Speech Recognition Model for Tamil Stops

A.Rathinavelu, G.Anupriya, A.S.Muthanantha murugavel

Department of Computer Science & Engineering,
Dr.Mahalingam College of Engineering and Technology, Pollachi, Tamilnadu, India
starvee@yahoo.com , anupriya_rajkumar@yahoo.co.in , as_ananth2k1@yahoo.com

Abstract—**In this paper, a novel approach for implementing Tamil isolated speech phoneme recognition is described. While most of the literature on Automatic Speech Recognition (ASR) is based on Hidden Markov Models (HMM) and other approaches, our system is implemented using Feedforward neural networks (FFNet) with backpropagation algorithm. Our model consists of two modules, one is for neural network training and another one is for Visual Feedback. The speech corpus is developed from ten children (5 boys and 5 girls) in the age group of 4-7 years for Tamil stops. The system has been trained with the speech corpus of 20 Tamil phonemes. This study includes the Visual Feedback module to respond to the utterance of children in front of Automatic Speech Recognition model.**

Index Terms—**backpropagation algorithm, Tamil stops, neural networks, speech recognition**

I. Introduction

Speech is human's most efficient mode of communication. Beyond efficiency, humans are comfortable and familiar with speech. Other modalities require more concentration, restrict movement and cause body strain due to unnatural positions. Research work on Tamil speech recognition, although lagging than other languages, is becoming more intensive than before and several researches have been published in the last few years. In spoken language, a phoneme is a basic, theoretical unit of sound that can change the meaning of a word. A phoneme may well represent categorically several phonetically similar or phonologically related sounds (the relationship may not be so phonetically obvious, which is one of the problems with this conceptual scheme). Depending on the language and the sounds used, a phoneme may be written consistently with one letter; however, there are many exceptions to this rule. The range of the possible applications is wide and includes: voice-controlled appliances, fully featured speech-to-text software, automation of operator-assisted services, and voice recognition aids for the handicapped [1]. In this paper we present the work on articulatory acoustic data analysis and training of a Neural Network model for recognition of Tamil Phonemes. Acoustic study on Tamil laterals, trills and fricatives was carried out with Normal Hearing and Hearing Impaired children [2]. The paper describes the testing of the trained network through Speech User Interface. Finally we present the results of the speech recognition system. The visual articulatory model is useful for training and improving the articulation of people who are not familiar with the articulation of phonemes.

II. Neural Network Model

A. Feedforward Neural Network (FFNet)

Neural Networks can be used in general-purpose applications like speech recognition. It can handle low quality, noisy data and speaker independence and can achieve greater accuracy based on more training. A feedforward neural network is a biologically inspired classification algorithm. It consists of a (possibly large) number of simple neuron-like processing units, organized in layers. Every unit in a layer is connected with all the units in the previous layer. These connections are not all equal, each connection may have a different strength or weight. The weights on these connections encode the knowledge of a network. Often the units in a neural network are also called nodes. Data enters at the inputs and passes through the network, layer by layer, until it arrives at the outputs. During normal operation, that is when it acts as a classifier, there is no feedback between layers [3].

B. The Learning Phase

During the learning phase the weights in the Feedforward network will be modified. All weights are modified in such a way that when a pattern is presented, the output unit with the correct category, hopefully, will have the largest output value. The Feedforward network uses a supervised learning algorithm: besides the input pattern, the neural net also needs to know to what category the pattern belongs. Learning proceeds as follows: a pattern is presented at the inputs. The pattern will be transformed in its passage through the layers of the network until it reaches the output layer. The units in the output layer all belong to a different category. The outputs of the network as they are now are compared with the outputs as they ideally would have been if this pattern were correctly classified: in the latter case the unit with the correct category would have had the largest output value and the output values of the other output

units would have been very small. The differences between the actual outputs and the idealized outputs are propagated back from the top layer to lower layers to be used at these layers to modify connection weights. This is why the term backpropagation network is also often used to describe this type of neural network [3].

C. The Testing Phase

In the testing phase the weights of the network are fixed. For classification the feedforward network and a pattern is needed [3]. A pattern, presented at the inputs, will be transformed from layer to layer until it reaches the output layer. Now classification can occur by selecting the category associated with the output unit that has the largest output value.

