Abstract— Magnetic Resonance Imaging technique makes it possible to measure the motion of tissues in our body organs more clearly than other medical imaging techniques. The aim of this paper is to build a co-articulatory model based on Magnetic Resonance Images (MRI). This work is a blend of various emerging technologies such as computer vision based visualization technologies, cognitive science, medical science, speech recognition. The sounds of human speech can be combined in many ways, and the associated articulator movements vary as the kinematic context changes. This kinematic variation, known as co-articulation is one of the most pervasive characteristics of speech production. Visualization of co-articulatory effects involved in the speech production will lead to a better understanding of the speech production process. MRI video obtained (from the subject AR) during the co-articulation of Tamil phonemes has been incorporated as input and processed to envision the movements of the key articulators involved in the speech production process. The Region of Interest for the articulators such as jaw, tongue, lower lip, and upper lip were obtained. The motion parameters for individual articulators and their positions in subsequent frames are estimated using Block matching algorithm. Estimated motion parameters are visualised and then reproduced. This system can act as an efficient tool to control the place of articulation visually to aid second language learners and also for the people suffering from mis-articulation to learn the correct method of articulation.

Keywords— Magnetic Resonance Imaging, Co-articulation, Articulators, Speech Production process, Block Matching Algorithm, Visualization, Motion Parameters

I. INTRODUCTION

Speech is the human’s most efficient mode of communication. Beyond efficiency, humans are very much comfortable and familiar with speech. Language acts as a master piece and helps in the process of synthesizing speech. Equally, co-articulation is also an important issue in Speech Production Process. It refers to the way in which the realization of phonetic segment [1] is influenced by neighbouring segments. In order to produce realistic speech, co-articulatory effects must also be considered.

Sound is caused because of the oscillation of air particles which includes the alternate regions of compression and rarefaction that move outwards. In free field, these waves radiate outwards from the sound source [10]. Therefore in the speech production process, while producing sound, the movements of the articulators should be significantly taken into consideration. From the above rationalization, it is obvious that by means of tracking the movements of the articulators involving in speech production process, a tool can be developed to support language learners.

MRI video was obtained from the subject (AR) during the co-articulation of complex speech sounds and words in Tamil. This MRI video has been used as an input to the proposed model. From the MRI video acquired, the regions of desired articulators were obtained as Region of Interest (ROI) that will be helpful for further processing.

After identifying Region of Interest, Block Matching Algorithm [2] was used to exactly figure out the positions of the desired articulators in the successive frames and the motion parameters were estimated. The exact positions of the articulators visible in the MR images will then be used to reconstruct an animated model. This can be carried out by means of mapping the motion parameters obtained from MRI to an animated object.

Hence this system addresses the problem of tracking the movements of the (inner) articulators obtained from MRI during co-articulation of Tamil words and focuses the attention on the design of vision-based perceptual model.

This paper is organized as follows. Section II reviews some of the related works in this context. Section III explains about the implementation details and techniques used. Section IV shows some of the experimental results obtained. Finally, Section V provides a discussion to extend the work of this system in future.

II. RELATED WORK

In this section, the discussion of the related works in the field of speech animation, motion tracking and motion estimation have been presented.

Reference [8] presents a novel method for transferring speech animation recorded in low quality videos to high resolution 3D face models. The basic idea is to synthesize the animated faces by an interpolation based on a small set of 3D key face shapes which span a 3D face space. The 3D key shapes are extracted by an unsupervised learning process in 2D video space to form a set of 2D visemes which are then mapped to the 3D face space [4]. Isomap-based Non Linear Dimensionality Reduction method was used to embed the video speech movements into a low-dimensional manifold. Here the main contribution is on the use of Isomap-based learning method to extract intrinsic geometry features of the
speech video space and thus to make it possible to define the 3D key viseme shapes. Then the 2D key visemes were extracted using k-Means clustering.

Mean Field Annealing is a method for obtaining dense displacement fields from a sequence of frames [27]. The algorithm is based on modelling the displacement field as a Markov random field, hence allowing us to use the equivalence of Markov random field and Gibbs distribution. The Gibbs distribution is used to define a mean field probability density function, and the displacement field is found using a Mean Field Annealing algorithm. Also, Mean Field Annealing is a technique which finds the global or near global minima in non-convex optimization problems.

Mean field theory in physics states that the interactions of molecules or particles comprising the system can be approximated by an effective or mean value. This allows the use of an approximate function that is much simpler, but still resembles the actual interactions between the molecules [16]. The idea behind MFA is to find a simpler energy function, or the Hamiltonian $H_o$, that approximates the given Hamiltonian $H$. This approximate Hamiltonian depends on a vector $p$, where the vector $p$ represents the mean field or effect of the neighbouring pixels on the pixel of interest. The task is then switched from minimization of the energy function $H$ to minimization of the difference between $H$ and $H_o$. A standard gradient-descent technique [27] can be used to reach the desired minimum.

