CLASSIFICATION OF ARTICULATION POSITION USING NEURAL NETWORKS

Paulraj M P[1], Mohd Shukry Abdul Majid[1], Sazali Yaacob[1], T.Manigandan[2]
R.Pranesh Krishnan[1]

[1] *School of Mechatronic Engineering, Universiti Malaysia Perlis (UniMAP), Perlis, Malaysia - 02600.*

[2] *Department of Electrical and Electronics Engineering, Kongu Engineering College,*
Perundurai, Tamilnadu, India -638 052.

manigandan_t@yahoo.com, praneshkrishnan@gmail.com

Abstract: *Conventional speech recognition systems classify speech sounds according to their phonemes. Phoneme is the basic unit of speech, which if changed would change the whole meaning of the word. A number of research works have already been done in the conventional systems. Articulation refers to the manner of production of sound and the placement of tongue, lips and teeth. Not all the possible combinations of tongue and lip positions are used in speech. There may be incorrect placement of lips, teeth, tongue or even soft palate during speech. This often affects the clarity of speech. This paper is intended to classify the place of articulation of English words for the English language spoken by the Malaysians. In this approach the articulatory acoustic features are considered for the classification. Difficulty in pronunciation happens because of voice disorders. Articulation disorders constitute the most numerous of all speech disorders. About 3 out of 5 of all speech and language disorders are related to articulatory problems. This system is proposed to be used as an aid in the diagnosis of articulatory disorders.*

Key words: *Place of Articulation, Articulation disorders, Articulatory Acoustic features, Neural Network*

1.0 INTRODUCTION

This paper investigates the use of articulatory-acoustic features for the classification of articulation position using neural networks. Place of articulation is defined in terms of the articulators involved in the speech gesture. It is common to refer to a speech gesture in terms of articulators. Articulation is the process by which sounds, syllables, and words are formed when the tongue, jaw, teeth, lips, and palate alter the air stream coming from the vocal folds. A person has an articulation problem when he or she produces sounds, syllables or words incorrectly so that listeners do not understand what is being said or pays more attention to the way the words sound than to what they mean. Articulation disorders are speech sound errors that do not change in different word contexts. These errors occur during the production of isolated speech sounds (phonemes) and are thus misarticulated at the syllable and word levels as well.

1.1 PRODUCTION OF VOWELS

Vowels are more difficult to describe accurately than consonants. This is largely because there is no noticeable obstruction in the vocal tract during their production. It is not easy to feel exactly where the vowel sound is made. The only reliable way of observing vowel production is using x-ray photography. But this is not only expensive; it is also dangerous and could not be carried out each time one wanted to describe a particular vowel.[1] So the vowel place of articulation is predicted based on the articulatory gestures (movements of the lips and tongue).

Tongue is a very important articulator and it can be moved into many different places and different shapes. It is usual to divide the tongue into different parts, though there are no clear dividing lines within the tongue [1]. Vowels produced with the highest point of the hump in the tongue close to the roof of the mouth are said to be HIGH and those produced with the highest point of the hump in the tongue barely rising above the floor of the mouth are said to be the LOW; the intermediate position is referred to as MID [1]. Depending on the location of the highest point of the tongue, vowels may be regarded as FRONT or BACK. Lips are important in speech. The quality of a vowel is affected by the shape of the lips.

2.0 ARTICULATION POSITION

Table 1: Articulation Position Representation based on Tongue Position, Tension and Lip Position

Tongue Position	Tongue Tension	Lip Position	
		Front	Back
High	Tense	*'beet'/iy/*	*'boot'/uw/*
	Relax	*'bit'/ih/*	*.'book'/uh/*
Med	Tense	*'bait'/ey/*	*.'boat'/ow/*
	Relax	*'bet'/eh/*	*but'/ah/, 'born'/ao/*
Low	.	*'bat'/ae/*	*'bob'/aa/*

Figure 1 shows the phonetic properties of the vowels.