III. Articulatory Acoustic Data

Table 1. *Speech parameters for Tamil Stops [5]*

Tamil Stops	Active Articu -lator	Passive Articu -lator	Constr -uction Degree	Status of Glottis	Air -stream Mech -anism
ka (க)	Tongue Body	Velum	Stop	Voice -less	Normal
ga (க)	Tongue Dorsum			Voiced	
ta (ட)	Tongue Tip			Voice -less	
da (ட)				Voiced	
tha (த)				Voice -less	
dha (த)				Voiced	
pa (ப)	Lower Lip	Upper Lip		Voice -less	
ba (ப)				Voiced	

Training input, articulatory acoustic data for 20 Tamil phonemes such as 4 stops (k, t, th and p) with a, e, i, o ,u combination (ka, ki, kai, ko, ku, ta, ti, tai to, tu, pa, pi, pai, po, pu, tha, thi, thai, tho, thu) is collected from ten children (5 boys and 5 girls) in the age group of 4-7 years to train the system for speech recognition to form a small meaningful corpus. Table 1 shows speech parameters for the collected Tamil stops. The following articulatory acoustic data analysis was performed, to aid in the development of a speech recognition system for Tamil stops. We collected the average first five formant frequencies (F0, F1, F2, F3 and F4) of 4 Tamil stops (ka, ta, pa and tha) from the Normal Hearing Children's (5 Boys and 5 Girls) speech for training.

IV. Speech Recognition System

The general scheme of the Tamil Phoneme Recognition System is depicted in Figure 1 and Figure 2. It consists of three parts: preprocessing phase, classification phase and visual feedback through speech user interface.

A. Preprocessing Phase

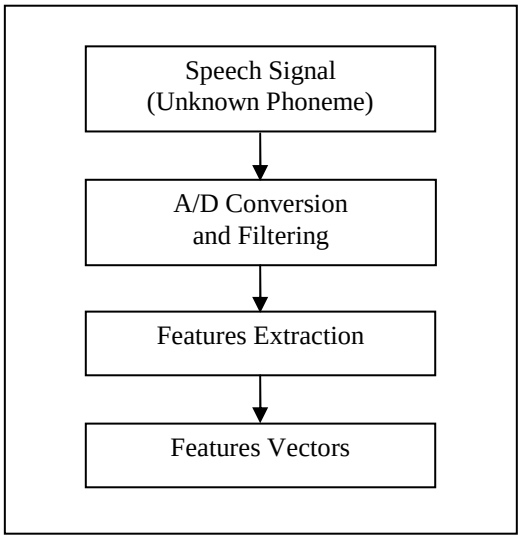

Figure 1: *Steps in Preprocessing Phase*

Table 2. *Formant values (in Hertz) for Tamil Phonemes*

Tamil Phoneme	IPA	F0	F1	F2	F3	F4
க	Ka	364	1125	1606	2530	3740
கி	Ki	389	499	1988	3204	3797
கை	Kai	348	860	2016	3033	3751
கொ	Ko	379	818	1445	2666	3510
கு	Ku	384	625	1288	2519	3564
ட	Ta	352	1089	1639	2638	3719
டி	Ti	379	503	1989	3191	3743
டை	Tai	362	840	2081	3040	3723
டொ	To	373	787	1475	2642	3590
டு	Tu	379	606	1291	2579	3633
ப	Pa	355	1066	1600	2599	3709
பி	Pi	384	453	2060	3174	3743
பை	Pai	362	817	1997	3017	3723
பொ	Po	361	763	1399	2608	3494
பு	Pu	377	598	1185	2563	3632
த	Tha	385	1095	1657	2441	3731
தி	Thi	436	535	1903	2968	3598
தை	Thai	415	853	1829	2734	3558
தொ	Tho	417	883	1515	2492	3589
து	Thu	419	624	1368	2422	3575

Figure 1 shows the steps in preprocessing phase in the implementation of the speech recognition system. Speech signal is digitized and also the important frequency component in the signal is filtered using Praat. Praat is a computer program which is used to analyze, synthesize, and manipulate speech. Currently, features such as first five formant values for the collected 20 Tamil sounds are extracted and analyzed using Praat. Table 2 shows the extracted speech features such as first five formant values (in Hertz) for 20 Tamil phonemes. Formants are the dominant acoustic components which determine the sound quality of particular vowels. They are formed by the air flowing through the vocal tract, and vibrating at different bands of frequencies as it responds to changes in the tract's shape.

