# Research Report

## Submitted

**To**

## Internal Quality Assurance Cell (IQAC)

## Kannur University


## CREATION OF STANDARD AUDIO SPEECH DATABASE IN MALAYALAM FOR THE DEVELOPMENT OF VOICE INTERACTIVE MACHINE

Dr. R.K. Sunil Kumar, Assistant Professor, Department of Information Technology

Kannur University

# CREATION OF STANDARD AUDIO SPEECH DATABASE IN MALAYALAM FOR THE DEVELOPMENT OF VOICE INTERACTIVE MACHINE

Dr. R.K. Sunil Kumar, Assistant Professor, Department of Information Technology

Kannur University

**Abstract**

Although audio-visual speech based application using resourced languages has achieved its acceleration towards robust outcomes, the realm of speech based application using under resourced languages has not gathered enough momentum in the research domain. One of the in-depth reason cited by the research community is the scarcity of application oriented speech database. The goal of this work is to present a new multi-application oriented speech database in Malayalam language and describing it under the background of available audio-visual speech database. This will be the first standard audio-visual speech database in Malayalam recorded in different condition to address specific problem. This database was recorded in 3 phases with unique speakers in each phase which makes it a speaker independent database. The first phase of recording creates audio-only speech database which contain 50 isolated phonemes captured in entirely two different condition one in isolated and noiseless environment and other in acoustically realistic environment. The second phase is utilized to capture audio and visual signal from 5 female speakers uttering isolated phonemes in acoustically and visually realistic environment. The third phase of recording contain audio-visual speech database recorded from 25 female and 5 male speakers uttering 50 Malayalam isolated phonemes and 207 connected words comprising of all allophonic variations in controlled environment. Each isolated phonemes and word in audio and video domain are properly segmented and labeled. The present database was the result of efforts taken in the past two years and it is expected to grow its dimension and content in recent years too.

## Introduction

In the past few years, humans have been interacting with a machine in every aspect of their life. For better human-computer interaction, the machine has to perform just like a human interacts with his surrounding. In speech processing aspect, human interaction with its surrounding is bimodal in nature. The audio signal from the mouth is the primary source for recognizing speech but it is well established that incorporating visual signals from the mouth has improved the efficiency of speech perception level. Even though the visual part of speech contains less speech information than the acoustic part, it helps better understanding when the acoustic part is less informative. Many technological advances have been taking place in speech recognition for various languages in recent years. The main focus of research groups remained around building ASR (automatic speech recognition) systems for European languages especially English. English is the language for which maximum work for speech recognition is done. But the fascination of speech recognition has made various research groups curious to develop systems in their native or local languages. In India, ASR systems have been developed for various languages like Hindi, Tamil, Kannada, Oriya, etc. However, speech recognition in Malayalam is still in its inception stage and very few works have been reported yet. The reported works in the Malayalam language mainly deal with audio signals and in addition to that, no standard audio-visual speech database has been reported yet. Researchers working in languages, which are less addressed by the speech processing group, are forced to do their work with their own created database which brings problematic condition when comparing with other ones.

Over the years of development and unceasing research in audio-visual speech recognition area has endorsed the dearth of standard audio-visual speech database. The need for diversity in resources and huge storage capacity are the main challenge that produces the remarkable hindrance in the development of bi-modal speech database. Creation of exhaustive and methodically arranged audio-visual speech database build up a worldwide acceptance in the research community. A diversity of standard database has been reported and most of them

claim to be beneficial for the specific task. The database to be invented should serve more than one goal so that it will be useful for alternative research works. The main criteria needed for speech database construction is that it should have a large phonetically balanced speech corpus uttered by a large number of unique speakers in an uncontrolled environment. For the development of an efficient ASR system, long duration of the annotated recording of the speech utterance is needed for both training and testing schemes. It is mandatory to figure out the peculiarities of the language of the database and its linguistic background and comparing it with other groups of languages which help to resolve the issues arise during the creation of the database. The database reported in this work aims to implement the above-mentioned approaches and makes an attempt to fill the gaps in Malayalam speech associated applications. Motivated by the demand for a regional database, this work will be the first audio-visual speech database in Malayalam.