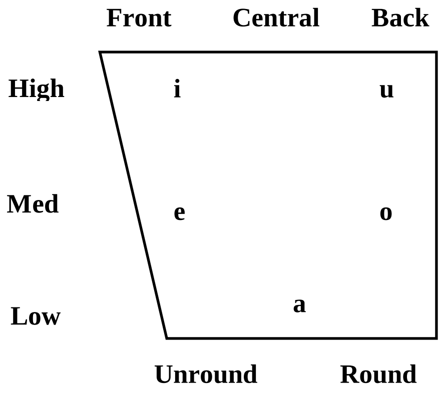

Figure 1: Phonetic Properties of the English Vowels

3.0 SPEECH DATABASE

In this paper the speech utterances of 16 speakers of three different races were considered. Speech samples uttered were collected from various speakers using a digital microphone connected to a laptop. The voice samples were recorded in a common meeting room at 44.10 kHz sampling frequency rate and 16 bits per sample using MATLAB. The conditions such as stuttering, stammering, hoarseness were monitored and each word was recorded for several attempts and the best three were picked. As the usual condition in meeting rooms or offices, noise sources are computer fans and air conditioning systems and moderate reverberation are present. Necessary steps were taken to avoid noise from these sources. An average unavoidable background noise level of 26 dB was maintained throughout the recording. All Speakers were asked to read out as set of diphthongs words. At the beginning of each recording session, a few seconds of the silence (background noise) are recorded for the background noise calculation.

3.1 FEATURE EXTRACTION

In this work 27 feature coefficients were evaluated and are intended for the articulation classification. The speech samples captured are pre emphasized and then grouped into frames of 256 samples. Each frame is hamming windowed, Fourier transformed and the Energy is calculated. The mean, standard deviation and the kurtosis are then computed from the Timbral and time domain features: Spectral Entropy, Spectral Centroid, Spectral Flux and the Zero Crossing Rate.

Pre-emphasis: Preemphasis refers to the process of increasing the magnitude of higher frequencies with respect to the magnitude of lower frequencies in order to improve the overall signal-to-noise ratio by minimizing the adverse effects of such as attenuation distortion or saturation of recording media. The captured voice signals are preemphasized to improve the signal to noise ratio.

In this paper, the signal S(t) was first pre-emphasized (a high pass filtering) using the following equation[3]

$$S_{pre}(t) = S(t) - \alpha S(t-1) \qquad ..(1)$$

where the value of α is chosen as 0.95.

Signal Processing Techniques [4]

The following timbral and time domain features are used in this paper.

- Spectral Entropy
- Spectral Centroid
- Spectral Flux
- Zero Crossing Rate

Spectral Entropy is a kind of Information Entropies which shows the spectral complexity of signal. The spectral entropy is calculated as

$$H = -\sum_{k=1}^{N} p_k \log p_k \qquad ..(2)$$

where p_k is the probability density function (PDF)

Spectral Centroid: It is a measure used in digital signal processing to characterize an audio spectrum. It indicates where the "center of mass" of the spectrum is present in the spectrum. It is calculated as the weighted mean of the frequencies present in the signal, with their magnitudes as the weights.

$$Centroid = \frac{\sum_{n=0}^{N-1} f(n)x(n)}{\sum_{n=0}^{N-1} x(n)} \qquad ..(3)$$

where *x(n)* represents the magnitude of bin number *n*, and *f(n)* represents the center frequency of that bin.

Spectral Flux: This feature measures frame-to-frame spectral difference. Thus, it characterizes the changes in the shape of the spectrum which is calculated by comparing the power spectrum for one frame against the power spectrum from the previous frame.

Zero Crossing Rate: Zero Crossings is the number of time domain zero voltage crossings with a speech frame. It is an indicator of the frequency at which the energy is concentrated in the signal spectrum.

$$Z(m) = \frac{1}{2}\left\{\sum_{n=0}^{N-1} \left| \mathrm{sgn}[s_w(n)] - \mathrm{sgn}[s_w(n-1)] \right| \right\} \quad ..(4)$$

where Z(m) is the Zero crossing rate of the mth frame.

4.0 NEURAL NETWORK

Neural Networks are the commonly used tool for the classification problems related to speech recognition. A backpropagation neural network is a multilayer, feedforward neural network with an input layer, an output layer and a hidden layer. The neurons in the hidden and output layers have biases (similar to weights) which are connections from units whose output is always 1[2].

4.1 THREE LAYER FFNN

An FFNN consists of three layers, namely input layer, hidden layer and the output layer as shown in Figure 2. A supervised learning method is employed in the FFNN, in which the calculated output is compared with the target value and then the error is fed back to the subsequent layers to modify the weights. The usage of bias neurons enhances the convergence process. The specified parameters such as learning rate and momentum factor control the change in weight. A tolerance level is fixed as the stopping condition for the training process.

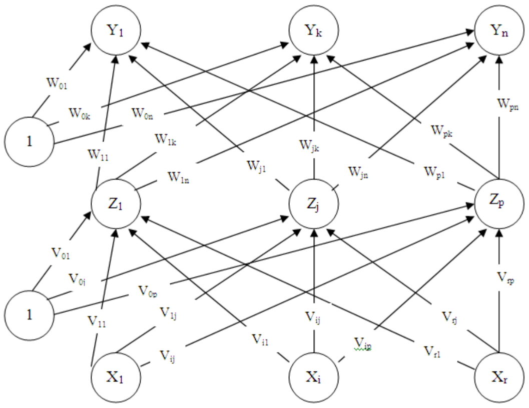

Figure 2. Three Layer Feed Forward Neural

The hidden and output layers have binary sigmoidal activation function, which are the activations of the corresponding neurons. The selection of an activation function for a specific application is a major criterion in achieving good performance of the BP algorithm[2].

