Collection of Korean Audio-video Speech Data

Cheol-Woo Jo* · Roland Goecke · Bruce Millar**

ABSTRACT

In this paper a detailed description of collecting Korean audio-video speech data is presented. The main aim of this experiment is to collect some audio-video materials which can be used for later experiments to estimate and model the actions of the visible human articulatory organs such as mouth, lips and jaw. We collect audio-video data from seven directions separately. Twelve markers are used to trace the movements.

Keywords: Audio-Video Speech, Modeling, Articulatory Organs

1. INTRODUCTION

Human processing of speech using both auditory and visual modalities and machine processing of speech using both audio and video signals are related activities which are attracting increasing interest from psychologists and information scientists and engineers. Knowledge of the processes which link the two sensory modalities of audition and vision can be used to improve the performance of speech recognition especially in adverse conditions and to enhance the message transfer capability of synthetic speech by adding appropriate visual cues. This knowledge may also be applied to enhance the performance of certain tasks in multimedia indexing and in human-computer interaction.

The methods for accumulating this knowledge start with the way in which the video data of speech activity is acquired. Worldwide the various groups working in this area use techniques ranging from sophisticated systems involving the tracking of surface-mounted sensors, through the detection of painted markers or completely "made-up" lips, to systems that analyse the natural face by means of image processing techniques.

The experiments described in this paper focus on methods for the acquisition of

* Dept. of Control & Instrumentation Eng., Changwon National University, Korea
** Computer Sciences Laboratory, RSISE, Australian National University, Australia

audio-video data in which the face is either unmarked or minimally marked. Finally korean audio-video data is collected. Data collection process is focused on the multi-view and multi-purpose applications.

2. CURRENT STATUS OF A-V SPEECH PROCESSING

2.1. Comparisons Between Measurement Methods

The following illustrates current auditory-visual speech processing research conducted in eight projects in France, Germany, Italy, Japan, Sweden and the USA. While applications and data capture methods vary quite considerably, there are certain key factors concerning the modelling of data which are worthy of global comment.

Guiard-Marigny et.al. [Gui92] measured width(A), height(B) and lip contact protrusion automatically using one camera and mirror. Based on those they constructed 22 basic lip contours(visemes) by applying polynomial approximation. They measured frontal view together with lateral view. Horiuchi et.al. [Horiuch97] from Chiba University, Japan performed facial gesture recognition using automatic detection of lips, eyebrow areas by line integration. Frame rates for the measurement was 30 frames per second. Jie Yang et.al. [Yang97] from Univ. of Karlsruhe, Germany, performed face and lip tracking using MS-TDNN with the frame rate of 20~30frames/sec. They used 24bit RBG and 16KHz audio and a Canon VC-C1 camera. Reveret et.al. [Reveret97] from ICP in Grenoble, France, constructed 3D active contour model of lips using optotrack marker tracking. Five markers were used to detect head position. Basu et.al. [Basu97] from MIT media laboratory recovered 3D lip structure from 2D observations. The frame rate for the experiment was 30 frames per second and total 17 markers are used including 16 markers on the lips. Vogt [Vogt97] from Univ. of Stuttgart, Germany, used 11 marker points to recognize lip shapes using TDNNs. The frame rate was 25 frames per second in this case. Lyberg et.al. [Lyberg98] from Telia Research AB, Sweden, used 29 markers using McReflec motion tracking system. IR Marker plus 4 camera was used. Cosi et.al. [Cosi98] from IFD, C.N.R., Italy used IR markers and 4 cameras to investigate 3D lip motions.

From this brief survey we know that there are no established standards for placing facial markers or for related measurement methods as yet. They have varied according to the available equipment or purposes of the experiment.

In terms of it's dimensions, models can be classified as in table 1. Table 1. shows the comparisons between 2D and 3D models. While 2D model is easy to measure, 3D model provides better information on orientation.

Table1. Comparisons between 2D and 3D model

	2D model	3D model
Good	Fast processing Relatively easy model	Useful for orientation free lip tracking, synthesis
Bad	Frontview image only Can detect only small variations of orientation (by C.Benoit, AVSP '98)	Relatively slow processing Complex model

In terms of dimensions of models used, there are two different models.

(1) Parametric modelling

The parametric modelling method tries to model the shape of lips or articulatory organs into some simple set of parameters which can represent its motion. In this case estimation of each parameter is required. ICP's model [Reveret97] is an example of such a parametric model.

(2) Key-frame modelling

Key-frame modelling tries to model the shape of lips or articulatory organs into a set of frames of images which has the characteristics of the movement of the articulatory organs rather than extracting some set of parameters. For example, many methods which use VQ or HMM can be classified into this category.

