VOICED/UNVOICED/SILENCE CLASSIFICATION OF SPEECH USING 2-STAGE NEURAL NETWORKS WITH DELAYED DECISION INPUT

Raphael Ahn* and W. Harvey Holmes*

*Speech & Signal Processing Laboratory
School of Electrical Engineering
The University of NSW

ABSTRACT - This paper proposes a two stage feed-forward neural network classifier capable of determining voiced, unvoiced and silence in the first stage and refining unvoiced and silence decisions in the second stage. Delayed decision from the previous frame's classification along with preliminary decision by the first stage network, normalised partial-sum-of-autocorrelation-coefficient ratio and energy ratio enables the second stage to correct the mistakes made by the first stage in classifying unvoiced and silence frames. Comparisons with a single stage classifier demonstrates the necessity of two stage classification techniques. It also shows that the proposed classifier performs favourably even in the presence of noise.

INTRODUCTION

Development of back-propagation training method by Rumelhart, Hinton & Williams (1986), has led to wide use of multi-layer perceptrons (MLP) in many pattern recognition problems. Voiced/unvoiced (V/UV) classification can be viewed as a pattern recognition and as such many researchers Bendiksen & Steiglitz (1990), Cohn (1991), and Ghiselli-Crippa & El-Jaroudi (1991) have applied neural networks in this area of speech processing. However, the problem of voiced/unvoiced/silence (V/UV/S) classification still remains a difficult one to solve.

Due to diverse statistical properties between V and UV or between V and S, V/UV or V/S classifications can be achieved with relative ease. Many such properties, however, are too similar between UV and S to be of use (Cao, Sridharan & Moody,1995). In order to overcome this problem, this paper will present two tandem feed-forward neural networks (NNs) capable of classifying V/(UV/S) in the first stage and UV/S in the second stage. The strength of this classifier rests with the features used in the second stage. They are delayed decision (DD) of previous frame, preliminary decision (PD) by the first stage NN, normalised partial sum of autocorrelation coefficient ratio (NPR) and RMS energy ratio (ER) between the current and the previous frames. Three of the features represent trends in a speech signal. As it shall be shown, these trends in speech prove to be major distinguishing factors between UV and S.

TWO STAGE NEURAL NETWORK CLASSIFIER

Figure 1 below illustrates the structure of the proposed classifier. There are two stages to the classifier. The first stage NN contains 5 input nodes, 3 output nodes (one for each classification of V, UV and S) and 15 hidden-layer nodes. Although the network performance is more than adequate with only 8 or 10 hidden-layer nodes, for reasons of convergence 15 nodes are chosen. The second stage NN contains 4 input nodes, 2 output nodes (for UV and S) and 8 hidden-layer nodes.

The first stage NN is equivalent to a single stage NN capable of classifying V, UV and S. In other words, results from the first stage alone can be used as the final classification if UV and S classifications need not be very accurate. The second stage NN requires almost negligible increase in computation as its input features are already available from the first stage or can be derived with minor operations.

As seen in Figure 1, if a speech segment is determined to be V by the first stage NN, the final output of the classifier will be V. In cases where a speech segment is either unvoiced or silent as determined by the first stage, the second stage NN is invoked to fine-tune the classification using the 4 features mentioned above. As the results will show, this two stage design makes a perfect sense since the first stage can accurately distinguish V from other classifications while the root of misclassification by the first stage resides in UV and S classification.