B. The Classification Phase

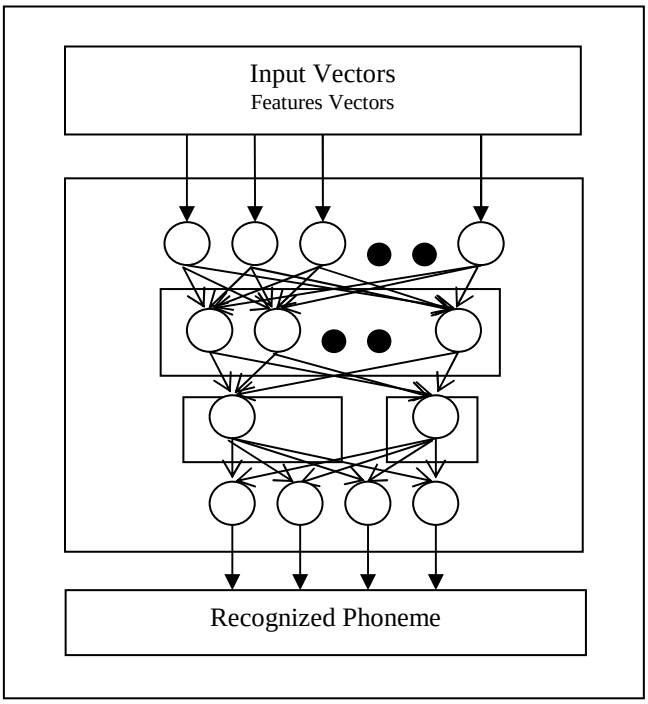

Figure 2: Steps in Classification Phase

Figure 2 shows the steps in classification phase in the implementation of the speech recognition system. The classification phase can be implemented using Feedforward Neural Network with Backpropagation Algorithm. Feed-forward neural network allow signals to travel one way only; from input to output. There is no feedback. The output of any layer does not affect that same layer. Feed-forward neural network tend to be straight forward networks that associate inputs with outputs. This type of organization is also referred to as bottom-up or top-down. Backpropagation is a systematic method for training multi-layer artificial neural networks. It has a mathematical foundation that is strong if not highly practical.

It is a multi-layer forward network using extend gradient-descend based delta-learning rule, commonly known as backpropagation (of errors) rule. Backpropagation provides a computationally efficient method for changing the weights in a feedforward network, with differentiable activation function units, to learn a training set of input-output examples. The aim of this network is to train the net to achieve a balance between the ability to respond correctly to the input patterns that are used for training and the ability to provide good responses to the input that are similar. Features such as first five formant values constitute the input vectors to the feedforward neural network used as classifier.

For Tamil phoneme recognition feedforward neural network with two hidden layers is designed using MATLab. Multiple layers of neurons with nonlinear transfer functions allow the network to learn nonlinear and linear relationships between input and output vectors. For training the network sigmoid transfer function-tansig is used. The neural network has been designed with 5 X 200 neurons corresponding to the first five formant values extracted from 200 samples (20 phonemes from 10 subjects). There are two hidden layers in the network. The first hidden layer contains 40 neurons and the second hidden layer contains 20 neurons. The output layer is designed with 20 neurons to classify 20 Tamil phonemes and it produces the values in the range -1 to +1. If the target value is 1 then the corresponding phoneme is recognized, otherwise the phoneme is not recognized. The network is trained by:

1. Propagating inputs forward in the usual way,
   - i.e. All outputs are computed using sigmoid thresholding of the inner product of the corresponding weight and input vectors.
   - All outputs at stage *n* are connected to all the inputs at stage *n*+1.
2. Propagating the errors backwards by apportioning them to each unit according to the amount of this error the unit is responsible for.

C. Visual Feedback through Speech User Interface

The trained model can be used for Speech Recognition, and visual feedback is given through the Speech User Interface. Figure 3 shows the speech user interface which has the options like record, play, save, load and recognize. Through the speech user interface the stored or recorded phoneme can be given as input to the trained network. . Before testing a phoneme for recognition, the speech features such as first five formant values are extracted using Praat. The error margin has been fixed as $1e^{-2}$ and the maximum number of epochs is 300. If the phoneme is successfully recognized, the Speech User Interface reports the results of the recognition with an accuracy rating of good, fair or poor.

