Application of Multilayer Feedforward Network to the Voiced- Unvoiced- Silence Detection Problem

Felix ALBU and Adelaida MATEESCU

Abstract ----Multilayer feedforward networks (MFN) are tested in detection and classification tasks. The classification of the speech signal into voiced, unvoiced, and silence (V/U/S) provides a preliminary acoustic segmentation of speech, which is important for speech analysis and may also have applications in speech-data mixed communication systems. The classification was made using a MFN with different hybrid features. The paper evaluates the efficiency of these feature vectors, the classification rate as a function of signal to noise ratio and simulation results are presented.

1. Introduction

The classification of the speech signal into voiced, unvoiced, and silence (V/U/S) provides a preliminary acoustic segmentation of speech, which is important for speech analysis. Systems that do this classification using the speech signal range from simply ones that threshold the short-time energy and measure the zero-crossing rate to systems using sophisticated pattern recognition techniques [3,6].

The nature of the classification is to determine whether a speech signal is present and, if so, whether the production of speech involves the vibration of the vocal folds. The voiced speech is produced by modulating puffs of air (creating by the vibrating vocal cords) by the vocal tract shape corresponding to the sounds being spoken[5]. The unvoiced speech has a noiselike character with no periodicity (the vocal cords are not vibrating) and no slowly changing temporal characteristics. When both quasi-periodic and noisy excitations are present simultaneously (mixed excitations) the speech is classified here as voiced because the vibration of vocal folds is part of the speech act[2]. .

In this study, a procedure is developed for making the V/U/S classification using a multilayer feedforward network (MFN)[1]. The feature vector for the classification is a combination of parametric features (cepstral coefficients, pondered cepstral coefficients or inverse sinus parameters) and waveform features (zero crossing rate, a function of rms energy). Additional waveform features are included to enhance the separation in pattern space when spectral information alone is not sufficient for making the classification.

2. Network training and classification

A block diagram for the network training and classification process is illustrated in Fig. 1. The speech was filtered at 3.5 kHz and digitized using a 12 - bit A/D at a sampling rate of 8 kHz.

The digitized signals were further high-pass filtered at 300 Hz by a fourth-order Butterworth digital filter to eliminate low-frequency hum or noise. The digitized speech is preemphasized using a simple first-order digital filter with a preemphasis factor (0.95). Each frame of speech is weighted by a Hamming window [1].

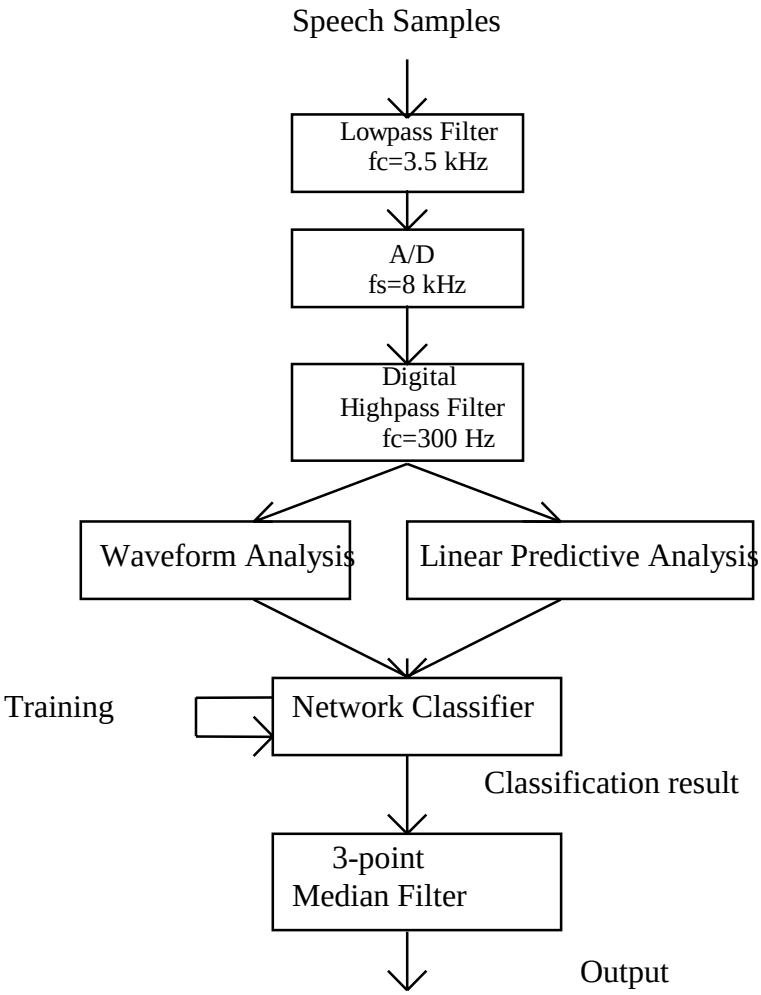

Fig.1. Flow chart of training and classification processes

The first feature vector was a combination of 12 cepstral coefficients derived from 12 LP coefficients, the zero-crossing rate and a function of rms energy (rms was evaluated taking into account N=200 samples having normalized values with respect to maximum value). An example set of the first feature vector are shown in Fig. 2.

