NEURAL NETWORKS FOR VOICED/UNVOICED SPEECH CLASSIFICATION†

Aage Bendiksen and *Kenneth Steiglitz*
Department of Computer Science
Princeton University
Princeton, NJ 08544

ABSTRACT

This paper describes the results of designing, training, and testing a neural network for the voiced/unvoiced speech classification problem. A feed-forward multilayer back-propagation network was used with 6 input, 10 internal, and 2 output nodes — for a binary decision. The six features are common and easily computed. Training was done with 72 frames from two speakers; testing was done with 479 frames from four speakers; a total of 2 errors (0.4%) occurred. Thus a small neural network performs well on the V/UV problem.

1. Review of Previous Research

Voiced/Unvoiced (V/UV) classifiers can been grouped into two general categories [SL77]: 1) classifiers which determine the V/UV content of a segment of speech as a byproduct of an attempt to determine the primary pitch of the segment, and 2) classifiers which determine the V/UV content of a segment of speech by examining one or more speech signal features which are known to be correlated with the V/UV distinction. This paper describes a neural network to do V/UV classification using the latter approach. Our motivation was to produce an accurate V/UV classifier for high-quality analysis-synthesis using linear predictive coding.

Multi-feature techniques have some drawbacks, and it is important to be aware of the practical limitations imposed by these methods. One problem with the techniques presented below is their lack of immunity to non-stationary noise effects [AR76, CB80]. This problem is largely due to the fact that these methods do not incorporate some form of continuous adaptation to the environment being sampled—they involve training a V/UV classifier in a fixed noise environment, to operate in that same environment. For a discussion of V/UV classification techniques designed to operate in the presence of varying noise conditions, see [CB80, KH87, BG87, KS79]. In addition, the tradeoff between time and accuracy is important in many applications. In particular, it may not be practical to use a large number of features in real-time applications, especially if the features used are computationally complex. The ongoing development of digital signal processing hardware should cause a continuous reconsideration of such tradeoffs. In the analysis below, the techniques considered are applicable when high accuracy is desired, computation time is not a severe constraint, and noise effects can be determined in advance.

Two issues are involved in building a V/UV classifier of this general class. First, it is necessary to determine a set of speech features adequate to the V/UV decision task, and second, it is necessary to find an explicit rule for making the V/UV decision based upon the values of the features. Although these two aspects cannot be completely isolated from one another, it is useful to view them as separate parts of the problem. (Sometimes, as in [SB80], the set of features used is tailored to the classification method.) Separation allows a more systematic approach to each part.

Among the features that have been used or proposed for V/UV classification are:

1) rms or log(rms) energy of the signal [AR76, SL79a]
2) rms energy of the preemphasized signal [SL79a, SL77]
3) normalized autocorrelation coefficient at unit sample delay [AR76, SL79a]
4) normalized autocorrelation coefficient at unit sample delay, of the preemphasized signal [SL79a, SL77]
5) LPC normalized minimum error [AR76]
6) LPC normalized minimum error, of the preemphasized signal [SL79a, SL77]
7) first LPC predictor coefficient [AR76]
8) first LPC predictor coefficient, for the preemphasized signal [SL79a, SL77]
9) number of zero crossings in the signal [AR76, SL79a]
10) ratio of high frequency (above 4kHz) signal energy to low frequency (below 2kHz) signal energy [SL79a, SL77]
11) bispectrum of the signal [WB85]
12) low frequency energy [NE78, CT86]
13) an LPC distance measure [RL77]
14) an energy distance measure [RL77]
15) bit alteration rate of linear delta-modulated signal [UC80]
16) peak amplitude of the signal [SL79a, SL77]

†This work was supported in part by NSF Grant MIP-8705454, U. S. Army Research Office - Durham Contract DAAG29-85-K-0191.

CH2847-2/90/0000-0521 $1.00 © 1990 IEEE

17) ratio of preemphasized signal energy to normal signal energy [CT86]

18) causal pitch prediction gain [CT86]

19) non-causal pitch prediction gain [CT86]

Many of these features are highly correlated with one another (see [SL77, AR76]). [SL79b] described additional speech features useful for the V/UV/Mixed problem.

