Audio-Visual Asynchrony in Malayalam Phonemes and Allophones

Bibish Kumar K T, Sunil John, Muraleedharan K M, R K Sunil Kumar

Abstract: Simulating the audio-visual asynchrony (AVA) is just one of those essential issues to be researched in the use of video signal along with the audio signal in the speech processing applications. AVA analysis deals with the estimation of the asynchrony between audio and visual speech signal produced during the articulation of phonemes and allophones. Just a few works of literature have discussed this specific dilemma that immediately reflects more exploration is needed to tackle this open research issue. An audio-visual Malayalam speech database containing of 50 phonemes along with 106 allophones of five indigenous speakers has been created. The listed visual information is made up of the complete facial area recorded in a frontal perspective. Time annotation of the audio and video signals is performed manually. Duration of audio signal and video signal of every phonemes and allophones are estimated from the time annotated audio visual database. Asynchrony is then estimated as their differences. Asynchrony analysis was performed individually for phonemes and allophones to underline the coarticulation effect. Multi modal speech recognition has greater accuracy than audio only speech recognition, especially in noisy environment. AVA plays a vital role in applications like multi modal speech recognition and synthesis, automatic redubbing, etc.

Keywords: Audio-Visual Asynchrony, Preservatory Coarticulation, Anticipatory Coarticulation, Phonemes and Allophones.

I. INTRODUCTION

For decades, audio speech signals was alone used in speech-based applications to emulate the human perception. From 1976 [1] onwards speech-based applications consider visual stimuli to improve the speech intelligibility in acoustically noisy condition. This integration creates significant improvement in the performance of audio-only speech system; however, it creates new computational problems in the fusion process. Visual signals, along with corresponding audio signals, improves the detection of speech segment information and speaker localization in the speech pre-processing context. However, one of the essential and critical problem when dealing with the fusion of two modalities of a different type is audio-visual speech asynchrony [2].

II. DATABASE

Malayalam linguistic properties are well described in [7]. The language material is made up of 50 phonemes: 10 monophthong vowel phonemes, two diphthong vowel phonemes and 38 consonant phonemes, and 106 allophones: 28 monophthong vowel allophones, three diphthong vowel allophones and 75 consonant allophones. The sound data is recorded with a sampling frequency of 44.1 kHz and movie is recorded using frame dimensions 1280 x 720 with 25 frames per second (fps). Manual time annotation is completed separately for both information for 5 speakers.
Sound speech signals are examined at a non-overlapping windows of dimensions 10 ms. The visual language signs were upsampled from 25 Hz (frame rate) to 100 Hz throughout the joint investigation of both streams so as to cop up with the sound frame rate of 100 Hz.

### III. AUDIO-VISUAL ASYNCHRONY IN MALAYALAM PHONEMES

This section discusses the audio-visual asynchrony due to articulation inertia alone. From the recorded data, audio signals are extracted from video and stored as separate files. The video file was then converted into frames. Each portion in the speech signal is encoded into a text file. Fig. 1 shows the time annotation of audio speech of monophthong vowel อำเภอ/əː/. The duration of the underlined phoneme is shown above the speech segment in seconds (in black colour, red coloured number indicates the time boundary).

![Fig. 1. Time annotation of audio speech of monophthong vowel phoneme-อำเภอ/əː/](image)

The visual speech annotation of the corresponding phoneme is shown in Fig. 2. The time duration of the visual speech segment is calculated by multiplying the number of frames (contain lip gestures of the underlined phoneme only) with the reciprocal of fps. The red line indicates the speech element boundary in visual speech.

![Fig. 2. Time annotation of video speech of monophthong vowel phoneme-อำเภอ/əː/](image)

Immense care is given while performing the time annotation of consonant phonemes. Since the consonants in Malayalam always appear like Consonant-Vowel (CV) syllable, the precise boundary detection is possible by repeatedly checking the sound of the selected consonant phoneme only. Fig. 3 shows the time annotation of a consonant phoneme.

![Fig. 3. Time annotation of audio speech of consonant phoneme อำเภอ/-b/](image)

The duration of the visual counterpart of consonant phoneme is shown in Fig. 4.

![Fig. 4. Time annotation of video data of consonant phoneme อำเภอ/-b/](image)

### IV. AUDIO-VISUAL ASYNCHRONY IN MALAYALAM ALLOPHONES

The coarticulation effect in Malayalam speech is embedded in Allophones. In a word, the occurrence of a speech element is highly influenced by the linguistic properties of the preceding and proceeding speech element and its relative position: at the beginning, intermediate, and end, in that word. In this paper, we estimated the audio and visual duration of each allophone in a different position in a word: initial, intermediate, and final. The time annotation of a monophthong allophone in three positions is shown in Fig. 5.