When the application is related to the automatic recognition of speech, then automatic detection of the lips from unmarked natural images is required. When the main focus is on building models for more effective audio-video synthesis then the lips may be detected either automatically or interactively using either a marked or an unmarked face.

3. SELECTION OF UTTERANCES

First, lists with required contents of phonemic environments are produced considering the shape of mouth, place of articulation, lip protrusion. As in other speech databases, the list must include all the possible conditions which can represent the characteristics of a specific language. In an audio-video speech database additional issues concerning the visual modality must be considered. The following is the method used to classify and select key utterances to represent Korean auditory-visual speech.

3.1. Classification of Korean Phoneme Set

Korean vowels are classified as follows according to the visible evidence of their articulation. Only monophthongs are included here. Figure 1 shows the classification of

Korean vowels according to the shape of mouth. Table 3 shows the classification of Korean consonants according to their place of articulation.

Table2. Pronunciations of Korean Vowels

Hangul	Sound	Hangul	Sound	Hangul	Sound
ㅣ	i	ㅡ	eu	ㅜ	u
ㅔ	e	ㅓ	eo	ㅗ	o
ㅐ	ae	ㅏ	a	ㅚ	oe

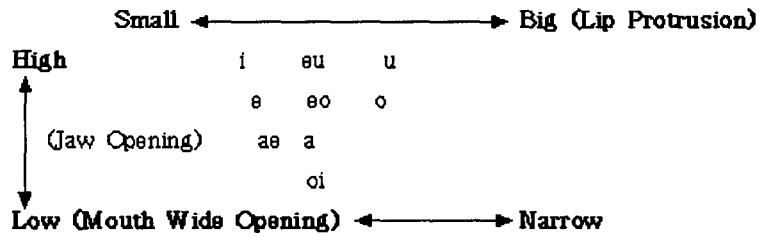

Fig. 1 Classification of Korean vowels according to the shape of mouth

Table 3. Classification of Korean consonants according to their place of articulation

Place of Articulation	Phonemes	Number
Bilabials	/b/, /p/, /pp/, /m/	4
Dentals	/d/, /t/, /dd/, /s/, /ss/, /n/, /l/	7
Dentals	/z/, /zz/, /ch/	3
Palatals	/g/, /k/, /gg/, /ng/	4
Finals	/K/, /P/, /T/, /n/, /ng/, /m/, /l/	7

Different consonants may show the same or similar shape of articulatory organs. Considering this, consonants are classified into 7 sets which are thought to generate the same shape of mouth. According to the classification shown in table 4, a pronunciation list was produced. The pronunciation list is constructed in the form of VCV contexts in which the surrounding vowel is chosen as /a/. Consonants used are limited to /b/, /d/, /g/, /s/, /z/, /n/,/l/.

Table4. Consonants correspondence list

Representing Phoneme	Subordinates
/b/	/pp/, /bb/, /m/
/d/	/t/, /dd/
/g/	/k/, /gg/
/s/	/ss/
/z/	/ch, zz/
/n/	
/l/	

3.2. Utterance List

All vowels and 16 syllables were pronounced from 7 different directions. Diphthongs were not included in this list in order to restrict the size of the data within manageable limits. The total number of utterances was (8+7+9)x7=168. Additional consonants from the front were recorded separately. Accordingly a total of 185 utterances were recorded. The table 5 provides the total list of utterances.

Table 5. Total utterance list

8 vowels
/a/, /e/, /i/, /u/, /o/, /eu/, /eo/, /oi/
VCV(7): aCa C=/b/, /d/, /g/, /s/, /z/, /n/, /l/
/aba/ /ada/ /aga/ /asa/ /aza/ /ana/ /ala/
CVC(9): V=/a/, /u/, /i/
/gaG/ /baB/ /gaT/ /daG/ /baT/ /gaB/ /daB/ /daT/ /baG/
Utterance lists for 17 consonants from the front only
/aga/ /agga/ /ana/ /ada/ /adda/ /ala/ /ama/ /aba/ /abba/ /asa/ /assa/ /aza/ /azza/ /acha/ /aka/ /ata/ /apa/

