A Neural Networks Approach to Voiced-Unvoice-Silence Classification Incorporating Amplitude Distribution

(음성 진폭분포로 신경망을 구동한 유-무-묵음 분류)

In-Seop Lee,* Jung-Ah Choi,* Myung-Jin Bae,* Sou-Guil Ann*

이 인 섭,* 최 정 아,* 배 명 진,* 안 수 길*

ABSTRACT

Voiced-unvoiced-silence classification is one of the most important problems on speech analysis. Various classification methods have been proposed. Many of them incorporate several parameters: energy, zero-crossing rate of the signal, autocorrelation coefficients, LPC coefficients, energy of the prediction error, etc. Basically, they need preprocessing to extract the above mentioned parameters. In this process, considerable computations are required and much of the information inherent in the speech signal may be lost. Thus, a new algorithm using the amplitude distribution for each frame is proposed. In the first stage, we obtain the frame amplitude distribution. Each of V-U-S regions has its specific amplitude distribution patterns. And in the second stage, we classify the first stage using neural networks. Neural networks do not require the threshold selection techniques and are robust to background noise and are suitable for real time processing.

요 약

유-무-묵음 분류과정은 음성분석시에 아주 중요한 문제중의 하나이다. 음성에너지, ZCR, 자기 상관계수, LPC 계수, 예측에러 에너지등을 퍼래미터로 사용하여 지금까지 많은 분류기법이 제안되어져 왔다.

이런 기법들은 기본적으로 퍼래미터를 추출해야 하고, 이때문에 많은 계산량이 요구되고, 이들 퍼래미터는 음성본래의 정보들의 대부분을 상실하게 된다. 이때문에 각 프레임의 진폭분포를 사용하는 새로운 앨고리즘을 제안하였다. 첫째로 V-U-S 영역은 개별 진폭분포형태를 가지기 때문에 주어진 프레임에서 진폭분포를 구한다. 그런 다음에는 신경망을 통해 분류를 하게 된다. 신경망은 문턱값을 별도로 선정할 필요없고, 배경잡음에 강력하며, 또한 실시간 처리에 적합하다.

I. INTRODUCTION

The need for deciding whether a given segment of a speech waveform should be classified as voiced speech, unvoiced speech or silence (absence of speech) arises in many speech analysis systems. And it is one of the most difficult problems in speech analysis. There are several reasons why this is so. One problem is the large dynamic range of the speech signal itself in which a 20-40 dB variation of signal level is not uncommon within the speech of a single talker. Compounded with this is a 20-40 dB variation in level among talkers. Another problem is that sometimes the acoustic waveform does not provide accurate information about the signal classification, e.g., the vocal cords

* Seoul National University
Dept. of Electronics Engineering

are vibrating (i.e. the signal is voiced speech) but no periodicity is seen in the acoustic waveform. Finally, all these problems are compunded by the degradations which include band-limiting, nonlinear phase distortion, center clipping, and noise addition.

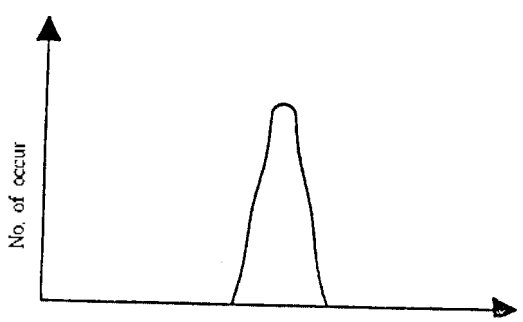
A variety of approaches has been described in the speech literature for making this decision. There is an approach using a logical decision based on the values of a certain measured feature of the signal, e.g., zero-crossings, energy, autocorrelation coefficients, prediction error, etc. When used in conjunction with pitch detection, features of the pitch detector are often used to supplement the voiced-unvoiced decision. All of these methods can be classified into two categories. One is a group which can be applied to a pitch detection problem directly. Methods which utilize the autocorrelation function of a speech signal (or its prediction error signal), and those which utilize cepstral peak values, are representatives of this category. The other includes methods whose approaches are similar to those of pattern recognition[1]. That is, they are based on differences of statistical distribution and characteristics of acoustic feature parameters between voiced and unvoiced speech. All of these methods have to adopt athreshold for decision, and the threshold depends on the signal-to-noise ratio(SNR), noise characteristics and speaker, etc. Thus the adaptation and optimization of threshold for V/U/S decision are complex and unreliable in a certain condition (e.g., nonstationary noise environments).