Table 2: Data Classification

Activation Function	$[1+\exp(-x/q)]^{-1}$
Learning Rate	0.1
Momentum Factor	0.9
Testing Tolerance	0.2
Training Tolerance	0.01
Number of Samples for Training	263
Number of Samples for Testing	329
Percentage of Classification	82.0669%
Number of Input Neuron	27
Number of Hidden Neuron	25, 25
Number of Output Neuron	7
Number of Epoch	88

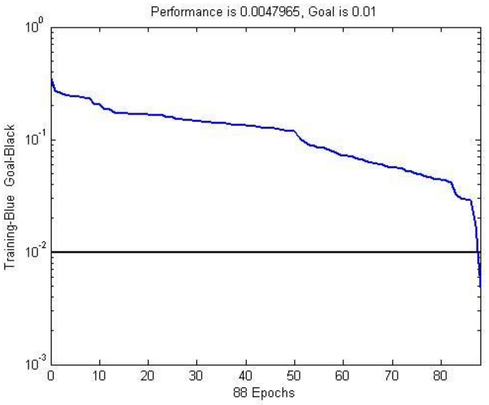

Figure 3. Cumulative Error Vs Epoch for Classification of Articulation Position.

5.0 RESULTS AND DISCUSSION

In the experimental study, the 11 classes of the voice sample classification are obtained from 10 individuals. Then the special features of the recorded sound waves are obtained. These coefficients are used as sample input patterns to the neural network. In this experiment the BPNN is a four-layer network with 27 input neurons, 25 neurons in each of the two hidden layers and 7 output neurons are considered. The activation function used for the hidden and the output neurons is binary sigmoidal function $[1+\exp(-x/q)]^{-1}$. The initial weights are randomized between -0.9 and +0.9. The cumulative error vs the epoch graph is shown in the Figure 3.

6.0 CONCLUSION

From Table 2 it can be observed that the timbral features namely Spectral Entropy, Spectral Centroid and Spectral Flux derived from the speech sample can be used to classify the articulation position which can be further used for the diagnosis of articulation disorders.

7.0 REFERENCES

1. Katamba, Francis (1989),“*An Introduction to Phonology”*. New York: Longman.

2. Sivanandam SN, Paulraj M (2003), “*Introduction to artificial neural networks”*. Vikas, Chennai.

3. O’Shaughnessy,D (1987),“*Speech Communication – Human and Machine”*, Addison Wesley.

4. E. Scheirer and M. Slaney, “Construction and evaluation of a robust multifeature speech/music discriminator,” in Proc. Int. Conf. Acoustics, Speech, Signal Processing (ICASSP), 1997, pp. 1331–1334.

5. Roach, Peter (2000), written at Cambridge, “*English Phonetics and Phonology: a Practical Course”*, Cambridge University Press.

6. Donald G. Kimber, Marcia A. Bush and Gary N.Tajchman, “*Speaker Independent Vowel Classification using Hidden Markov models and LVQ2*”. IEEE Transactions on Neural Networks, 1990.

7. Harris Drucker and John Preusse, “*Real time Recognition of Ten Vowel – Like Sounds in Continuous Speech*”, IEEE Transactions on Acoustics Speech and Signal Processing, February, 1994.

8. Osamu Fujimura, “*Syllable as a Unit of Speech Recognition*”, IEEE Transactions on Acoustics, Speech and Signal Processing, Vol, ASSP-23, No.1, February, 1975.

9. Nicolaos, B. Karayiannis, Anastasioos N. Venetsanpoulos, “*Fast Learning Algorithm for Neural Networks*”, IEEE Transactions on Circuits and Systems –II, Vol.39,No.7, July, 1992.

10. John Laver, “*Principles of Phonetics,*” Cambridge University Press, Great Britain, 1994.

11. Dimitris A. Karras, Stavros J.Perentonis., “*An Efficient Constrained Training Algorithm for Feed Forward Networks*”, IEEE Transactions on Neural Networks, Vol.6, No.6, November, 1995.

12. Yoshua Bengio and Renato De Mori, "*Use of Neural Networks for the Recognition of Place of Articulation*", IEEE, ICASP, S3.2, 1988, pp. 103-106.