Reference [9] deals with the process of identifying the air tissue boundary of the articulators in each image, which are called open contours. That requires first of all finding the start and end points of the boundary sections. The first frame of a given image sequence was manually initialized, and the start and end point locations of the boundary segments were then consecutively estimated for the subsequent images.

In contrast to open contour framework approach, a closed contour processing framework appears to be much more attractive, since it is area-based and would be expected to be more noise robust. This approach implies [21] a powerful algorithm to segment from an image, a single region with a constant level of intensity, and hence the region’s boundary against the background region can be detected [11].

In Regularized Complexity based approach [24], a Fourier contour descriptor is used to capture the outline of a single region of interest in a given scene. The region is modelled to have unity amplitude, and a discretized spatial frequency domain representation of the object model is computed. A nonlinear optimization is subsequently carried out, which aims at matching the frequency domain representation of the model to the frequency domain representation of the observed image by adjusting the boundary contour of the region of interest in the image model.

Block matching motion estimation was one of the most important aspects in the design of any video encoder which is also closely related to this proposed system which uses MRI video data [19] [26]. The block-matching algorithms eliminate the temporal redundancy, which is found predominantly in any video sequence [6]. It divides frames into equal sized non-overlapping blocks and calculates the displacement of the best-matched block from the previous frame as the motion vector of the block in the current frame within the search window.

During the process of block matching, each target block of the current frame is compared with a previous frame in order to find the best matching block of all blocks. Block-matching algorithms calculate the best match by calculating the cost (usually Mean Absolute Difference (MAD)) and the least cost block is considered as the best match block [20].

The Full search Block Matching algorithm provides the best result by matching all possible blocks within the search window. The target block of the current window is compared with each and every block of the previous window and the one which has the least MAD is selected as the best matched block. Full search block matching algorithm is also called Exhaustive search block matching algorithm. Even though, the full search block matching algorithm provides best results, it lacks significantly in computation time, which necessitates improvement.

To improve the motion estimation search time, there has been a tremendous contribution by researchers, experts from various institutions and research laboratories for the past few decades for refining the block-matching algorithms [12] [13]. As a consequence, some few fast block matching motion estimation algorithms [3] were developed. Some of them are listed here: Two-Dimensional Logarithmic Search, Three Step Search [29], Four Step Search [18], and Block-based Gradient Descent Search, Cross-Diamond Search [7], Efficient Three Step Search, Novel Hexagon-based Search [5] and many more.

All the above block-matching algorithms minimize the search time either by having different search patterns or less number of searching points to find the best matching block among the available blocks.

III. IMPLEMENTATION

There are two main driving forces behind the research in motion analysis using Magnetic Resonance Imaging (MRI).

Quantitative measurement of blood flow

Analysis of heart and other tissue motion [15] (tongue, lung, upper airway, bones, and connective tissues at a joint and skeletal muscle)

In this model, some aspects of second driving force were considered. The movements of the inner articulators have been tracked. This work has been implemented in three phases: Pre-Processing phase, Motion Tracking Phase, Visualization Phase.

A. Pre-Processing Phase

Pre-processing phase involves two preliminary tasks: MRI Acquisition and MRI Segmentation.

In MRI Acquisition of Pre-processing phase, MRI data were acquired for 30 Tamil phonemes that include stops, laterals and trills from the subject (AR). The acquired MRI has been typically in the form of video that was recorded in
the format of Audio and Video Interleaving (i.e. with avi extension).

For further processing and motion estimation, conversion of captured video into sequence of images is mandatory and this is considered as MRI Segmentation task. The sample image obtained after MRI Segmentation is shown in the Fig. 1.

Fig. 1  Sample Image obtained after MRI Segmentation (40th frame for the Tamil word KA-N-RU)

B. Motion Tracking Phase

At the end of the Pre-processing phase, sequence of images was obtained. In Motion Tracking phase, these images were processed to locate the positions of the articulators as well as to track their motions at different time periods.

Motion tracking phase entails two major tasks: ROI Identification and Motion estimation.

1) ROI Identification: From the sequence of images obtained as a result of Pre-processing phase, the region of desired articulators has to be found out in each image to track the position of articulators in successive images which in turn helpful to estimate the motion of the articulators. Manual initialization for Region of Interest for the articulators was obtained. After initializing the Region of Interest for articulators manually, automatic detection of Region of Interest was done. For this purpose, a binary mask was created and mapped. The process of Region of Interest Identification process is depicted in the Fig. 2.