A new algorithm is suggested in this paper, which measures the frame amplitude distributions and uses artificial neural networks to decide V/U/S frame. The proposed algorithm needs no threshold selection technique and is a robust V/U classifier.

II. Amplitude Distribution Characteristics of Speech Signals in Each Frame

When applying statistical notions to speech signals, it is necessary to estimate the probability density from speech waveforms. The probability density is estimated by determining a histogram of amplitudes for a large number of samples, i.e., over a long time. Many extensive measurements of this kind have shown that a good approximation to measured speech amplitude densities is a gamma distribution, or somewhat simpler approximation is the Laplacian density. The gamma density is clearly a better approximation than the Laplacian density, but both are reasonably close. But, if we meausre the amplitude histogram of speech signals by the frame(in this paper 256 samples with 128 sample sliding window), it is apparent that voiced, unvoiced, silence speech signals have different types of distribution. The amplitude histogram of voiced speech signal tends to be distributed over a wide range of levles. In the silence region, the histogram tends to be concentrated around some low lelvels. The unvoiced speech signals are more likely to be concentrated around some low levels than the voiced region and a few of them have higher levels than silence. It can be seen in Fig.1[2].

We use these characteristics of the amplitude distribution in a frame to classify V/U/S speech. These characteristics are robust to nonstationary



(a) silence region

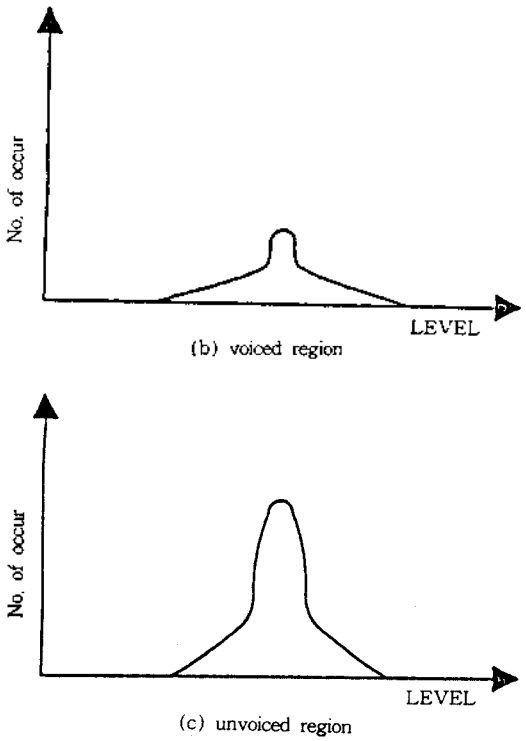


Fig. 1. Amplitude characteristics of V/U/S speech.

noise environments, e.g. burst noise, impulse noise as well as additive white gaussian noise

II. Artificial Neural Net

1. ANN Architecture

The feedforward ANN used for V/U/S classification is illustrated schematically in Fig.2[5].

The outputs of nodes in one layer are transmitted to nodes in the upper layer via links that amplify or attenuate such outputs through weighting factors. Except for the input layer nodes, the net input to each node is the weighted sum of the outputs of the nodes in the lower layer. Each node is activated according to the input to the node, the activation function, and the threshold of the node.

