ON THE USE OF STRUCTURED LIGHT IN SPEECH RESEARCH

John N. Carter¹ Chrstine H. Shadle¹² and Colin J. Davies¹

¹Department of Electronics and Computer Science, University of Southampton, S)17 1BJ, UK
²ATR HIP Research Laboratories, 2-2 Hikaridai, Selka-cho Soraku-gun, Kyoto 619-02, Japan

ABSTRACT

The colour-encoded structured light system developed at Southampton, and previously described for its use in enhancing EPG and acquiring 3D images of the face, has now been used to obtain quantitative measurements of lip and cheek shape during speech. 2D and 3D lip parameters are shown, as are cheek depth measures indicating expansion during the closed phase of [p].

RESUME

Le systeme de lumiere structuree codee en couleur developpe a Southampton et decrit precedentement pour son utilisation dans l'amelioration de l'EPG et dans l'acquisition des images 3D du visage est maintenant utilise pour obtenir des mesures quantitatives de la forme des lèvres et des joues en parole. Des paramètres labiaux 2D et 3D sont presents, ainsi que des mesures de la profondeur des joues, mettant en évidence une expansion pendant la phase de fermeture du [p].

1. INTRODUCTION

Knowledge of the shape of the vocal tract is necessary for understanding the mechanisms involved in producing human speech but the tract is an extremely difficult piece of anatomy to measure quantitatively. No single measurement technology will suffice to answer all questions. Most accessible are the external components comprising the lips and cheeks and many techniques have been developed to probe these areas. The Southampton structured light system has been described for its quantitative measurement of the electropalate, useful in enhancing EPG (Chiu et al., 1995) and its qualitative use in images of the external face during speech (Shadle et al. 1993). In this paper we explore its use for quantitative measurements of the external face during speech.

2. METHOD

The Southampton Colour Encoded Structured Light (CESL) system is an active optical measuring system with a 20ms time resolution and ~1mm. spatial resolution throughout its active volume. The system uses the deformation of patterns of coloured light projected onto a subject to determine the shape of their face in absolute coordinates. The patterns consist of either vertical stripes (Monks 1995) or a regular array of circular spots (Davies 1996). In each case, a 35mm. slide projector provides the coloured pattern used to illuminate the subject while a high resolution colour camera records the scene. The audio and video signals are simultaneously recorded on a SVHS Video recorder. The subject is placed approximately 1m. away from both camera and projector, which are separated by an angle of approximately 20°. The alignment of the subject, camera and projector is adjusted until the region of interest fills the field of view and is normal to the mid-line between camera and projector. Both audio and video are then acquired, in a darkened quiet room. The optical properties of human skin are such that the subjects face is covered with white makeup to obtain the best results. Synchronization between the audio and video channels is obtained using light flashes and audio beeps synchronized to the vertical synchronization pulse of the video signal. A full camera and projector calibration, including lens distortion, is used to compute the position in three dimensions of any pixel in the recorded image, interpolating between stripes or spots as required.
Once recorded the video data is digitized by stepping the SVHS recorder frame by frame, under computer control, and storing the digitized images to disk. Subsequently each picture is processed, detecting and matching coloured regions against the known pattern to produce a three dimensional grid representing the shape of the subject's face. This processing takes between 5 and 30 minutes per video frame depending on the exact technique, the accuracy desired and the computing system employed.

3. LIP SHAPE
Simple geometric measures, such as lip rounding and protrusion, have proved to be useful parameters for articulatory synthesis and modeling. The Grenoble double video system (Lallouache et al. 1988) is thus designed to compute such parameters in a form used by the Maeda articulatory model. However it is limited to measuring only 2-D projections and as such may be prone to ambiguity. In some cases 2D parameters can be measured more accurately by the structured light system, such as when the cheek fold is obscuring the corner of the lip opening. Further, some sounds require a specification of the curvature, which cannot be uniquely recovered from 2-D parameters. For instance, the shape of the lip horn, (as opposed to only the area of the opening) has been shown to affect the radiation impedance and, in particular, the cut-on frequency for higher modes (Pelorsom 1995; Pelorsom et al 1995). It also appears that lip shape affects the turbulence noise source strongly for front fricatives (Shadle 1996).

