PREPROCESSING AND NEURAL CLASSIFICATION OF ENGLISH STOP CONSONANTS [b, d, g, p, t, k][1]

A. Esposito, C. E. Ezin, M. Ceccarelli

International Institute for Advanced Scientific Studies (IIASS)
Via G. Pellegrino 19, 84019 Vietri Sul Mare (SA) Italy
Tel +39 89.761.167 Fax: +39 89. 761.189

ABSTRACT

Neural networks are accepted as powerful learning tools in pattern recognition in which they proved their performance. Nevertheless, many problems like phoneme classification with multi-speaker continuous speech database are hard even for Neural Networks. Our aim is to propose an Artificial Neural Network architecture that detects acoustic features in speech signals and classifies them correctly. We reached this goal with English stop consonants [b, d, g, p, t, k] extracted from the general multi-speaker database (TIMIT) by modifying some parameter values in the preprocessing algorithm and by using a modified TDNN (Time Delay Neural Network) architecture.
Our net performed a good classification giving as testing recognition percentage the following results: 92.9 for [b], 91.8 for [d], 92.4 for [g], 80.3 for [p], 90.2 for [t], 94.2 for [k].

1. INTRODUCTION

Phoneme recognition is a hard task because these small speech units are very variable and it is difficult to find features that remain stable and allow to discriminate among them in their corresponding acoustic signal. Neural Networks for phoneme recognition firstly were used by Waibel et al. in 1987. The results published were enough good but they were obtained using an elaborated network architecture and a small number of speakers and therefore a specific database. When researchers try to realize stop consonant recognition with general database (like TIMIT), the network performance decreases significantly [3]. With the present work, we faced the phoneme recognition problem again with new approaches. We used as database TIMIT and as preprocessing algorithm RASTA-PLP [1].

The paper is organized as follows. In the first section, we describe the data and the preprocessing performed on them. The second section describes some acoustic features of the phonemes we want to classify, whereas the third section describes the network architecture and the learning algorithm. Finally we present the experimental results in the fourth section.

2. SPEECH MATERIALS AND PREPROCESSING PROCEDURE

The continuous speech database used is extracted from the TIMIT database made up of English sentence-texts produced by speakers from different US regions.
Each speaker had read ten different English sentence-texts. Our data (the stop consonants [b, d, g, p, t, k]) are extracted from such sentences. The training data are produced by 38 speakers (24 males and 14 females) from the same US region (dr1 in TIMIT) whereas the testing data are obtained from 35 speakers (22 males and 13 females) coming from three different US regions (dr1-dr2-dr3 in TIMIT). **Table 1** summarizes, the number of phonemes used to train and test the net.

Phonemes	Training	Testing
[b]	183	176
[d]	300	265
[g]	166	157
[p]	211	188
[t]	329	326
[k]	389	352
Total	1578	1464

Table 1: Size of the training and testing data

The speech signal is preprocessed using the Rasta-PLP algorithm due to its performance over Linear Prediction and Perceptual Linear Prediction algorithms [1, 2]. Moreover, we modified some parameter values of the RASTA-PLP algorithm in order to capture the acoustic features of the stop phonemes. Indeed, the speech signal was sampled at 16 kHz; at each step of 5 msec rate, the speech segment is weighted by a Hamming window of 10 msec. In the original algorithm, the speech signal was sampled at 20 kHz and the Hamming window 20 msec long is moved over the speech signal every 10 msec.

The resulting waveform file is processed to produce nine acoustic features for each one of the stop consonants [b, d, g, p, t, k]. This vector of features has been used as input of the network.

This work has been supported by IIASS, ICSC-World Laboratory, INFM, Salerno

3. ACOUSTIC FEATURES OF STOP PHONEMES

English stop consonants [b, d, g, p, t, k] are divided into two classes: the voiced [b, d, g] and the voiceless [p, t, k]. Stop consonants are always preceded by a closure interval prior to the release and they are released with an explosive burst when produced before a vowel in a monosyllable. The silent interval is an essential cue for the identification of stop consonant [5].
There is little vowel transition following [p], [t], or [k] because most of the articulatory movements of the vocal apparatus to the vowel configuration occur during the VOT period [5]. Defined as the interval between the onset of the stop burst and onset of the vowel voicing, the Voice Onset Time (VOT) reliably discriminates between the voiced and voiceless stops [11] in initial position of isolated words and in short sentences.
Hovewer, in other phonetic environments, VOT values were found to be compressed for both voiced and voiceless stops, and the separation is less sharp [12]. The Rasta-PLP algorithm is capable of extracting from the raw speech signal such acoustic features and they can be appropriately used by our Neural Network in order to perform a good classification of [b, d, g, p, t, k].