The net input to a node in layer j is

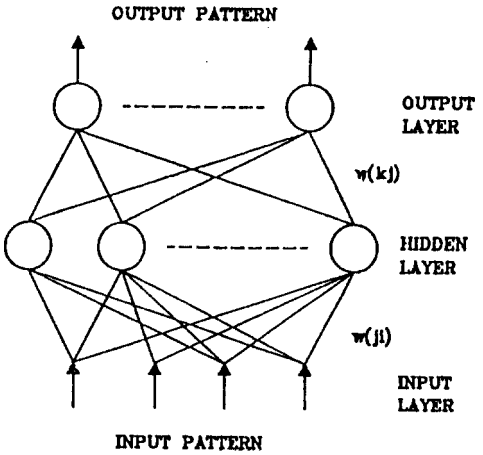


Fig. 2. A schematic depiction of a semilinear feedforward ANN.

$$net_j = \sum w_{ji} o_i \tag{1}$$

The output of node j is

$$o_j = f(net_j) \tag{2}$$

where f is the activation function. For the sigmoidal activation function used in our ANN, we have.

$$o_j = \frac{1}{1+\exp(-(net_j-\theta_j)/\theta_0)} \tag{3}$$

where the parameter θ_j serves as a threshold. The effect of a positive θ_j is to shift the activation function to the right along the horizontal axis, and the effect of θ_0 is to modify the shape of the sigmoid. The low value of θ_0 makes the sigmoid take on a threshold-logic unit. The sigmoidal activation function is illustrated in Fig.3.

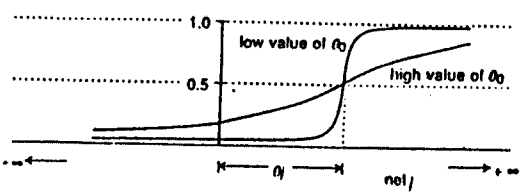


Fig. 3. The sigmoidal activation function with threshold and shape modification.

We adopt the generalized delta rule for learning the weights and the biases. In this procedure, we present the input pattern and ask the net to adjust the weights and the thresholds such that the desired outputs are otained at the output nodes. For all the input-output pairs presented, we ask the net to find a single set of weights and biases that will satisfy all the input-output pairs. However, the outputs will not be the same as th desired values. Let o_{pk} be its desired value. For each pattern, the square of the error is

$$E_p=\frac{1}{2}\sum_k (d_{pk}-o_{pk})^2 \quad (4)$$

and the average system error is

$$E=\frac{1}{2}\,1/P\sum^{P} E_p=\frac{1}{2}\,1/P\sum_p\sum_k (d_{pk}-o_{pk})^2 \quad (5)$$

The factor $\frac{1}{2}$ is used for mathematical convenience.

2. Learning in ANN

In the learning phase, the weights are adjusted in a manner to reduce the error E as rapidly as possible. We adjust weights by

$$w_{kj}(n+1)=w_{kj}(n)+\eta\delta_k O_j \quad (6)$$

where η is a gain term, δ_k is an error term of node k, and the quantity (n+1) is used to indicate the (n+1), th step. The δ_k's are given by the following two expressions:

$$\delta_k=(d_k-o_k)\,o_k(1-o_k) \quad (7)$$

if k is an output node or

$$\delta_j=o_j(1-o_j)\sum_k \delta_k w_{jk} \quad (7)$$

where, k is over all nodes in the layers above node j if j is a hidden node.

In the expression(6), a large value of η corresponds to rapid learning but might result in oscillations. This can be avoided by incluiding a sort of momentum term. That is,

$$w_{kj}(n+1)=w_{kj}(n)+\eta\delta_k O_j+\alpha(w_{kj}(n)-w_{kj}(n-1)) \quad (8)$$

where α is a proportionality constant between 0 and 1. The momentum terms specifies that the change in w_{kj} at the (n+1), th step should be similar to the change of the n-th step. A finite α tends to dampen the oscillation but it may slow the rate of learning.