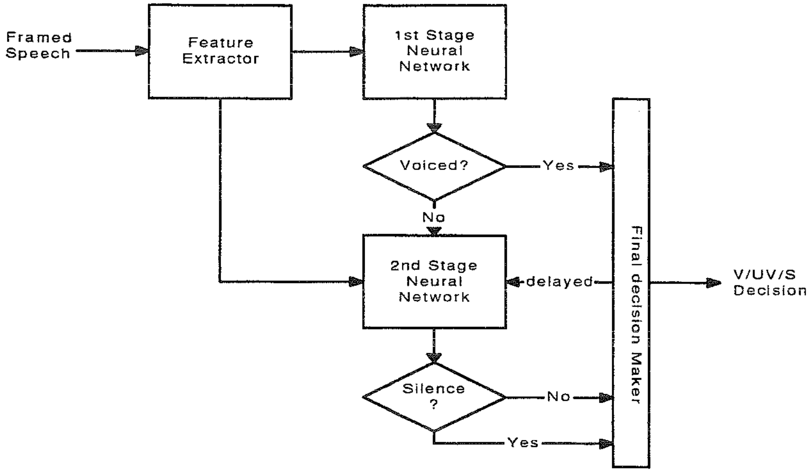

Figure 1. Structure of two-stage neural network classifier

FEATURE SELECTION

Though many possible features exist, five features are selected for their ability to differentiate well between V and UV or V and S classifications. They are as follows:

RMS Energy of Signal (RMSE) :

$$f_1 = \left(\frac{1}{N} \sum_{i=1}^{N} s_i^{\,2} \right)^{1/2} ;$$

Zero-crossing Count (ZC) :

$$f_2 = \sum_{i=0}^{N} Z_{ci} \text{ , where } Z_{ci} = \begin{cases} 1 & \text{if } sign(s_i) - sign(s_{i-1}) \neq 0 \\ 0 & \text{otherwise} \end{cases} ;$$

Normalised Partial Sum of Autocorrelation Coefficients (NPSAC) :

$$f_3 = \sum_{p=1}^{P} R(p)/R(0) ,$$

where $R(\tau) = \sum_{n=1}^{N} s(n)s(n+\tau)$ for $\tau = 0, \pm 1, \ldots, \pm(N-1)$ and $P = 3$ for 8 kHz sampling;

LPC Normalised Prediction Error (LPCE) :

$$f_4 = E_p = 10\log_{10}\left(10^{-6} + \left| \sum_{k=1}^{P} a_k \phi(0,k) + \phi(0,0) \right| \right) - E_s ,$$

where $E_s = 10\log_{10}\left(10^{-6} + \frac{1}{N} \sum_{n=1}^{N} s^2(n) \right)$, and $\phi(i,k) = \frac{1}{N} \sum_{n=1}^{N} s(n-i)s(n-k)$;

First LPC Predictor Coefficient (LPC1) :

$$f_5 = a_1 .$$

Four of the features are traditionally favoured parameters used in many classifiers including an early method by Rabiner et al. Atal & Rabiner (1976). The third parameter proposed in Cao, Sridharan & Moody (1995) provides robust classification of V and UV stop consonants.

As mentioned earlier in Introduction, DD, PD, NPR (Current Frame's NPSAC / Previous Frame's NPSAC) and ER (Current Frame's RMSE / Previous Frame's RMSE) are used as inputs to the second stage. NPR and ER are selected for their sensitivity to the changes or the lack of changes in the speech signal. When combined with DD and PD in training, these features allow the network to learn from the mistakes made by the first stage NN and correct them.

CLASSIFIER TRAINING

Speech segments from TIMIT database were used to train and test the classifier. Four male-spoken and Four female-spoken sentences were down-sampled to 8 kHz and segmented into 60-sample frames. Following table shows the number of frames used to train the classifier according to their classification by TIMIT.

Classification	Voiced	Unvoiced	Silence	Total
No. of Frames	1520	1310	670	3500

Table 1. Training vector size according to classification

Each frame was passed through feature detector producing 5 element training vector for the first stage NN. All frames with UV or S decision from the first stage NN are then collected to form input vectors for the second stage NN along with newly obtained ZCR, ER and DD.

Both stages were trained using back-error propagation technique with adaptive learning rate and momentum. Fast convergence was observed for both training cases. Although an iteration value of 3000 were used in training each of the two stages, rapid convergence toward final errors were demonstrated within 100 iterations. This can be seen in Figure 2 and 3.

CLASSIFIER TESTING

Three male-spoken and Three female-spoken sentences were used in testing the proposed classifier against a single stage NN with 20 hidden-layer nodes and 7-feature input vectors (composed of 5 features from the first stage NN plus NPR and ER). Table 2 below lists input speech vector sizes by classification.

