz, "Leveraging Native Language Speech for Accent Identification using Deep Siamese Networks", IEEE Automatic Speech Recognition and Understanding, 2017.

[3] Alexandros Lazaridis, Elie Khoury, Jean-Philippe Goldman, Mathieu Avanzi, Sebastien Marcel and Philip N. Garner, "Swiss French Regional Accent Identification", Odyssey: The Speaker and Language Recognition, 2014.

[4] DejanStantic, Jun Jo, "Accent Identification by Clustering and Scoring Formants", International Journal of Computer and Systems Engineering, 2012.

[5] Morgan Bryant, Amanda Chow, Sydney Li, "Classification of Accents of English Speakers by Native Language", Stanford University, 2014.

[6] Matthew Seal, Matthew Murray, Ziyad Khaleq, "Accent Recognition with Neural Network", Stanford University, 2011.

[7] Paul Chen, Julia Lee, Julia Neidert, "Foreign Accent Classification", Stanford University, 2011.

[8] Tao Chen, Chao Huang, Eric Chang, Jingchun Wang, "Automatic Accent Identification using Gaussian Mixture Models", IEEE Xplore, 2002.

[9] DweepaHonnavalli, Shylaja S S, "Supervised Machine Learning Model for Accent Recognition in English Speech using Sequential MFCC Features", PES University, 2021.

[10] LIU Wai Kat, Pascale FUNG, "Fast accent identification and accented speech recognition", IEEE Xplore, 2002.

[11] Azam Rabiee, Saeed Setayeshi, "Persian Accents Identification Using an Adaptive Neural Network", IEEE Xplore 2010.

[12] M.A. Yusnita, M.P. Paulraj, Sazali Yaacob, Shahriman Abu Bakar, A. Saidatul, "Malaysian English accents identification using LPC and formant analysis", IEEE Xplore, 2011.

[13] Rasoul Mahdavi, Azam Bastanfard, Dariush Amirkhani, "Persian Accents Identification Using Modeling of Speech Articulatory Features", IEEE Xplore, 2020.

[14] AbualsoudHanani, Hanna Basha, Yasmeen Sharaf, Stephen Taylor, "Palestinian Arabic regional accent recognition", IEEE Xplore, 2015.

[15] Guntur Radha Krishna, Ramakrishnan Krishnan, Vinay Kumar Mittal, "Foreign Accent Recognition with South Indian Spoken English", IEEE Xplore, 2020

[16] Shaikh, A. A., Kumar, D. K., & Gubbi, J, "Visual speech recognition using optical flow and support vector machines", International Journal of Computational Intelligence and Applications, 10 (2), 167–187, 2011.

[17] Zhou, Z., Guoying, Z., Xiaopeng, H., & Matti, P, "A review of recent advances in visual speech decoding", Image and Vision Computing, 32(9), 590–605, 2014.

[18] Borde, P., Varpe, A., Manza, R., & Yannawar, P., "Recognition of isolated words using Zernike and MFCC features for audio visual speech recognition", International Journal of Speech Technology, 18 (2), 167–175, 2014.

[19] Radha N., Shahina A., Prabha P., Preethi Sri B.T., Nayeemulla Khan A., "An analysis of the effect of combining standard and alternate sensor signals on recognition of syllabic units for multimodal speech recognition", Pattern Recognition Letters, Vol 115, pp. 39-49, 2018.

[20] Potamianos, G., Neti, C., Luettin, J., & Matthews, I., "Audio-visual automatic speech recognition: An overview", Issues in visual and audio-visual speech processing. Cambridge: MIT Press, 2014.

[21] Yau W.C., Kumar D.K., Weghorn H, "Visual Speech Recognition Using Motion Features and Hidden Markov Models", Computer Analysis of Images and Patterns. CAIP. Lecture Notes in Computer Science, vol 4673. Springer, Berlin, Heidelberg, 2007.

[22] Radha N, Shahina A, Nayeemulla Khan A, "Visual Speech Recognition using Fusion of

Motion and Geometric Features", Procedia Computer Science, Vol. 171, pp. 924-933, ISSN 1877-0509, 2020.

[23] Triantafyllos Afouras, Joon Son Chung, Andrew Senior, Oriol Vinyals, Andrew Zisserman, "Deep Audio-Visual Speech Recognition", Computer Vision and Pattern Recognition, 2018.

Dr. N. Radha, Assistant Professor in the Department of Information Technology, SSN College of Engineering, Chennai. She has 14 years of teaching experience, including 1 year of industry experience. She received her Bachelor of Engineering degree in Computer Science and Engineering from Bharathiyar University, Coimbatore. She has completed her Master of Engineering in Computer Science and Engineering from Anna University, Chennai with distinction. Currently, she has completed her Ph. D degree (Part-Time Programme) from Anna University, Chennai in the area of Speech Recognition. She guided the projects in the area audio-visual speech recognition, speaker recognition, video processing and networks for Under Graduate and Post Graduate students. She has published about 20 research papers in her area of research both international journals and conferences.

Mr. R. M. SACHIN MADHAVAN, received his Bachelor of Technology degree in Information Technology from Sri Siva Subramaniya Nadar College of Engineering. He has completed Internal Funded Project on "Women's Safety Wearable Device". He has published 2 research papers in her area of research in international conferences. His research interest includes video processing, speech recognition, speech synthesis and machine learning.