Ⅳ. Computer Simulation

1. Experimental Conditions

Speech signals used in this simulation were isolated words of about 3 year old female speaker. The V/U/S decision for each analysis frame was made by visual inspection of its speech waveform. The speech samples were recorded in a sound-proof room. A noisy speech was created on a computer by adding the sampled white Gaussian noise sequence to the sampled speech sequence. The SNR was defined as the ratio of the average power of speech to the average power of additive noise of each data lengh. Experimented SNR's in this paper were -3, 0, 3, 10, 20 dB. The length of frame to calculate the histogram of amplitude is 256 samples and its overlap samples are 128 sample. Each of the speech signals was sampled at 8 kHz with 12-bit accuracy and band-limited to 3.4 KHz.

2. ANN architecture for V/U/S Classification

For V/U/S classification of speech, a 3 layer net is constructed. It is composed of an input layer.

two hidden layers, and an output layer. Its overall architecture is shown in Fig.4.

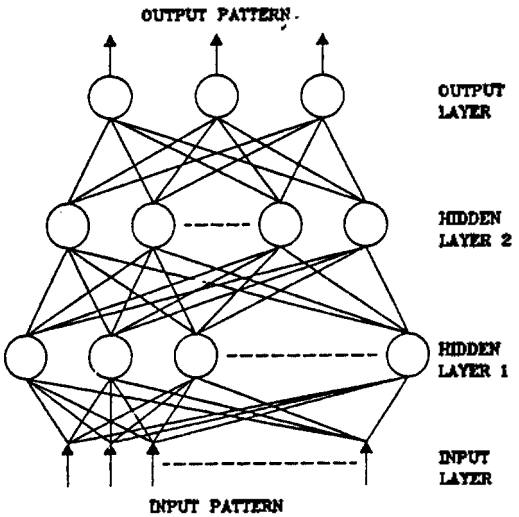


Fig. 4. The architecture of the ANN used for V/U/S classification.

In the first step, input speech is analyzed to yield the histogram for each frame. The histogram consists of 256 levels, which is used as input to the ANN. This input layer is fully interconnected to the unit of the first hidden layer.

In the second hidden layer, 8 hidden units are interconnected to the units of the first hidden layer. Finally, the output is obtained from the weighted sum of the outputs of the second hidden layer. The output layer consists of 3 units, one for the voiced speech, another for the unvoiced speech, and the other for the silence.

3. learning in our ANN

In our ANN, we have adopted the generalized delta rule formulated by Rumehart, Hinton, and Williams[6]. A gain term η is set to 0.7 and a momentum term is introduced. The database used for the experiment is split into two parts: the training and the testing set. The training set consists of 300, each classified as V/U/S regions. The training data are randomly scrambled to reduce the effect of input order. The system error is limited to 0.0000001 and the maximum number of iterations is 2000. The iteration stops if the system error reaches to the limited value or the maximum number of iteration is reached. The normalized error plot is shown in Fig.5.

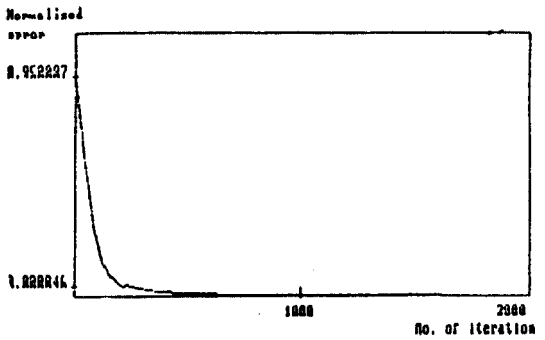


Fig. 5. Plot of normalized error vs. iterations.

4. Experiments

We have used 60 segment of Korean words as the testing data for output generation. Each word is analyzed on a frame basis to get the histogram and this histogram is input to the ANN. Output is generated immediately; the largest of the three output values indicates that the frame is voiced, unvoiced or silence classification result. The result are shown in table 1 and table 2.