The corpus for our investigation of lip shape consists of approximately 12 seconds of nonsense words of the form [pVCV], where V = [a,i,u,y] and C = voiced fricatives and stops, liquids, and nasals. The subject was a male speaker of French (PB). Over 600 video frames have been digitized and calibrated. A significantly smaller number have been manually analyzed to derive the outlines of the lips as shown in Figure 1.

The outer border of the lips is taken from a dark line drawn on the face of the subject, while the inner border is determined by the point at which the stripe on the lip curls inside the mouth and is no longer visible. Four curves are derived from each video frame, and are converted to a 3-dimensional coordinate system. In labeling the lips great care was taken so that when the upper and lower lips are in contact the curves are identical.

The area measure in Figure 3 is calculated with a simple surface integral stepping along the upper and lower boundaries of the lip opening. The 3-D measure is larger, as expected geometrically, tracking the flat approximation in a smooth manner.
Figure 3 Area of the lip opening.

The lip separation measure in Figure 4 is similar to that of Figure 3, the 2-D measure is again less than its counterpart. The difference in distance is due, not only to geometric effects, but also to the possibility that one lip protrudes further than the other, by as much as 6mm. Again the data broadly track each other with time. However, careful examination of the diagram shows that at 4 and 4.3 seconds respectively, the three dimensional measurement reaches its maximum before the other curve. This indicates that there may be another significant component of motion, other than the vertical motion measure in projection. Both these effects are the subject of further investigation.

4. THE CHEEKS
Qualitative observations have indicated that there is significant detectable motion of the neck and cheeks during speech. Preliminary studies using structured light based on stripes have indicated the neck to be much more difficult to measure than the face and lips, because of both its concavity and the beating of the stripes with a mid-sagital reference line. The cheeks on the other hand have proved to be significantly easier to measure and equally interesting.

Figure 4 Separation between the bottom of the top lip and the top of the bottom lip.

By measuring the amount of cheek inflation, we may be able to estimate volume changes during a plosive. If so, simultaneous measurement of intra-oral air pressure and the externally-visible volume change would allow us to measure accurately the wall impedance of the vocal tract (Ishizaka 1975). An experiment has been performed to test the feasibility of this measurement. The variation of CESL using circles rather than stripes was employed, as this has advantages when imaging smooth surfaces (Davies 1996). Three sound sequences were recorded, [afa], [apa] and an exaggerated [p] sound, preceded by deliberately blowing up the cheeks. Audio and video recording were made, and the optical system calibrated to determine the 3-dimensional shape. Reviewing the video tape in slow motion has shown that there is a very noticeable expansion of the cheek around the [p] in [apa]. This can be demonstrated graphically by comparing sections across the face for different sounds. Figure 5 shows an image of the subject's face(SM), close to the occurrence of the [p]
in [apa]. The z coordinate of the section marked has been plotted in Figure 6 for the three sounds under investigation. The three curves in Figure 6 share some common characteristics. Firstly, the z value at the left hand side of each curve is similar, confirming that the subjects head was not parallel to the reference plane and finally the stepped nature of the curves, is due to a simple interpolation algorithm. It is clear, however, that the inflation of the cheeks for the [p] in [apa] is greater than that measured for [f] in [afa], indicating the applicability of this technique for estimating the volume change of the vocal tract. Such an analysis is underway.

5. SUMMARY AND FURTHER WORK

In this paper we have illustrated the use of our structured light system to obtain quantitative measurements of lip shape during speech, including the area of the curved surface enclosed by the lips, and cheek shape, which is potentially useful for estimating volume change of the vocal tract during speech. Other applications to speech research exist; those we are currently investigating include:

1. measurement of the teeth and tongue tip position when these are visible, as in Fig. 1;
2. a comparison of the shape of plaster casts of the lip area for one of our subjects (PB) and the shape of his lips during continuous speech, to assess deformations induced by the plaster;
3. acquisition of the neck shape during speech. Use of colour-encoded spots rather than stripes makes this possible, but the acquired shape is complex and difficult to interpret due to the apparent motion of the underlying musculature.

This work has been supported in part by a European CEC-ESPRIT BRA Project - Speech Maps. The authors also acknowledge the patience and willingness of their subjects, Pierre Badin and Sheila Mair.

6. REFERENCES