Cole and Scott showed that the identification of a particular stop in each class involves identification of either invariant or transitional cues (Invariant cues are defined as acoustic cues which accompany a particular phoneme in any vowel environment.Transitional cues are defined as acoustic cues which accompany a particular phoneme in a specific environment) depending upon the position of the consonant in the syllable and the position of the syllable in the utterance. Moreover they showed that stop consonants occurring in word initial position if substituted with another consonant were misperceived except for [g] which could be replaced by [b] whenever it occurs. When a stop consonant occurs in any other position, all voiced stops could be replaced by [d] while all unvoiced are replaced by [t] without loosing the meaning of the sentences. The results of this experiments suggested that some speech segments could be changed by production error or by phonological rules without changing the listener's perception. Moreover, they showed that the perception of [b, d] is much stable than the perception og [g], and the perception of [t, k] is much stable than the perception of [p]. Since neural networks can only extract features from their inputs in order to realize the recognition process, the phonemes that more often undergo changes likes that mentioned above are not classified so well as the others. These observations can explain the different performance obtained on the phonemes we try to recognize. Indeed, our experimental results, reported through the tables which follow, showed that net recognition percentage for [p, g] are always lower than for the other consonants.

4. DESCRIPTION OF THE NETWORK ARCHITECTURE AND LEARNING ALGORITHM

To perform the task mentioned above, we used Time-Delay Neural Networks which have turned out to be very suitable for phoneme recognition because , as Waibel showed the features learnt by such networks are invariant under translation in time [8, 10]. However our net architecture is more simplified compared to the one proposed by Waibel et al. In fact, our input layer is a vector of nine components whereas Waibel used a matrix with 16×15 components. Our net architecture that performs best is 9-256-6-6, i.e. the input layer has 9 units, the first hidden layer contains 256 nodes, the second hidden layer has 6 units and the output layer contains 6 neurons. The delays are short in order to capture the variable acoustic features of English stops. Each of the six output units corresponds to [b], [d], [g], [p], [t], [k]. Our net operates in the same way as the TDNN descibed by Waibel [8, 9, 10].
The network is trained with normalized input data using an on-line back-propagation algorithm (which is more appropriate for speech recognition [3]) without using the momentum for the updating phase. .Indeed, this factor does not seem really efficient when used with an on-line learning algorithm [3, 4]. The learning rate is low (0.03)_in order to avoid as much as possible the local minima problem. The backpropagation learning algorithm is a gradient descent of the mean squared error as a function of the weights i.e.

$$\Delta\omega_{ji} = -\eta \times \frac{\partial E_{pat}}{\partial \omega_{ji}}$$

where η is the learning rate, ω_{ji} the connection from node i in the layer $[s]$ to node j in the successive layer $[s+1]$, and

$$E_{pat} = \frac{1}{2}\sum_{j}(tar_{pat,j} - out_{pat,j})^2$$

defined the error function for pattern pat. Expressions $tar_{pat,j}$ and $out_{pat,j}$ define respectively the target and the output for the pattern pat on node j [7]. The choice of sigmoid function as activation function is motivated by its mathematical properties [6] and its use means that enough information about the output is available to units in earlier layers [7].

5. RECOGNITION EXPERIMENTS

The first experiment is done in order to explain our modification in the RASTA processing algorithm. We run the net with an architecture of 9-24-6-6 and summarized the performance rate depending upon the number of epochs and algorithms used (see **Table 2** and **Table 3**).

Phonemes	epochs 200	epochs 1000	epochs 1800	epochs 2600
[b]	67.1	73.2	69.1	65.0
[b]	59.0	56.3	63.1	70.0
[g]	32.0	31.3	33.0	35.0
[p]	34.1	45.1	47.4	53.1
[t]	80.0	84.2	84.2	81.5
[k]	62.2	67.1	67.0	66.3
Total	46.0	59.5	60.6	61.6

Table 2: Performance rate with the original RASTA-PLP algorithm during the training phase.

Phonemes	epochs 200	epochs 1000	epochs 1800	epochs 2600
[b]	74.3	73.2	69.4	73.4
[b]	76.0	81.3	79.1	76.0
[g]	54.0	45.2	49.1	58.2
[p]	51.0	63.0	65.0	64.2
[t]	86.3	89.4	88.2	88.5
[k]	79.4	77.4	84.1	82.0
Total	70.2	72.2	72.5	73.7

Table 3: Performance rate with the modified RASTA-PLP algorithm during the training phase.