![Fig. 5. Time annotation of different allophones of อำเภอ/-a/ at different positions.](image)

Fig. 5. Time annotation of different allophones of อำเภอ/-a/ at different positions.

The upper and lower speech elements represent the allophone of the phoneme อำเภอ/-a/ situated in initial and final position in a word.
The middle speech segments represent another allophone of the phoneme /a/ situated in the intermediate position. The name and IPA of corresponding speech element are shown in the top portion. The visual counterpart of the time annotation of a monophthong allophone in three-position is shown in Fig. 6.

V. RESULT AND DISCUSSION

From the time annotation process, the time lag between the audio and visual streams was estimated separately for phonemes and allophones. After estimating the time lag, the histogram of asynchrony distributions is tabulated for phonemes and allophones as in Fig. 7 and 8, respectively. For phonemes, the histogram is centered in the visual lead region with few distributions in the audio lead region. On average, the audio signals are maximally correlated with visual signals 83 ms in the past. The effect of articulator inertia is reflected in the wide range of the histogram, ranging from -157 ms to 343 ms.

For allophones, the histogram is centered near the boundary between the synchronous and visual lead region with comparatively few distributions in the audio lead region when compared to Fig. 7. On average, the audio signals are maximally correlated with visual signals 12 ms in the past. The effect of coarticulation effect is reflected in the narrow range of the histogram, ranging from -37 ms to 203 ms.

In general, anticipatory coarticulation is prominent in Malayalam language. The asynchrony between the audio and visual speech is relatively protruding in phonemes than in allophones.

REFERENCES

1. McGurk, Harry, and John MacDonald. "Hearing lips and seeing voices." Nature 264, no. 5588 (1976): 746.
2. Katsaggelos, Aggelos K., Sara Bahaadini, and Rafael Molina. "Audiovisual fusion: Challenges and new approaches." Proceedings of the IEEE 103, no. 9 (2015): 1635-1653.
3. Mattheyses, Wesley, Lukas Latacz, and Werner Verhelst. "Comprehensive many-to-many phoneme-to-viseme mapping and its application for concatenative visual speech synthesis." Speech Communication 55, no. 7-8 (2013): 857-876.
4. Vatakis, Argiro, Petros Maragos, Isidoros Rodomagoulakis, and Charles Spence. "Assessing the effect of physical differences in the articulation of consonants and vowels on audiovisual perceptual temporal perception." Frontiers in integrative neuroscience 6 (2012): 71.
5. Kent, R. D. "Coarticulation in recent speech production." Journal of Phonetics 5, no. 1 (1977): 15-133.
6. Keating, Patricia A. "Underspecification in phonetics." Phonology 5, no. 2 (1988): 275-292.
7. http://www.cmillemu.in/phonetic/#/

AUTHORS PROFILE

Bibish Kumar K T earned his Master's Degree in Physics from the University of Calicut in 2010. He is an active research scholar working in the area of audio-visual speech analysis under the supervision of Dr. R K Sunil Kumar. He has six research publications in peer-reviewed journals and international and national conferences to his credit. He joined in PhD program in 2015. He guided 30 post graduate projects in the area of audio and visual speech processing. He developed audio visual Malayalam speech data base. His area of interests are Digital Processing of Speech and Image and Digital Electronics.

Sunil John earned his Master’s degree in Physics from the University of Calicut in 2008. He is an active research scholar working in the area of multi accent speech analysis under the supervision of Dr. R K Sunil Kumar. He co-authored three books on computational physics, Methodology of science and Physics practical for undergraduate.
Audio-Visual Asynchrony in Malayalam Phonemes and Allophones

He has three research publications in peer-reviewed journals and international and national conferences to his credit. He is currently working as Assistant Professor in the department of physics, St. Mary’s College, Sulthan Bathery, Kerala, India. His area of interests are Digital Signal Processing and Computational Physics.

Muraleedharan K M earned his Master’s in Physics from the University of Calicut in 2003. He is an active research scholar working in the area of Nonlinear modelling of speech production system under the supervision of Dr. R K Sunil Kumar. He has five research publications in peer-reviewed international journals and national conferences to his credit. He is currently working as Assistant Professor in the department of physics, SARBTM Government College, Koyilandy, Kerala, India. His area of interests are Digital Signal Processing, Nonlinear dynamics and Classical Mechanics. He guided 15 post graduate projects in the area of Digital Signal processing.

Dr. R K Sunil Kumar earned his PhD in Speech Signal Processing from University of Calicut, Kerala, India in 2004. He has published several research papers in peer-reviewed journals and national and international conferences in the area of signal processing and artificial neural networks. His research interest includes speech signal processing and neural networks and active noise cancellation. He worked as Assistant Professor in the department of physics, Government College Madappally, Kerala, India for four years. Currently he is working as assistant professor in the department of IT, Kannur University, Kerala, India. He is guiding four Research Scholars in the area of speech processing.