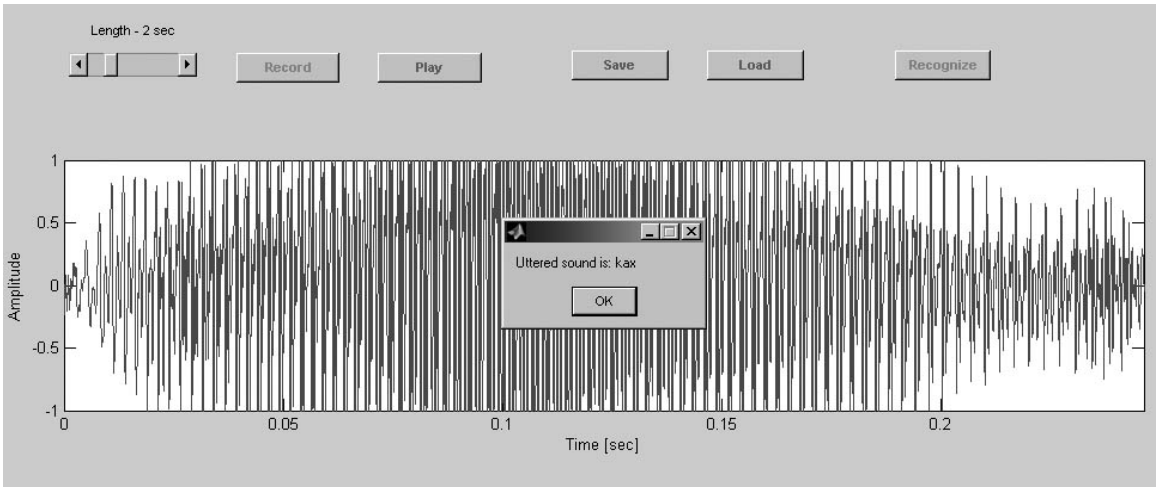

Figure 3: *Speech User Interface for Tamil stops recognition*

V. Results And Discussion

The test data for the system has been acquired from 10 subjects (5 boys and 5 girls in the age group of 4-7 years) for 20 Tamil phonemes. The Recognition % for each phoneme is calculated by taking the ratio of the number of successful recognitions to the total number of test samples. Table 3 shows the results of 20 Tamil phonemes from the experiments conducted. The low rate of recognition for some of the phonemes may be due to the misarticulation of the children or due to the pronunciation similarities of Tamil phonemes. A visual articulatory model may be used to improve the articulation of the children [4, 5].

Table 3. *Recognition Results*

Tamil Phoneme	IPA	Recognition %
க	Ka	90
கி	Ki	90
கை	Kai	80
கொ	Ko	70
கு	Ku	80
ட	Ta	70
டி	Ti	80
டை	Tai	80
டொ	To	90
டு	Tu	80
ப	Pa	70
பி	Pi	90
பை	Pai	90
பொ	Po	70
பு	Pu	70
த	Tha	80
தி	Thi	80
தை	Thai	90
தொ	Tho	90
து	Thu	80

VI. Future Work And Conclusion

Currently development of speech recognition is widely used in industrial software market. The main contribution of the proposed speech recognition system is to recognize the Tamil phonemes properly and respond with Visual Feedback through Speech User Interface based on the accuracy level of the user's speech. A Neural Network model has been designed to recognize 20 Tamil phonemes based on the training input. The trained network has been tested through Speech User Interface. An average accuracy level of 81% has been achieved in the experiments conducted using the trained neural network. The number of phonemes for recognition can be increased by using the feedforward neural network model with backpropagation algorithm and the accuracy level can also be further improved by giving more training. Our future work will focus on gender selection in the speech user interface, option for more phonemes and large corpus size. This preliminary study will help us to develop Automatic Speech Recognition system for Normal Hearing and Hearing Impaired children by providing visual feedback to improve their articulation.

References

[1] El Choubassi, M.M.; El Khoury, H.E.; Alagha, C.E.J.; Skaf, J.A.; Al-Alaoui, M.A., "Arabic speech recognition using recurrent neural networks", Signal Processing and Information Technology, 2003. ISSPIT 2003. Proceedings of the 3rd IEEE International Symposium, 14-17 Dec. 2003, Page(s): 543 – 547.

[2] Rathinavelu, A., Hemalatha, T., Savithri, S.R., and Chitra, R. Interactive multimedia tool to help vocabulary learning of hearing impaired children by using 3D VR objects as visual cues. Nat J Technol 2006; 2(1), Page(s): 25-32.

[3] http://www.praat.org © May, 2004.

[4] Rathinavelu.A, et al: Computer aided articulation tutor using three dimensional visual cues for the child with hearing loss Journal of Computer Science, 2(1), 2006, Page(s): 76-82.

[5] Rathinavelu, A., Hemalatha, T, and Anupriya, R., Three dimensional Articulatory Model for speech acquisition by children with hearing loss (To appear in Springer LNCS and selected for HCI 2007, China).