Although a variety of methods have been used to perform V/UV analysis based on a set of speech features, the discussion here will focus on pattern recognition techniques that have been successfully applied to the problem. These techniques have yielded classifiers that are among the most accurate available [AR76, SL77, SB80]. Other decision methods have been less general, and therefore applicable to only the specific speech features used (as in [CT86] or [NE78]), or have been unable to achieve accuracy as high as pattern recognition techniques have achieved (as in [RV80]). Pattern recognition techniques have a distinct advantage in that they are sufficiently general to be adapted to any set of available speech features.

2. Neural Network Classifiers

The feed-forward multilayer back-propagation network is particularly well suited to the V/UV classification problem. We use the logistic activation function

$$o_j(\mathbf{x}) = \frac{1}{1 + e^{-(\mathbf{w}_j^t \mathbf{x} + w_{0j})}}$$

where $o_j(\mathbf{x})$ is unit j's activation value when presented the input vector $\mathbf{x}$, $\mathbf{w_j}$ is a column vector containing unit j's input weights, and w_{0j} is unit j's scalar bias weight (see [RM86] for more details).

To train the network, we use the following back-propagation formula to modify a unit's weights:

$$\Delta w_{ji} = \eta \delta_j o_i,$$

in which Δw_{ji} is the change to be made to the weight of unit j's input from unit i, η is the learning rate, δ_j is an error signal available at unit j, and o_i is the activation value of unit i after the input training vector has been presented to the network as input and the network has settled. The way of computing the error signal δ_j for each unit depends on the type of unit involved. If unit j is an output unit, the error signal is computed as

$$\delta_j = (t_j - o_j) o_j (1 - o_j),$$

where t_j is the target vector element corresponding to output unit j. If unit j is a hidden unit, the error signal is computed as

$$\delta_j = o_j (1 - o_j) \sum_k \delta_k w_{kj}.$$

Here k indexes all units which have unit j as an input. δ_k is the error signal available at unit k, and w_{kj} is the weight of unit k's input from unit j. We use a feed-forward network, so all the units indexed by k are in layers succeeding the layer of unit j, and error signals propagate back from the output layer.

3. Feature Selection, Network Structure and Training

Six speech features were used with ten hidden units, and two output units. Each input unit output was connected to an input of every hidden unit and to an input of every output unit. Each hidden unit output was connected to an input of every output unit. Connection weights were initially set to small real pseudo-random numbers. The number of hidden units used in the network was near the high end of what seemed reasonable for the problem; a relatively large number of hidden units was used to safeguard against training difficulties.

The output units were meant to be binary indicators of voicing, the first output indicating unvoiced excitation, and the second output indicating voiced excitation. With this convention, an output vector of (1,0) would indicate a purely unvoiced frame, and a vector of (0,1) would indicate a purely voiced frame. Thresholds of 0.1 and 0.9 were used to make binary decisions.

Speakers were recorded on standard grade audio cassette tape, using a low fidelity portable tape recorder. Speech was digitized to 16 bits per sample, at a sampling rate of 28000 Hz. Features were computed for input frames of 256 samples, corresponding to a frame duration of 9.1 milliseconds, and a frame rate of 109 frames per second. The six input features used were

1) the rms energy of the signal:

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

2) the rms energy of the preemphasized signal:

$$f_2 = \left(\frac{1}{N}\sum_{i=1}^{N} p_i^2\right)^{1/2}$$

3) the signal's normalized autocorrelation coefficient at unit sample delay:

$$f_3 = \frac{\sum_{i=1}^{N} s_i s_{i-1}}{\left(\left(\sum_{i=1}^{N} s_i^2\right)\left(\sum_{i=0}^{N-1} s_i^2\right)\right)^{1/2}}$$

4) the preemphasized signal's normalized autocorrelation coefficient at unit sample delay:

$$f_4 = \frac{\sum_{i=1}^{N} p_i p_{i-1}}{\left(\left(\sum_{i=1}^{N} p_i^2\right)\left(\sum_{i=0}^{N-1} p_i^2\right)\right)^{1/2}}$$