## 2. Database Design

The audio-visual speech database contains audio and visual recordings simultaneously captured while a speaker is speaking. The purpose of a speech database depends deeply on the language material, speaker population and the quality of recording which should be optimally adapted to attain the prime goal. It is better to consider a continuous speech corpus for speech recognition and isolated words for speaker verification. A large number of speakers are required for the verification task than compared to the recognition task. Extracting minute variations in the visual features of the speaker may improve the recognition task which can be implemented by using high definition camera rather than cameras with high fps (frames per second) under better illumination condition.

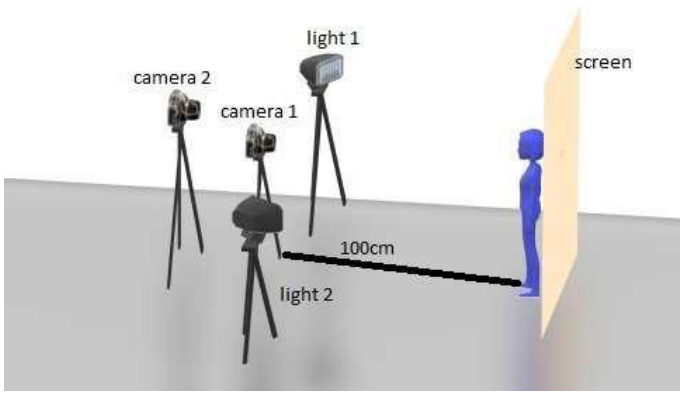
This work introduces a new audio-visual speech database in the Malayalam language. The database is recorded by native speakers in Kerala living in northern Kerala. The visual speech database is mainly a female-oriented database by keeping in mind about the low contrast between the lip and skin colour tone in the Indian context and the presence of facial hairs in major male population in Kerala. But in order to avoid gender discrimination,

5 male speakers are included in the visual speech database which carries the visual complexity like facial hairs partially covering the lip region. The language material consists of 50 Malayalam isolated phonemes and 207 connected words comprising of all allophonic variation in this language. This database is recorded in 3 phases with unique speakers in each phase. The first phase consists of audio-only speech database captured for implementing two task. The aim of the first task is to create a clear audio speech database recorded in the laboratory in a closed environment uttering 50 Malayalam isolated phonemes 10 times each by 10 male and 20 female speakers in the age group 21-25. The second task is chosen to study the effects of ageing in human speech organs by capturing 5 short vowel phonemes repeated 10 times each by 10 male and 10 female speakers in each age ranging from 5 to 60 (at present) age group which is recorded in acoustically realistic conditions. The second phase contains audio-visual speech database captured in an uncontrolled environment with an intention to capture the acoustically and visually realistic part of a speech recognition system. This phase contains 5 female speakers uttering 50 Malayalam isolated phonemes 10 times each with a complex background and also contain moving and multiple speakers in the background. The third phase is the audio-visual speech database recorded in a controlled environment. This phase contains 25 female and 5 male speakers uttering 50 Malayalam isolated phonemes and 207 connected words comprising of all allophonic variations 3 times each. In total, there are 30 female and 5 male speakers participate in the audio-visual speech database from the last two phases. The database is made available on the website *www.malayalamavspeechdb.in*. India has 23 constitutionally recognized official languages. Hindi and English are typically used as an official language by the Central Government. State governments use respective official languages. Malayalam is a Dravidian language spoken across Kerala, Lakshadweep, and Mahe spoke by 38 million people worldwide.