4. COLLECTION PROCESS

4.1. General Descriptions

According to our knowledge there has been no previous systematic collection of

audio-video data for Korean which is suited to a wide range of scientific investigations. Data collections have been made for specific clinical experiments and modelling of the interactions across those domains has been based on perceptual measurements. In this study we examine two methods of acquisition and the prospects for modelling of cross-domain interactions directly from the audio and video signals. Both methods of acquisition used the same transducers and data logging equipment. Audio signals were acquired using a clip-on Sennheisser MKE-10 microphone mounted approximately 20cm below the speakers lips. Video signals were acquired via a Flexicam 1/3" colour CCD camera positioned approximately 30cm in front of the lips. These signals were combined into uncompressed AVI files via the Adobe Premiere LE software package and a Fast Electronics FPS60 framegrabber operating at 320x240 pixels and 20 frames per second and a sound card operating at 16bits (mono) and 11050 samples per second. These facilities were implemented on a 300MHz Pentium-II computer operating under Windows 95. The methods of acquisition differed in the way in which three-dimensional data was gathered involving two different room set-ups. These set-ups were recorded in detail via the web-based CSLD system [Millar98] and are described in general terms below. Within each environment, data was collected from a Korean native speaker. The Korean speaker is 37 yr old native Korean who uses south eastern dialect of Korea.

4.2.Experimental setups for audio-visual speech collection

The speaker was seated in a swivel chair which was adjusted in height to maintain alignment of face position in front of a fixed camera position. Markers both for placement of feet and for direction of gaze were established in carefully measured positions on the floor and on suitable vertical surfaces such that the speaker could quickly orientate himself to one of the following 7 positions: facing the camera position, 45′ to the left, 45′ to the right, 90′ to the left, 90′ to the right, 45′to the leftback, 45′ to the rightback.. These directions were given numerical tags as specified in figure 2 to enable precise referencing on all data files. Natural light through large windows was used for illumination within this setup.

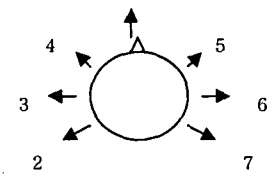

Figure 2. Seven orientations

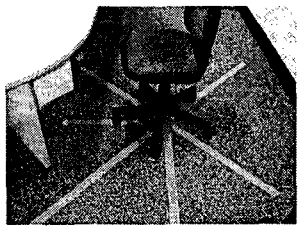

Figure 3. Markers on the floor

4.3. Facial Markers

To specify the reference points, 12 markers are attached on the important positions of the subject's face.

The positions are as follows;

(1) Around the lips: five markers are put to measure the movements of the lips.

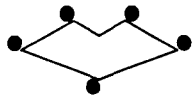

Figure 4. Positions of Markers around the Lips

(2) On the horizontal line connecting nose and ear. About 1 cm apart from nose.
Left and right 2 points each. 4 points.
These markers are put to get the horizontal reference lines.

(3) At chin, 3 points : these markers are put to measure the movement of the jaw.

(4) The size of the markers are approximately 5 mm in diameter. Their colour is dark blue.

4.4. Collection process

Each speaker was asked to study the utterance list and be ready to pronounce each item on the list in succession with a distinct pause between the utterances.

(1) Subject sits on chair and start speaking from direction 1.

(2) Start speaking from direction 1.

(3) Speak whole list with enough pause between the utterance

(4) Save the data into file.

(5) Move to next position

(6) Repeat steps (1) to (7).

4.5. Way of speaking

Speaker was asked to pronounce the suggested list once at each specified position with enough pause between the utterances.

4.6. Recording Process

To prepare recording the following actions were taken before recording.

(1) Start Adobe Premiere
(2) Set project video size to 320 x 240
(3) Click on "File>Capture>Movie Capture"
(4) Click on "Movie Capture". Check "Record Video" and "Record Audio" to set proper frame rate and sampling rate.
(5) Set video options in "Movie Capture>Audio Options". Use 20 fps, no compression. Verify the capture limit is switched off.
(6) Set audio options in "Movie Capture>Audio Options". Use 11 kHz, 16 bits mono

Note: We had no dropouts using these settings.

The recording procedures were as follows

(1) When ready, click on "Record" button and start utterance.
(2) To stop, simply press either the left mouse button or ESC.
(3)I f capturing was successful, the resulting AVI video can be viewed.
(4) Use "SAVE" or "SAVE As" to save file.

4.7. Experimental setup for dual image audio-visual speech collection

The speaker was seated in a chair which was adjusted in height to achieve alignment of the face position in front of a fixed camera position, and which was adjusted in placement to achieve alignment with a fixed mirror position. The mirror was mounted vertically on a fixed structure such that its reflective surface subtended an angle of 45 degrees to the axis of the camera. The speakers cheek was positioned adjacent to the mirror such that it contained an image of the face profile when viewed from the position of the camera. The facial area was illuminated by light generated by an overhead projector angled to provide both direct and reflected augmentation of the natural light in the room and to generate minimal shadowing on the face. The natural light in the room was reduced from that used in section 4.2 by the use of blinds so that the two-dimensional background of the mirror images were suitably contrastive with the illuminated face.