Classification	Voiced	Unvoiced	Silence	Total
No. of Frames	1096	726	678	2500

Table 2. Test vector size according to classification

Table 3 shows the percentile error rate in each error categories according to the total input vector size. This table is organised such that the relationship between the total error rate and each individual error rate can be easily seen. However, it does not represent the 'true' error rate in each error category as the number of input vectors in each classification are not equal. For this reason, a second table is presented. Table 4 shows the error rate according to vector sizes in each classification.

Error Type	V → UV	V → S	UV → V	UV → S	S → V	S → UV	Total
Single stage	0.92%	0.12%	0.84%	3.56%	0.08%	2.76%	8.28%
First stage	0.96%	0.08%	0.88%	3.80%	0.04%	2.88%	8.64%
Proposed	0.96%	0.08%	0.88%	0.16%	0.04%	0.12%	2.24%

Table 3. Percentile error rate according to total input vector size

Error Type	V → UV	V → S	UV → V	UV → S	S → V	S → UV
Single stage	2.10%	0.27%	2.89%	12.26%	0.29%	10.18%
First stage	2.20%	0.18%	3.03%	13.09%	0.15%	10.62%
Proposed	2.20%	0.18%	3.03%	0.55%	0.15%	0.44%

Table 4. Percentile error rate according to vector sizes in each classification

Both tables demonstrate inadequacy of a single stage classifier to distinguish UV and S. In comparison, our proposed method handled UV/S classification with much higher accuracy.

Further tests were conducted to examine robustness of the proposed classifier in noisy speech. The same set of six sentences used in the above testing were tested with varying degree of degradation by white noise. The results are tabulated in Table 5.

SNR	UV/S Error	V/S Error	V/UV Error	Total Error
Original	0.28%	0.12%	1.84%	2.24%
30dB	0.28%	0.12%	1.84%	2.24%
20dB	0.84%	0.16%	2.26%	3.26%
10dB	4.26%	0.48%	6.46%	11.20%
0dB	9.52%	1.96%	18.58%	30.06%

Table 5. Percentile error rate with noisy inputs

Error rates in each category are computed with the total input vector size for convenience and for ease in comparison to the total error. Figure 4 illustrates the noise test results. Table 5 shows that the proposed classifier is capable of handling degraded speech up to 20 dB SNR without significant fault. Increase in V/UV error becomes noticeable between 10 and 20 dB. As V/S and V/UV is determined by the first stage NN, it seems that the selected features and NN weights are capable of handling down to 10 dB error rate before succumbing to high noise level. In contrast, there seems to be a large increase in UV/S error between 20 and 10 dB SNR. As UV/S is classified by the second stage, it seems to indicate that further training with a larger number of vectors or more features may be necessary to increase the robustness in the second stage.

CONCLUDING REMARKS

As the test results show, the proposed classifier performs particularly well in classifying UV/S. One can also observe from the results that V/UV and V/S errors for both single stage classifiers are much smaller when compared to UV/S errors. This demonstrates that most inputs features to the first stage are capable of distinguishing V/UV and V/S cases with relative ease. However, when the same features are used to classify UV/S, the results are less than satisfactory. The results in Table 3 show a marked improvement in UV/S classification when 2-stage classifier is used. These findings clearly support the argument for the necessity of 2-stage classifier with the second stage solely dedicated to classifying UV/S.

In order to improve robustness against error, further investigation is currently underway. In particular, we are examining noisy training vectors for both stages of classifier. Preliminary studies show promising results.

REFERENCES

Rumelhart, H. W., Hinton, G. E. & Williams, R. J. (1986) "Learning Internal Representations by Error Propagation," *Parallel Distributed Processing*, Vol. 1: Foundations, 318-362, (MIT Press, Cambridge, MA).

Bendiksen, A. & Steiglitz, K. (1990) "Neural Networks for Voiced/Unvoiced Speech Classification," *ICASSP-90*, 521-524.

