Available online at www.sciencedirect.com

ScienceDirect

Procedia Computer Science 93 (2016) 668 – 675

Procedia Computer Science

6th International Conference On Advances In Computing & Communications, ICACC 2016, 6-8 September 2016, Cochin, India

Neural network based Gujarati Speech Recognition for dataset collected by in-ear microphone

Desai Vijayendra A[a,*], Dr. Vishvjit K. Thakar[b]

[a]*Electronics and Communication dept,C.K.Pithawalla College of Engg and Tech,Surat, Gujarat , India*
[b]*Electronics and Communication Dept, A.D. Patel Institute of Technology, New vallabh Vidhyanager, Gujarat India*

Abstract

This paper addresses different configurations of two layers and three layers neural network approach for the low resource language like Gujarati. The speech data are collected with the in-ear microphone compare to conventional microphone system and results are compared. Different end point detection algorithms are also tested to remove an unwanted silence portion where maximum chances of noise take place. Word boundary detection is used to separate out the different words form sentences. Detected words are then passed to the feature extraction block. Feature extractions are done with the help of the Mel-Frequency Cepstral Coefficients (MFCCs) and Real Cepstral Coefficients (RC). Results are tested and compared to them. Two layers and three layers neural networks approach are used for the classification.

Peer-review under responsibility of the Organizing Committee of ICACC 2016

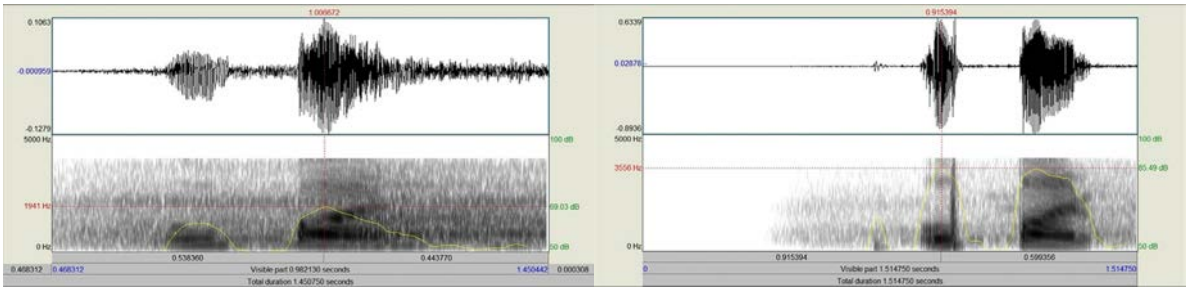
Keywords: Gujarati Speech recognition, low resource languages, speech recognition, in-ear microphone, End point detection, Energy Entropy feature, two layers and three layers neural networks, MFCC, conjugate gradient (CG), Levenberg-Marquardt (LM)

1. INTRODUCTION

In today's world, language is one of the few barriers that hinder the human to human interaction. An event such as armed conflict or natural disasters might create a situation to communicate with speaker of less prevalent languages. So it is very important and useful to develop a speech recognition system for low resource languages. Various applications of local language speech recognition are agriculture, automatic telephone system, voice operated services. The creation of language and acoustic resources, for any given spoken language, is typically a costly task. For example, a large amount of time and money is required to properly create annotated speech corpora for automatic speech recognition (ASR), domain-specific text corpora for language modeling (LM), etc. Very limited amount of work had been done

* Desai Vijayendra A. Tel.: +91-9408118010.
E-mail address: vijayendra.desai@ckpcet.ac.in

Peer-review under responsibility of the Organizing Committee of ICACC 2016
doi:10.1016/j.procs.2016.07.259



(a) By keeping microphone inside the ear (b) By keeping microphone outside the mouth

Figure 2.1: Waveform and spectrogram for word upper spoken

in the Gujarati language. So in our work, we implemented the Gujarati speech recognition system using different combinations of two and three layer neural. In our work we have generated database of the speakers of various age and gender. A vocabulary consisting of ten isolated Gujarati words recorded as follows: ડાબી,જમણી,ઉપર,નીચે,આગળ,પાછળ,આજુ,બાજુ,આમ,તેમ. For speech input, we have used in-ear microphone concept rather than typical microphone used for speech collection. Results show that, it improved the recognition rate very much. Different methods of the word boundary detection are used. For feature extraction and classification different approaches are used and results are compared. The results are tested and compared with the English words like up,down,left,right,pan and move.

2. IN-EAR MICROPHONE

Most of the speech recognition system collects speech data by placing a microphone outside the mouth. In many real time application speech data required to collect in the noisy environment such as agricultural, industrial, office environments, on the road. In such conditions, if speech data collected from the mouth then it may pick-up such surrounding noise also. In this kind of scenarios speech collected from the ear can give a better quality speech. Different ways of speech collection other than mouth are tried previously also [2-6]. In this work I have used in-ear microphone to collect the speech data. Following results suggest that in-ear microphone is better then conventional speech data collection method. From spectrogram results using [PRRAT] shown in figure 2.1a and 2.1b. We can observe that, speech data collected using in-ear microphone, concentrated up to 2.3 kHz only, while it will go up to 4 kHz for speech collected outside mouth. Muscles and tissues provide an attenuation effect to higher frequencies so it behaves like low pass filter [2]. From figure 2.2a and 2.2b, we can observe that typical noise hum will go below -90db at around 200 Hz for in-ear microphone data, which will be continue till 1.25 kHz for speech recorded by keeping the mic outside mouth. So, we can say that speech is less affected by noise in case of in-ear microphone. For the testing speakers of different ages and different genders are selected. For analysis ten different words with rich phonetic meanings are used, as shown in table 2.1. The results are compared to the data collected from inside ear and outside ear.

3. WORD BOUNDARY DETECTION.

On-the fly speech activity detection remains a challenging problem in real-life applications where the lowest-energy portions of the speech signal, such as weak fricatives for example, fall under the energy level of the background noise or silence. In many cases, sensitive microphone easily picked up speaker generated artifacts such as a heavy breath release, coughing, some clicks, and surrounding low frequency noises. Speaker-generated artifacts not necessarily shorter in duration and energy than the speech segments were, which made the speech detection phase quite challenging. Figure 3.1a and 3.1b shows these kind of cases. The scheme applied here is, a two-step search algorithm. First, the short-time average magnitude quantity