

Visual Speech Synthesis: How to ?

- Rohit Kumar -
December 08, 2003

Part - I

Animated Talking Faces have been around for some time now and several approaches to go develop such faces have been tried out. People have experimented with modeling the human face by 3D Parametric Models which try to generate the shapes of facial muscles based on some input parameters. As with all other parametric models, these never reach sufficient degree of perfection (atleast as of now) in terms of naturalness. Basically there are 2 problems. First the faces are not realistic i.e. they are animated characters. What is wrong with that is that they can never have the appeal that real faces can have. Second, just like in speech synthesis parametric models have a massive difficulty in modeling co – articulation i.e. transition between one shape of face to another. Its not a mere curve sketching problem. Often the outputs are jerky and ofcourse unrealistic.

Recent developments [Ezzat and Poggio] have shown an approach to develop video-realistic animated characters which are based on data driven approaches. They involve collection of a visual corpus by recording a real speaking individual and then processing it to extract the required sequence of facial images. These facial images are then put together with some processing to achieve realistic facial animation (of speech). MARY101 is the current state of art system developed at CBCL, MIT using video-realistic facial animation techniques.

The MARY101 system is capable of doing very realistic animation of the lips in synchrony with a spoken audio track which is phonetically annotated. Also the high degree of realism of the system is achieved by synchronous eye and head movements. This system is capable of generating facial animation for any kind of spoken audio track including announcements, songs, etc. independent of language in which in the actual database was recorded.

What do we want to do ?

We want to nearly replicate the efforts of Ezzat and Poggio to develop a realistic Visual Speech Synthesis System. More specifically, we want to come up with techniques described by them in preprocessing of the corpus and we would do the synthesis by a combination of a new approach (to select the image sequence) and their approach (for finally putting together the image sequence). Given ahead is an overview how we intend to proceed with this effort.

Viseme: Visual Phoneme. Viseme is the name given to refer characteristic shape of lip while uttering a particular sound.

As the first step of all, we have must come up with a set of visemes that we want to work with. The set of Indian Languages phonemes should be mapped to this viseme set.

This will be a many to one mapping. We have completed this step and have come up with a list of 15 visemes which can cover all the 45 phonemes that we interested in.

Now considering the problem of co – articulation in visual speech synthesis, we intend to experiment with divisemes (which is a transition from the stable position of one viseme to the stable position of another viseme). Divisemes cover the region of co – articulation effects and the lip transition in this region can be got from the divisemes. Hypothetically this should be able to synthesize much more natural and smoother lip transitions. We want to prove the validity of this hypothesis.

Hence, we must collect a text corpus that covers all the valid divisemes. Theoretically, there are $15 \times 15 = 225$ possible divisemes. But due to several phonotactic restrictions and other language restrictions, some of these divisemes will not be possible. An optimal text selection needs to be done to come up with a minimal representative corpus comprising of all the possible viseme transitions. We have completed this step and have implemented a greedy algorithm that does an optimal text selection over a corpus to come up with a minimal set of sentences covering the divisemes. Finally a set of 68 sentences have been selected.

The next step involves collection of a visual corpus by a real person. The selected 68 Sentences were recorded as AVI files of 640 x 480 resolution at 30 fps with a synchronized audio 22050 Hz track storing 1 byte per sample. We have used a well lit room and a native hindi speaking female to record the video corpus. The total corpus is of around 9 to 10 minutes has been recorded.

Following the corpus collection, pre-processing steps are involved. We need to do the following steps in pre – processing of the corpus either automatically or manually.

1. *Phonetic Annotation:* The collected corpus needs to be segmented at Phone Level and annotated with label marking the phone corresponding to each segment. From the marked phone we can get the corresponding viseme since it is many to one mapping from phones to visemes. At the moment we intend to do this manually and the process is underway.
2. *Residual Head Movement Normalization:* The head of the speaker does not remain stable while recording even though instructions we given to keep the head as still as comfortably possible. These residual head movement needs to be normalized either by implementing a fully automatic approach as suggested to Ezzat and Poggio using Optical Flow Methods or by other intuitive partially automated methods which may be involve a bit more manual work but not extensively. The ultimate result of this normalization will get us for every frame of the video, a set of parameter dx and dy which would be the amount the frame needs to be shifted for the head to align with our agreed base head frame. We might do this automatically within each sentence and across sentence we can give these parameters manually. Also the result should be in a form such that manual adjustment can easily be done.

3. *Brightness Normalization:* The Brightness of the recording has varied across sentences we can apply simple Image Processing to normalize it during pre – processing it so that we don't have to such effect while synthesis. This step may be fully automated.
4. *Extraction of the Lips and Eyes from the Frames:* This part may be done manually and would involve marking the masks for eyes and lips. The masks will mark the region which we will cut out to get the eyes and the lips which will then be fitted upon the base face during synthesis.

After these pre processing steps are done, a fully automated feature extraction script may be run which will build up an inventory of the visual units we have along with features like duration, context, etc.

Further in part II of this report, we will discuss the unit selection and the synthesis approach we intend to work with for Data Driven Visual Speech Synthesis.