It appears that the average recognition performances of the net with the modified algorithm were better than those obtained with the original algorithm. This result confirms that net performances depend upon how the data are preprocessed [3]. Therefore, all the subsequent training and testing experiments were done using data preprocessed with our modified RASTA-PLP algorithm.

6. OPTIMAL NUMBER OF NEURONS FOR IMPROVED PERFORMANCE

It is largely accepted that neural performances improve when increasing the number of neurons [3]. Our goal in doing this experiment is to propose an optimal net architecture for obtaining the best performance. Our TDNN net architecture only offers the possibility to increase the number of hidden neurons in the first layer. For practical and computational reasons, we think it is not possible to increase such a number indefinitely, there must exist a limit. As it is possible to see in **Table 4**, the performance increases according to the number of hidden neurons in the first layer with the existence of a maximum.

Number of nodes in the first layer	24	32	64	128	256	350
Performance (%)	72.1	72.8	81.5	85.0	88.1	85.6

Table 4: Average net performance rate during the training phase as a function of the number of hidden neurons. Such performances are obtained with data preprocessed with the modified RASTA-PLP algorithm .

From these results, the net architecture 9-256-6-6 is an optimal choice. Both the training and testing results on such net architecture are given in **Table 5**. The net has been trained over 1600 epochs. A greater number of epochs did not improve the net performances. As the number of epochs grows overfitting may occur and, as consequence, the net performances on testing set gets worst.

Phonemes	Performance %	
	Training	Testing
[b]	93.7	92.9
[d]	92.3	91.8
[g]	77.5	92.4
[p]	78.4	80.3
[t]	92.4	90.8
[k]	94.4	94.2
Total	88.1	90.4

Table 5: Net performance rate during the training and testing phases for a 9-256-6-6 TDNN architecture.

7. CONCLUDING REMARKS

We showed that a net architecture 9-256-6-6 is an optimal choice to improve the English stop consonants classification using Time Delay Neural Network. Moreover, we proved that the Hamming window of 10 msec moved at every 5 msec rate and a sampling rate of 16 kHz are more appropriate when the speech signal is preprocessed using the RASTA-PLP algorithm.

8. ACKNOWLEDGMENTS

The authors are grateful to Professor Maria Marinaro, President of IIASS and Professor Antonino Zichichi, President of ICSC-World Laboratory.

9. REFERENCES

1. H. Hermansky, " Perceptual Linear Predictive (PLP) Analysis of Speech", *J. Acoust. Am.* vol 87(4), 1990, p. 1738-1752.

2. H. Hermansky and N. Morgan, " RASTA Processing of Speech", *IEEE Transactions on Speech and Audio Processing*, vol. 2, N°4, 1994, p. 478-589.

3. Y. Bengio, "Artificial Neural Networks and their Application to Sequence Recognition", Ph. D. Thesis, Department of Computer Science, McGill University, Montreal, 1991.

4. L. Bottou et al. Scott, " Speaker Independent Isolated Digit Recognition : Multilayer Perceptrons vs Dynamic Time Warping", *Neural Networks*, Vol. 3, 1990, p. 453-456..

5. R. A. Cole and B. Scott, " Toward a Theory on Speech Perception", *Psychological Review*, Vol. 81, N°4, 1974, p. 348-374.

6. S. W. Ellacott., "Aspects of The Numerical Analysis of Neural Networks", *Acta Numerica* 1994, p. 145-202.

7. R. Beale and T. Jackson, *Neural Computing: An Introduction*, Institute of Physics Publishing Bristol and Philadelphia, IOP Publishing Ltd, 1992.

8. A. Waibel et al., "Phoneme Recognition Using Time-Delay Neural Networks", *IEEE Transactions on Acoustics, Speech and Signal Processing*, Vol. 37(2), 1989, p..328-339.

9. A. Waibel et al., "Consonant Recognition by Modular Construction of Large Phonemic Time-Delay Neural Networks", *IEEE International Conference on Acoustics, Speech and Signal Processing*, May 1989.

10. A. Waibel, "Connectionist Glue: Modular design of Neural Speech Systems, *Proceeding on Connectionist Models*. Summer School, Morgan Kaufmann, 1988.

11. L. Lisker, A. S. Abramson, "A Cross Language Study of Voicing in Initial Stop: Acoustic Measurements", *Word*, Vol. 20, 1964, p.384-422.

12. L. Lisker, A. S. Abramson, "Some Effects of Context on Voice Onset Time in English Stops", *Language and Speech*, Vol. 10(3), 1967, p. 1-28.