5) the ratio of signal energy above 4000Hz to signal energy below 2000Hz:

$$f_5 = \frac{e_+(4000)}{e_-(2000)}$$

6) the product of signal energy above 4000Hz and signal

energy below 2000Hz:

$$f_6 = e_+(4000)e_-(2000)$$

Here s_i was the i^{th} sample of the analysis frame; s_0 was the last sample of the previous frame, and s_1 through s_{255} were samples taken from the input. Similarly, p_i was the i^{th} sample of the preemphasized analysis frame, which was obtained by using a preemphasis filter on the original signal:

$$p_i = s_i - .96s_{i-1}$$

In computing f_5 and f_6, $e_+(4000)$ was the frame's signal energy at frequencies above 4000Hz, and $e_-(2000)$ was the frame's signal energy at frequencies below 2000Hz. These signal energies were computed using a 256-point FFT with a Hamming window. Except for the autocorrelation coefficients, the input features were scaled so that their peak values over the duration of a sentence were equal to the fixed value 1.

The particular features used were selected for a number of reasons. Most compelling was the fact that f_1 through f_5 were used previously in successful classifiers [SS77]. f_6 is an indicator of mixed excitation [SL79b], and was included in the hope that the network could be taught to recognize mixed cases. Also, in comparison to other feature sets considered, this set of features was relatively straightforward to compute, requiring no computation more difficult than an FFT.

To build a training set, input frames were manually classified by examining graphs of their waveforms, as suggested in [AR76] and [SL77].

Seventy-two frames of training data were used in the training set. These frames were divided into two groups of 36 frames, one group from a male speaker, and one group from a female speaker. Each group contained 18 voiced frames and 18 unvoiced frames. The training frames were chosen so as to form as diverse a set as was practical. Each speaker's training data was taken from a single enunciation of the sentence "Chapter sixteen described a weird fictitious hut in which a pathetic knight pathetically polished a magical yellow gong," used in [SL77].

The training set was built up piece by piece. That is, a small group of frames was added to the set, and the network was allowed to train before more frames were added, then another group of frames were added, and the network was trained more, etc. This allowed some observation of the training behavior of the network. All of the frames for the male speaker were put into the training set before any of the frames from the female speaker were added. The first group of frames was completely misclassified. After a few thousand training presentations, with a learning rate of $\eta = 0.35$, the frames in the first group of training frames were all correctly classified to within about five percent of the target values (either 0 or 1). Subsequently, each time new training frames were added, most of the new frames were correctly classified without additional training. About 1,000 training set presentations were done after each addition of new frames to the training set. If the new frames were largely classified correctly, then this number was reduced to about 500. On some occasions when the new frames were poorly classified, the number of iterations was raised to 2,000. When the first group of frames from the female speaker was added, there were a number of classification errors. After this first group from the female speaker was used in training, the network again exhibited an ability to correctly classify most new frames at the time when they were added to the training set. When all the frames had been put into the training set, the network was correctly responding to all of the frames, with an error of no more than 0.05 or 0.06 on either network output for any given frame. That is, output units whose target value was 1.0 had output values of at least 0.94, and output units whose target value was 0.0 had output values of no more than 0.06. Once all of the frames were included, the network was allowed to train for an additional 25,000 iterations, in order to push the outputs even closer to the targets. Output errors were reduced to no more than 0.013 for any given output unit. The mean output error was 0.00068.

4. Testing the Classifier

A performance test was made of the classifier obtained. Test data was taken from four speakers, using the same sentence as was used in training. The frames used for testing were manually classified, in the same manner as the frames used for training. Two of the speakers, AB and LM, were the ones used to train the network. Two others, JE and VB, were speakers who were not included in the training exercise. AB and JE were male, VB and LM were female. The results of the test are summarized in table 1.

	JE	VB	AB	LM
Voiced Frames:	101	76	61	52
Unvoiced Frames:	68	48	38	35
Total Frames:	169	124	99	87
Errors:	0	2	0	0

Table 1: Experimental Test Results

There were a total of 479 frames included in the test set. Only two classification errors were observed; in further training these two misclassified frames would be added to the training set. Both errors occurred for speaker VB, and both occurred in unvoiced frames with low energy. This represents an error rate of 0.4% over the entire set of test data, and 1.6% over the data for VB.