Designated a classical language status in 2013. Due to its lineage deriving from both Tamil and Sanskrit, the Malayalam alphabet has the largest number of letters among the Indian Language orthographies. The Malayalam language demands the speakers to speak all the vocal instruments inside the mouth along with the nose. It demands more frequent muscle movements and better air regulation inside the mouth. So for the people who speak the languages that require lesser of these would find it really difficult to catch up with the local Malayalam. Malayalam is a language which is used by masses and less accessed in computational point of view. So developing a speech based applications in the Malayalam language brings the benefit to enjoy the recent technological revolution in the native language rather than in English and placing India along with the developed countries.

**Recording Setup**

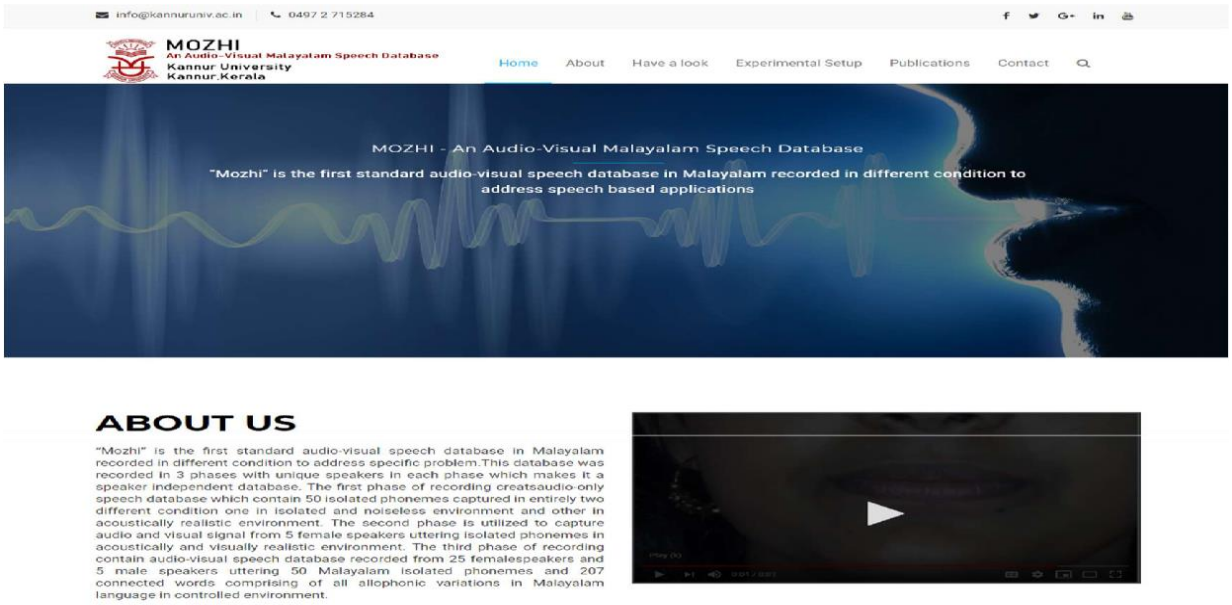
The database entitled "MOZHI" consists of 3 phase of recording. A single audio-only speech database acquisition phase and two audio-visual speech database acquisition phase. Recording in each phase begins with an explanation of the objectives to be achieved. During each utterance, the speakers are advised to close their mouth in the beginning and end of each utterance. The major location in the database is a lab environment in a college which have a provision to capture the audio and video samples in both controlled and uncontrolled condition. Thereby enriching the database for implementing in alternate research works in the Malayalam language. Fig. 1 shows the recording setup of this database.