The second feature vector was a combination of 12 pondered cepstral coefficients, the zero-crossing rate and a function of rms energy. Each coefficient of the of the cepstral vector $c_l(m)$, at time frame l is weighted by a window $W(m)$ of the form

$$W(m) = \left[1 + 6 \cdot \sin\frac{\pi \cdot m}{12}\right] \qquad 1 \le m \le 12 \qquad (1)$$

to give $\hat{c}_l(m) = c_l(m) \cdot W(m)$ (2)

The raised sine lifter, $W(m)$, was used successfully for digit recognition [4]. An example set of the second feature vector are shown in Fig. 3.

The third feature vector was a combination of 12 inverse sine (IS) parameters, the zero-crossing rate and a function of rms energy. The inverse sine parameters were derived by parcor coefficients (they are often called reflection coefficients) $k_l(m)$. We can obtain the reflection coefficients from the LP parameters [3]. The inverse sine parameters are given by

$$\sigma_l(m) = \frac{2}{\pi} \cdot \sin^{-1} k_l(m) \qquad 1 \le m \le 12 \qquad (3)$$

The IS parameters have the advantage of remaining bounded in magnitude by unity

$$|\sigma_l(m)| \le 1 \qquad (4)$$

for any *l*.

An example set of the third feature vector are shown in Fig. 4.

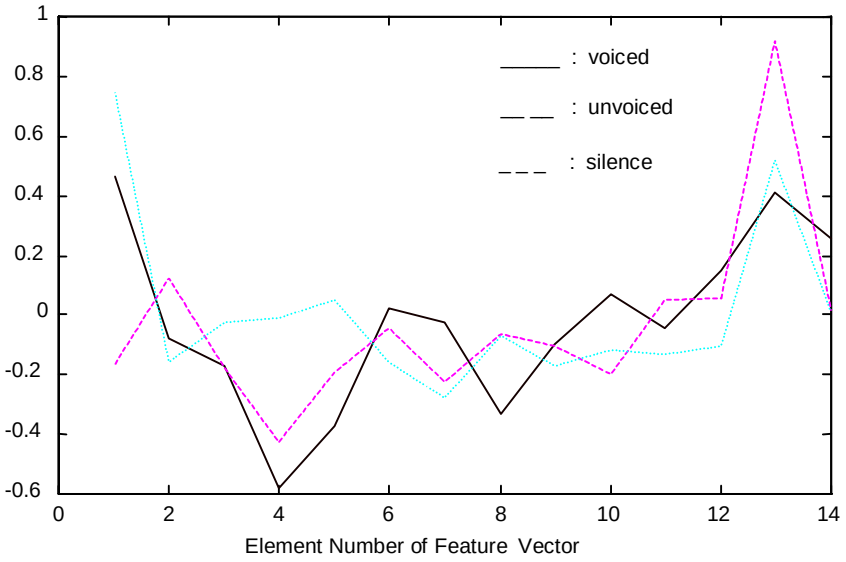

Fig. 2. Example first average feature vectors (element 1-12: cepstral coefficients, element 13: zero-crossing rate, element 14: normalized rms energy)

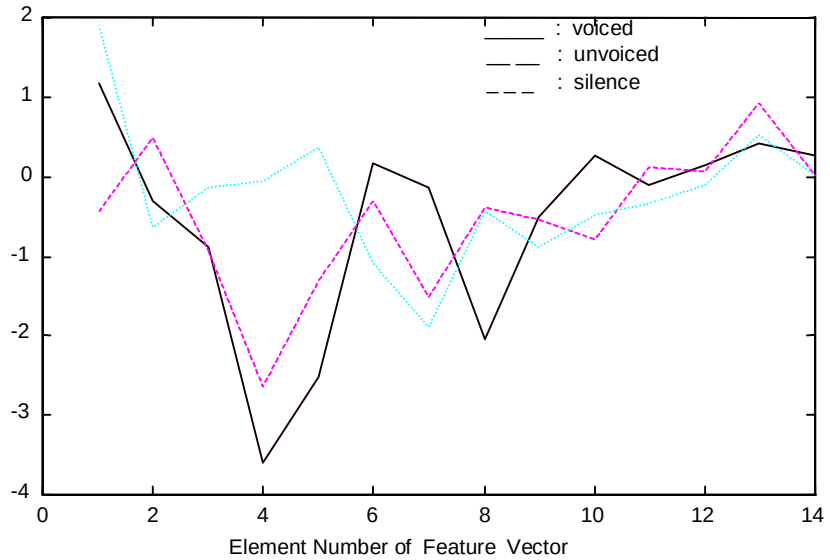

Fig. 3. Example second average feature vectors (element 1-12: pondered cepstral coefficients, element 13: zero-crossing rate, element 14: normalized rms energy).