5. Conclusions and Future Work

A small multilayer neural network appears to work well for the V/UV problem. A network with 6 input, 10 hidden, and 2 output nodes did not encounter difficulties, such as slow learning or local minima of the error function, that could have prevented successful training. The network trained on two speakers correctly classified 186 additional frames from those same test speakers, plus all but two frames out of 293 additional frames from two other speak-

ers. The observed error rate is at least as good as other reported classifiers based on pattern recognition, and the network is simple to implement and easy to train.

In future work we will study the relative power of additional internal nodes, sensitivity to choice of feature set, and extensions to the V/UV/Mixed problem.

References

[AR76] Atal, B. S, Rabiner, L. R., "A Pattern Recognition Approach to Voiced/Unvoiced/Silence Classification with Applications to Speech Recognition," *IEEE Trans. on ASSP*, vol. 24, pp. 201-212, June 1976.

[BG87] Bruno, G., et. al., "A Bayesian-Adaptive Decision Method for the Voiced/Unvoiced/Silence Classification of Segments of a Speech Signal," *IEEE Trans. on ASSP*, vol. 35, pp. 556-559, April 1987.

[CB80] Cox, B., "Nonparametric Rank-Order Statistics Applied to Robust Voiced-Unvoiced-Silence Classification," *IEEE Trans. on ASSP*, vol. 28, pp. 550-561, Oct. 1980.

[CT86] Campbell, P., Jr., "Voiced/Unvoiced Classification of Speech with Applications to the US Government LPC-10E Algorithm," *Proceedings of 1987 IEEE Int. Conf. on ASSP*, pp. 473-476.

[KH87] Kabatake, H., "Optimization of Voiced/Unvoiced Decisions in Nonstationary Noise Environments," *IEEE Trans. on ASSP*, vol. 35, pp. 9-18, Jan. 1987.

[KS79] Knorr, S., "Reliable Voiced/Unvoiced Decision," *IEEE Trans. on ASSP*, vol. 27, pp. 263-267, June 1979.

[NE78] Neuberg, E., "Improvement of Voicing Decisions by Use of Context," *Proceedings of 1978 IEEE Int. Conf. on ASSP*, pp. 5-7.

[RL76] Rabiner, et. al., "A Comparative Performance Study of Several Pitch Detection Algorithms," *IEEE Trans. on ASSP*, vol. 24, pp. 399-418, Oct. 1976.

[RM86] Rumelhart, D.E., McClelland, J.L., *Parallel Distributed Processing*, vol. 1, MIT Press, Cambridge Mass., 1986, pp. 318-362.

[RV80] Ramamoorthy, V., "Voiced/Unvoice Detection Based on a Composite-Gaussian Source Model of Speech," *Proceedings of 1980 IEEE Int. Conf. on ASSP*, pp. 57-60.

[SB80] Siegel, L., Bessey, A., "A Decision Tree Procedure for Voiced/Unvoiced/Mixed Excitation Classification of Speech," *Proceedings of 1980 IEEE Int. Conf. on ASSP*, pp. 53-56.

[SL77] Siegel, L., *A Pattern Classification Algorithm for the Voiced/Unvoiced Decision in Speech Analysis*, Ph.D. dissertation, Princeton University Dept. of EECS, May 1977.

[SL79a] Siegel, L., "A Procedure for Using Pattern Classification Techniques to Obtain a Voiced/Unvoiced Classifier," *IEEE Trans. on ASSP*, vol. 27, pp. 83-88, Feb. 1979.

[SL79b] Siegel, L., "Features for the Identification of Mixed Excitation in Speech Analysis," *Proceedings of 1979 IEEE Int. Conf. on ASSP*, pp. 752-755.

[SS76] Siegel, L., Steiglitz, K., "A Pattern Classification Algorithm for the Voiced/Unvoiced Decision," *Proceedings of 1976 Int. Conf. on ASSP*, pp. 326-329.

[UC80] Un, C., "Voiced/Unvoiced/Silence Discrimination of Speech by Delta Modulation," *IEEE Trans. on ASSP*, vol. 28, pp. 398-407, Aug. 1980.

[WB85] Wells, B. "Voiced/Unvoiced Decision based on the Bispectrum," *Proceedings of 1985 IEEE Int. Conf. on ASSP*, pp. 1589-1592.