The first phase of recording consists of audio-only speech database captured in two different conditions for implementing two task. The first task is captured in an acoustically isolated lab environment using a standard headphone with mic (with foam cover) located near the mouth region enabling a good quality audio acquisition. The language material is provided in a printed form (which contain 50 Malayalam isolated phonemes and its corresponding name in English for saving the audio file) and allows the speakers to repeat each utterance 10 times and save as a single wav file for each utterances. Speakers are requested to repeat the utterance if they felt it necessary, either because of a mistake during articulation or if the recorded sound contain noise due to respiration or channel problem. The age group utilized for this task was 21-25 belonging to post-graduate students. The second task is recorded in an acoustically realistic environment (office, school and house) using an ordinary mobile headset with mic and a laptop monitored by an operator. The speakers are allowed to repeat each utterance 10 times and is saved and monitored by the operator since the age group utilized for this task is 5-60. This database is chosen to study the effects of ageing in human speech organs and also to investigate better noise robust techniques in speech based applications.

**Accessibility of Database**

The database has been available through the servers of Kannur University, Kerala, India at the web address: http://pgcap.kannuruniversity.ac.in/mozhi/index. This database is used for research purpose only and will be provided only through request. The user is compelled to accept the licence agreement. The snapshot of the website is shown in fig. 2.




**Summary**

A new multimodal Malayalam audio-visual speech database is developed and presented. This database is recorded in 3 phases with unique speakers in each phase which makes it a speaker independent database. The first phase of recording creats audio-only speech database which contain 50 isolated phonemes captured in entirely two different condition one in isolated and noiseless environment by 20 female and 10 male speakers and other in acoustically realistic environment by 10 female and 10 male speakers in the age group 5 to 60. The second phase is utilized to capture audio and visual signal from 5 female speakers uttering isolated phonemes in acoustically and visually realistic environment. The third phase of recording contain audio-visual speech database recorded from 25 female and 5 male speakers uttering 50 Malayalam isolated phonemes and 207 connected words comprising of all allophonic variations in controlled environment.. This database can be used for different speech based applications like uni-modal

and bi-modal speech recognition, person identification, lip synchronization and lip tracking in Malayalam context.

**Reference**


[01] http://www.cmltemu.in/phonetic/#/

[02] https://www.britannica.com/topic/allophone

[3] Mattheyses, Wesley, and Werner Verhelst. "Audiovisual speech synthesis: An overview of the state-of-the-art." Speech Communication 66 (2015): 182-217.

[4] 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.

[5] Chourasia, Vishal, K. Samudravijaya, and Manohar Chandwani. "Phonetically rich Hindi sentence corpus for creation of speech database." Proc. O-Cocosda (2005): 132-137.

[06] Shrishrimal, Pukhraj P., Ratnadeep R. Deshmukh, and Vishal B. Waghmare. "Indian language speech database: A review." International Journal of Computer applications 47, no. 5 (2012): 17-21.

[7] Balyan, Archana. "Resources for Development of Hindi Speech Synthesis System: An Overview." Open Journal of Applied Sciences 7, no. 06 (2017): 233.

[08] Kurian, Cini. "A Survey on Speech Recognition in Indian Languages." International Journal of Computer Science and Information Technologies 5, no. 5 (2014): 6169-6175.

[9] Upadhyaya, Prashant, Omar Farooq, Priyanka Varshney, and Amit Upadhyaya. "Enhancement of VSR using low dimension visual feature." In Multimedia, Signal Processing and Communication Technologies (IMPACT), 2013 International Conference on, pp. 71-74. IEEE, 2013.

[10] Biswas, Astik, P. K. Sahu, Anirban Bhowmick, and Mahesh Chandra. "Audio Visual Isolated Oriya Digit Recognition Using HMM and DWT." In Proceedings of the Conference on Advances in Communication and Control Systems, pp. 234-238. 2013.

[11] Borde, Prashant, Ramesh Manza, Bharti Gawali, and Pravin Yannawar. "'vVISWa'–A Multilingual Multi-Pose Audio-Visual Database for Robust Human-Computer Interaction." International Journal of Computer Applications 137, no. 4 (2016).

[12] Kandagal, Amaresh P., and V. Udayashankara. "Visual Speech Recognition Based on Lip Movement for Indian Languages." International Journal of Computational Intelligence Research 13, no. 8 (2017): 2029-2041.