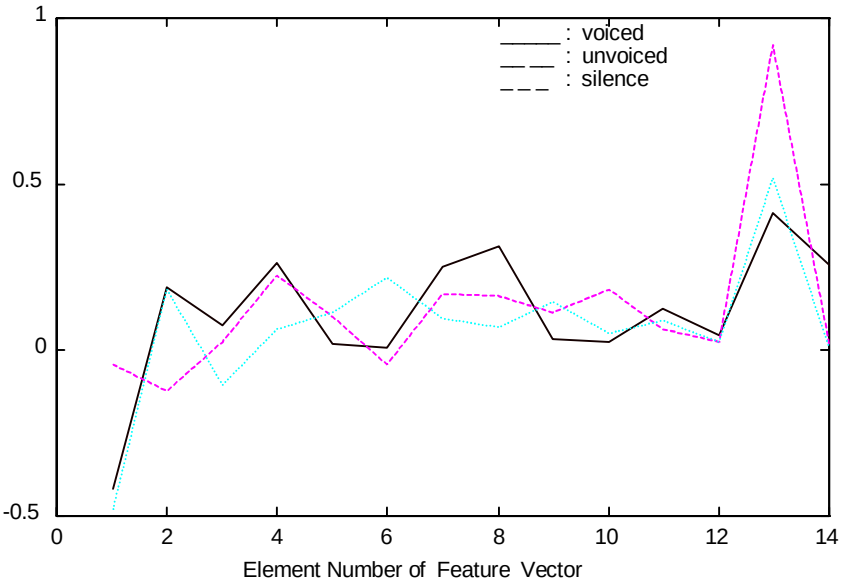

Fig. 4. Example first average feature vectors (element 1-12: inverse sine parameters, element 13: zero-crossing rate, element 14: normalized rms energy)

The V/U/S classification was made for each input feature vector after training was completed. The classification output was further decoded and passed through a three-point median filter to eliminate isolated "impulse" noise [1].

The network was trained using an adaptive learning rate with learning rate (0.01), learning rate increase (1.05), learning rate decrease (0.7), and momentum (0.9). The training loop would not terminate until a total of 20000 training iteration had been exhausted.

The input and output layer had fixed numbers of PE's (14 and respectively 3). We used a single hidden layer with 20 nodes. The output vector was coded as [100] for voiced sound, [010] for unvoiced sound, [001] for silence. This coding was selected to maximize the code differences between categories. Because of the minimum and maximum of the activation function could only be reached at infinity, 0 and 1 were replaced by 0.1 and 0.9, respectively, in practical calculations [1].

3.Database

The speech database used in our experiment included Romanian digit numbers and other words spoken by three talkers (two males and one female). The speech recordings were pre-processing (see Fig. 1.) and were interactively labeled in three sound categories using waveform and spectrographic displays and audio output as feedback. The network classification rate was obtained using a three-step procedure: (1) a set of training samples of a given size was randomly selected from the database, (2) all data samples (excluding the training samples) were classified once training was completed, and (3) an error was counted whenever the network classification differed from manual classification [1].

4. Network performance

The V/U/S classification is susceptible to noise corruption because the unvoiced speech itself is a noise and the corruptive noise will significantly obscure the distinction between silence and unvoiced speech [1]. The noise added was a Gaussian random noise whose variance was manipulated to control the signal-to-noise ratio. 90 training samples (30 from each sound category) were randomly selected after the noise of the appropriate level (depended on the signal level of each speaker) was added. Additive noise contaminates the speech signal and changes the feature vectors representing the speech. These phenomena produce mismatches between the training and classification conditions that result in degradation in accuracy.

As can be seen all classifiers were degraded in a comparable rate when the signal to ratio was reduced. All classifiers had practically failed when the signal to noise ratio was reduced to 0 dB. The worst comportment for low signal to noise ratio was obtained if we used the sinus inverse parameters for feature vector. The results are presented in Fig 5.