Cohn, R. P. (1991) "Robust Voiced/Unvoiced Speech Classification using a Neural Net," *ICASSP-91*, 437-440.

Ghiselli-Crippa, T. & El-Jaroudi, A. (1991) "A Fast Neural Net Training Algorithm and Its Application to Voiced-Unvoiced-Silence Classification of Speech," *ICASSP-91*, 441-444.

Cao, Y., Sridharan, S. & Moody, M. (1995) "Voiced/Unvoiced/Silence Classification of Noisy Speech in Real Time Audio Signal Processing," *AES - 5th Australian Regional Convention*, Preprint 4045, Apr.

Atal, B. S. & Rabiner, L. R. (1976) "A Pattern Recognition Approach to Voiced-Unvoiced-Silence Classification with Applications to Speech Recognition," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. ASSP-24, 201-212.

MORE FIGURES

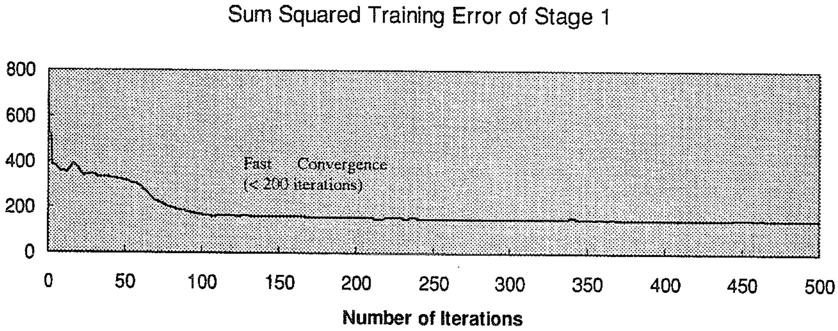

Figure 2. Plot of 1st stage training error vs. number of iterations

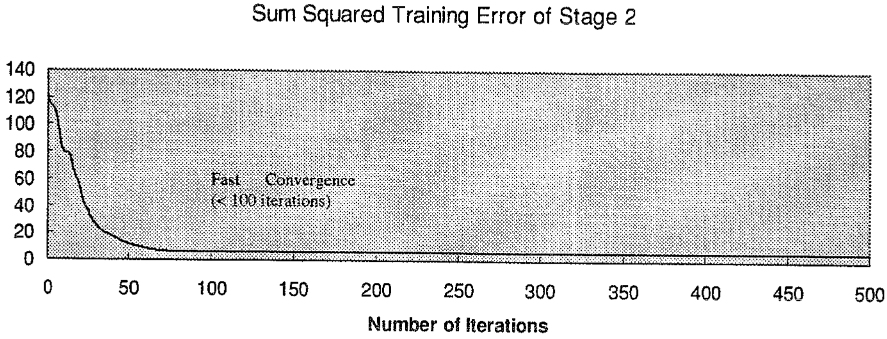

Figure 3. Plot of 2nd stage training error vs. number of iterations

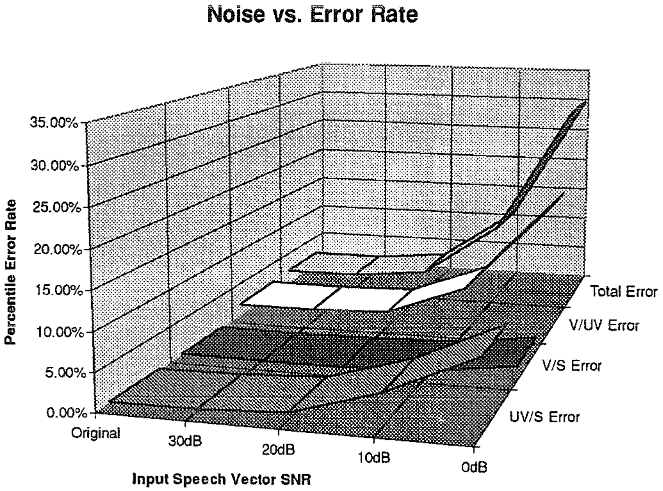

Figure 4. Noise test results