Fig. 2  Flow Diagram for ROI Identification

2) Motion Estimation: The next step is to track the movements of the articulators, for which the motion of the articulators are to be estimated and motion parameters should be calculated. In order to estimate the motion parameters, Three Step Search (TSS) Block Matching algorithm was used. The principle of motion estimation is based on the assumption that the patterns corresponding to the objects and background in a frame of video sequence move within the frame to form corresponding objects on the subsequent frame. The idea behind block matching is to divide the current frame into a matrix of macro blocks and then to compare with corresponding blocks and its adjacent neighbours in the previous frame to create a vector that stipulates the movement of a macro block from one location to another in the previous frame [2]. This movement constitutes the estimated motion in the current frame. The search area for the good macro block match is constrained up to p pixels on all four sides of the corresponding macro block in the previous frame, and the parameter p is called search parameter. In this work, the value of p is considered as 7 pixels. The best match block is then found using cost function. The macro block that results in the least cost is considered as the one closest to the current block. The cost function used here is Mean Absolute Difference (MAD) which can be calculated using the equation given below:

\[
\text{MAD} = \frac{1}{N^2} \sum_{i=0}^{N-1} \sum_{j=0}^{N-1} |C_{ij} - R_{ij}|
\]

where,  
N – Size of the macro block  
C – Current block  
R – Reference block

Three Step Search: Three Step Search Block Matching Algorithm is one of the earliest fast block matching algorithms and has been used in this work to estimate the motion vectors. In the first step, the algorithm starts with the search location at the centre and initially the step size (S) was assigned as 4. Then 8 locations +/- S pixels around centre were searched. Then the best match block was chosen among these nine locations (including centre) using cost function given in the equation (i) and the resultant block was considered as a new search origin. In the second step, new step size was 2,
(S=S/2) and the similar search with respect to new search origin was repeated. Again, in the second step, one more new origin was found out which act as a centre for the third step. In the third step, similar search is made with step size 1 (S=S/1). At this point, the macro block with the least cost is considered as the best match block and the motion vector was calculated which corresponds to the best match location. The general idea of this algorithm is depicted in the Fig. 3.

For each desired articulator, some five points of interests were obtained manually. Block Matching algorithm explained above was then implemented to track the positions of these points of interests and the corresponding motion vectors for the subsequent frames were stored.

C. Reconstruction Phase

Motion vectors for the points of interest of articulators were obtained in motion tracking phase. Reconstruction phase covers two major tasks: Visualization and Mapping. In visualization, the estimated motion vectors of Motion Tracking phase are visualised through graph which shows the movements of the individual articulators at different time in each frame. Then these values are reconstructed so that we can get an exact shape of the corresponding articulators. It is also intended that the motion vectors estimated are used to map the movements to an animated object and to build a control interface so that it will be more helpful to the language learners.

IV. EXPERIMENTAL RESULTS

For this co-articulation model, MRI video of the subject (AR) acted as an input. It was then divided into sequence of images as shown in the Fig. 1 (sample image). The Region of Interest was obtained manually at the initial stage and then automatically with the help of binary mask.

Then, the image was divided into macro blocks and Three Step Search Block Matching algorithm (Section III) was implemented. The screen shot obtained after choosing the point of interest has been shown in Fig. 4.

The motion vectors were calculated and stored using Block Matching algorithm. Motion vectors provide the displacement in terms of pixels for each block under consideration. From the displacement, we can track the movements of the articulators. A sample screen shot of calculated motion vectors have been shown in the Fig. 5.

V. CONCLUSION AND DISCUSSION

It has been planned to develop a co-articulatory model for visualizing the movements of the articulators. MRI video focussing the movements of inner articulators was obtained and processed. The Regions of Interest for the articulators were identified in the sequence of frames. Some points of interest were chosen manually in the first frame and the positions of those points in successive frames were tracked using Three Step Search Block Matching algorithm. Motion vectors were estimated and stored in a database as desired parameters of the articulators. It is also intended to visualize the motion vectors of the articulators in each frame graphically at different time slices and to reconstruct the shape of the articulators in each frame with the help of estimated and stored parameters.
It is also intended to visualize the motion vectors of the articulators in each frame graphically at different time slices and to reconstruct the shape of the articulators in each frame with the help of estimated and stored parameters. It is also planned to map the estimated movements to an animated object so that the system may act as an efficient tool for second language learners, speech therapists, and also for the people who suffered from mis-articulation to learn the correct way of articulation. A visual control interface can also be developed to manually control the place of the articulators by the learners.

REFERENCES