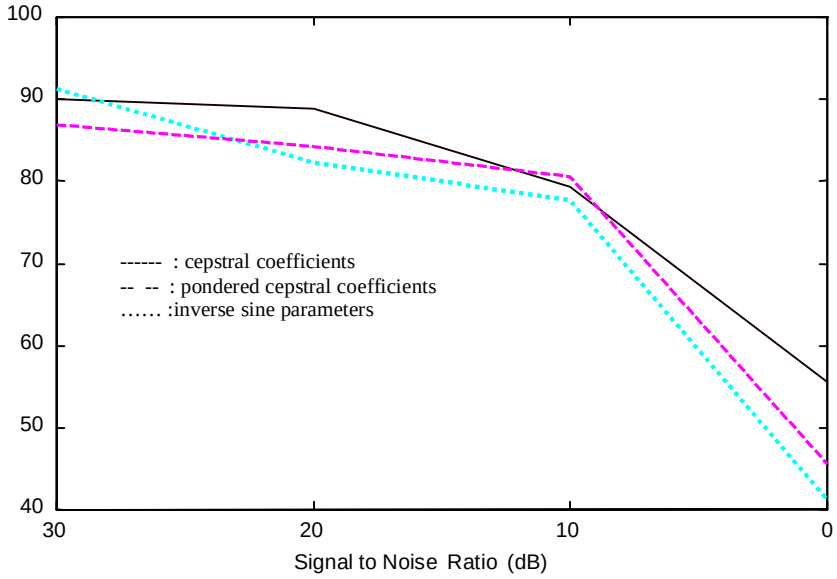

Fig. 5. Classification rate as a function of signal to noise ratio.

With a first feature vector we were achieved an overall classification rate of 90.6 %, with a second feature vector we were achieved an overall classification rate of 87 % and with a third feature vector we were achieved an overall classification rate of 91.4 %. Sample classification results when are used the first feature vector are shown in Fig. 6.

High accuracy might be obtained with other features (such introducing a nonlinear function of rms energy) or by increasing the number of hidden layer.

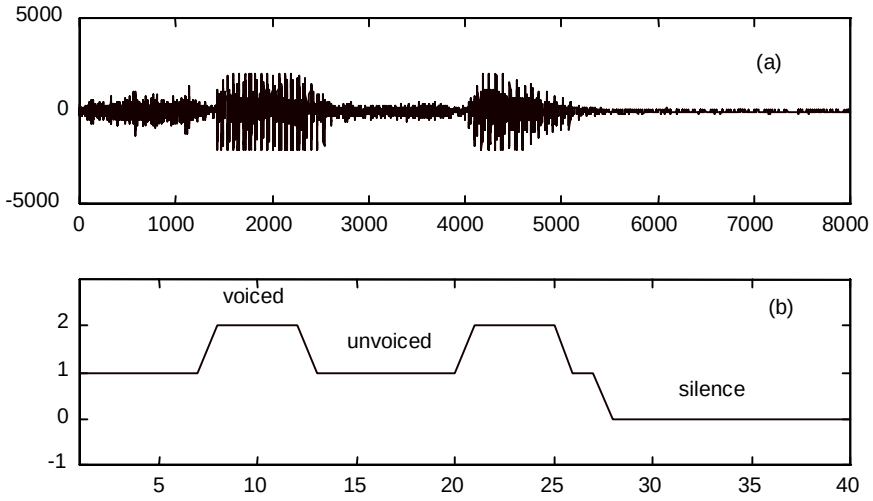

Fig. 6. (a) Speech waveform. (b) Network classification for the Romanian word "SASE " spoken by one male speaker

Bibliography

[1]. Y. Qi , B. Hunt , " Voiced- Unvoiced- Silence Classifications of Speech Using Hybrid Features and a Network Classifier ", *IEEE Transactions on Speech and Audio Processing*, vol. 1, No. 2 , April 1993

[2]. L. Siegel and A.Bessey, " Voiced/Unvoiced/Mixed excitation Classification of Speech ", *IEEE Trans. Acoustic, Speech, Signal Processing*, vol. ASSP-30, pp. 451-460, June 1982

[3]. J. R. Deller, Jr., J. G. Proakis, J. H. L. Hansen, " *Discrete-Time Processing of Speech Signals* " Macmillan, New York, 1993

[4]. B. H. Juang, L. R. Rabiner, and J. G. Wilpon , " On the Use of the Bandpass Liftering in Speech Recognition ," *IEEE Trans. Acoustic, Speech, Signal Processing*, vol. 35, pp. 947 - 954, July 1987

[5]. D. P. Morgan, and C. L. Scotfield, "*Neural Networks and Speech Processing* " Norwell, Mass. :Kluwer, 1991

[6]. B. Atal and L. Rabiner , " Apattern recognition approach to Voiced- Unvoiced- Silence Classification with Applications to Speech Recognition, "*IEEE Trans.* Acoustic, *Speech, Signal Processing*, vol. ASSP-24, pp. 201-212, June 1976