International Journal of Speech Technology
https://doi.org/10.1007/s10772-019-09655-0

Viseme set identification from Malayalam phonemes and allophones

K. T. Bibish Kumar[1] · R. K. Sunil Kumar[2] · E. P. A. Sandesh[3] · S. Sourabh[1] · V. L. Lajish[3]

Received: 31 March 2019 / Accepted: 24 October 2019

Abstract
Knowledge about phoneme and viseme in a language is a vital component in the making of any speech-based applications in that language. A phoneme is an atomic unit in an acoustic speech that can differentiate meaning. Viseme is the equivalent atomic unit in the visual realm which describes distinct dynamic visual speech gestures. The initial phase of the paper introduces a many-to-one phoneme-to-viseme mapping for the Malayalam language based on linguistic knowledge and data-driven approach. At the next stage, the coarticulation effect in the visual speech studied by creating many-to-many allophone-to-viseme mapping based on the data-driven approach only. Since the linguistic history in the visual realm was less explored in the Malayalam language, both mapping methods make use of K-mean data clustering algorithm. The optimum cluster determined by using the Gap statistic method with prior knowledge about the range of clusters. This work was carried out on Malayalam audio-visual speech database created by the authors of this paper with consist of 50 isolated phonemes and 106 connected words. From 50 isolated Malayalam phonemes, 14 viseme were linguistically identified and compared with results obtained from a data-driven approach as whole phonemes and consonant phonemes. The many-to-many mapping studied as a whole allophone, vowel allophones, and consonant allophones. Geometric and DCT based parameters are extracted and examined to find the parametric phoneme and allophone clustering in the visual domain.

Keywords Phonemes · Allophones · Visemes · K-mean clustering · Gap statistic

1 Introduction

Speech is bimodal; that is the most frequent communication system between humans and involves the understanding of the auditory and visual channels. The contribution of the visual part in the judgment of speech, especially in a noisy environment, is a fact. The visible organs of the articulatory system of human speech production consist of upper and lower lips, teeth, tongue, and lower jaw. The lips, tongue, and jaw will be the actively visible articulators used in language production. Analyzing the most dynamic active visible articulator, the lips, is the most crucial component in the visual speech analytics framework for recognition and synthesis.

Phonemes in a speech would be the nuclear sound units necessary to symbolize all words in that speech. On the other hand, the visual equivalent of a phoneme has several features which require a comprehensive study of this phoneme-to-viseme mapping region. For many years of study in visual language, it has gained considerable alterations in its definition. A viseme can contemplate regarding articulatory gestures such as mouth opening, teeth, and tongue vulnerability

✉ K. T. Bibish Kumar
bibishkrishna@gmail.com

R. K. Sunil Kumar
seuron74@gmail.com

E. P. A. Sandesh
sandeshepa@gmail.com

S. Sourabh
sourabhsuresh5@gmail.com

V. L. Lajish
lajishvl@gmail.com

[1] Computer Speech & Intelligence Research Centre, Department of Physics, Govt. College, Madappally, Vadakara, Calicut, Kerala 673102, India

[2] School of Information Science and Technology, Kannur University, Kannur, India

[3] Department of Computer Science, University of Calicut, Calicut, Kerala 673 635, India

Published online: 04 November 2019