Table 1. Classification results(open test of original data.)

Output / True	Voiced	Unvoiced	Silence	Error(%)
Voiced	30	0	0	0.0
Unvoiced	0	29	1	3.3
Silence	0	0	30	0.0

Table 2. Classification results(open test 10dB SNR).

Output / True	Voiced	Unvoiced	Silence	Error(%)
Voiced	30	0	0	0.0
Unvoiced	0	30	0	0.0
Silence	0	3	27	10.0

Table 3. Classification results(open test 3dB SNR).

Output / True	Voiced	Unvoiced	Silence	Error(%)
Voiced	30	0	0	0.0
Unvoiced	0	30	0	0.0
Silence	0	30	0	100.0

Table 4. Classification results(open test 0dB SNR).

Output / True	Voiced	Unvoiced	Silence	Error(%)
Voiced	30	0	0	0.0
Unvoiced	0	30	0	0.0
Silence	0	30	0	100.0

V. Conclusion

In this paper we have proposed the new V/U/S classification algorithm which uses neural netwroks as decision procedure and amplitude distribution characteristics of a frame as input. The computer simulation result shows that the proposed algorithm works well under extremely noisy environments in V/U classification, but fails in U/S classification. Even if we classify by the visual analysis or listen to the speech, it is almost impossible to classify U/S in very noisy environments.

The proposed algorithm needs no preprocessing of speech data and no threshold selection techniques and is robust to noise and is suitable for real time implementation.

Reference

1. B.S. Atal, and L.R. Rabiner, "A pattern recognition approach to voiced-unvoiced-silence classification with applications to speech recognition", IEEE Trans. ASSP, Vol. 24, No.3, pp.201-212, Jun., 1976.
2. I.S. LEE, J.A. CHOI, M.J. BAE, Souguil ANN, "An Algorithm for Detecting the Endpoints of Speech Signals by Amplitude Distribution", Proceedings of KITE Summer Conference '89, Vo. 12, No.1, pp.626-628, 1989.
3. H.Kobatake, "Optimization of Voiced / Unvoiced Decisions in Nonstationary Noise Environments', IEEE Trans. ASSP, Vol. 35, No.1, pp.9-18, Jan., 1987.
4. L.R. Rabiner, "Application of an LPC Distance Measure to the Voiced-Vnvoiced-Silence Detection Problem", IEEE Tran. ASSP, Vol.25, No.4, pp.338-343, Aug., 1977.
5. Y.H. Pao, Adaptive Pattern Recoghition and Neural Networks, Addison-Wesley, 1989.
6. D.E. Rumehart, Parallel Distributed Processing: Explorations in the Microstructures of Cognition. Vol. 1: Foundations, MIT Press.

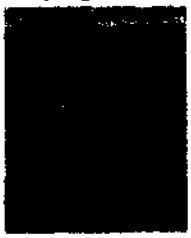
▲Inseop Lee

1986.3–1988.2 : Ph.D Course from Department of Electronics Engineering Seoul National University

▲Jungah Choi

1988.3–1990.2 : Ph.D Course from Dept. of Electronics Eng., Seoul National University

▲Myung Jin Bae



1981.2 : Department of Electronics Engineering, Soongsil University (B.S)

1983.2 : Department of Electronics Engineering, Seoul National University (M.S)

1987.8 : Ph.D course from Department of Electronics Engineering, Seoul National University

1986.3~ : Department of Electronics Engineering, Hoseo- University

▲Souguil Ann

1956.5 : Department of Electronics Engineering, Seoul National University (B.S)

1957.4 : Department of Electronics Engineering, Seoul National University (M.S)

1974.3 : Department of Electronics Engineering, Seoul National University (Ph.D)

1989.1~ : the president of the Acoustical Society of KOrea