4.8. Experimental Procedure

The experiment proceeded in an identical manner to that described in section 4.6 except that just one reading of the material was carried out for each condition of "marked" face and "unmarked" face with the speaker maintaining the same position throughout.

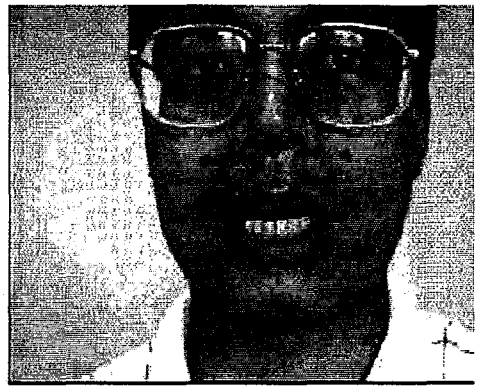

Figure 5. Example image of the collected data

5. CONCLUSION

In this paper, details of the collection procedures of Korean audio-video speech data are described. Korean vowels and consonants are classified into categories according to their visible articulatory characteristics. Based on this classification, representative data for Korean speech was collected.

The collected data provides us with 7 different views for the same repeated utterance. Dual orthogonal images of identical utterances using a mirror technique were also collected. These audio-video data will be useful to analyse the movements of the outer articulatory organs in various ways. The data can be used in various experiments requiring auditory-visual information. This is a pilot collection which we expect to lead to further data collection using refined methods and multiple speakers.

REFERENCES

Guiard-Marigny. T. 1992. "Animation en temps reel d'un modele parametrise de levres." In Memoire de D.E.A. *Signal Image Parole*, l'Institut National Polytechnique, Grenoble, France.

Ichikawa, A., Okada, Y., Imiya, A., and Y. Horiuchi. 1997. "Analytical method for linguistic information of facial gestures in natural dialogue language." 105-108, *Proceedings on AVSP'97*, Rhodes, Sept.

Meier, Uwe., Stiefelhagen, Rainer., and Jie Yang. 1997. "Preprocessing of visual speech under real world conditions." 113-116, *Proceedings on AVSP'97*, Rhodes, Sept.

Reveret, L., Garcia, F., Benoit, C., and E. Vatikiotis-Bateson. 1997. "An hybrid orientation-free liptracking." 117-120, *Proceedings on AVSP'97*, Rhodes, Sept.

Basu, Sumit and Alex Pentland. 1997. "Recovering 3D lip structure from 2D observations using a model trained from video." 121-124, *Proceedings on AVSP'97*, Rhodes, Sept.

Vogt. M. 1997. "Interpreted multi-state lip models for audio-visual speech recognition." 125-128, *Proceedings on AVSP'97*, Rhodes, Sept.

Hallgren, Asa and Bertil Lyberg. 1998. "Visual speech synthesis with concatenative speech." 181-183, *Proceedings on AVSP'98*, Terrigal, Dec.

Caldognetto, Emanuela Magno and Claudio Zmarich, Piero Cosi. 1998. "Statistical definition of visual information for Italian vowels and consonants." 135-140, *Proceedings on AVSP'98*, Terrigal, Dec.

Millar, 1998. "Customisation and quality assessment of spoken language description", 1575-1578, *Proceedings of ICSLP'98*, Sydney, Dec.

Received : Jan. 20. 2000.
Accepted : Feb. 15. 2000.

▲ Cheol-Woo Jo
Dept. of Control and Instrumentation Eng.
Changwon National University
Changwon, Kyeongnam, 641-773 KOREA
e-mail: cwjo@sarim.changwon.ac.kr
Also, Speech Research Laboratory
Applied Science & Engineering Labs
duPont Hospital for Children
P.O. Box 269
Wilmington, DE 19899, USA
e-mail: jo@asel.udel.edu

▲ Roland Goecke
Computer Sciences Laboratory
Research School of Information Sciences and Engineering
Australian National University
Canberra, ACT 0200, AUSTRALIA
e-mail: Roland.Goecke@discus.anu.edu.au

▲ J. Bruce Millar
Computer Sciences Laboratory
Research School of Information Sciences and Engineering
Australian National University
Canberra, ACT 0200, AUSTRALIA
e-mail: Bruce.Millar@anu.edu.au