A PATTERN CLASSIFICATION ALGORITHM FOR THE VOICED/UNVOICED DECISION

L. J. Siegel and K. Steiglitz
Electrical Engineering Department
Princeton University, Princeton, New Jersey

Abstract: An algorithm for making the voiced/unvoiced decision in speech analysis is presented. Three features (LPC normalized minimum error, ratio of energy content at high to low frequencies, and input RMS) define a three-dimensional space in which the decision making process is viewed as a pattern classification problem. This is formulated as a linear program which runs on a training set to find a hyperplane dividing the V/UV regions if they are separable, or minimizing the distance by which misclassification occurs if they are not. A procedure is given for selecting the features and constructing the training set.

Introduction

Deciding whether a short segment of speech is voiced or unvoiced is an important problem in speech analysis. Most V/UV decision methods have been based on one parameter, sometimes accompanied by additional logic to take into account the V/UV decision for surrounding segments [1-6]. No single feature seems to give consistently reliable performance in making the V/UV decision, so it is desirable to combine several features to obtain a good characterization of voiced and unvoiced segments of speech. One way to incorporate a number of features is to view the V/UV decision process as a pattern classification problem. Atal and Rabiner [7] have used a statistical model to design a minimum distance classifier. This requires assuming a particular distribution function for the features and computing the mean and covariance matrix for each class using a large enough set of data to obtain an accurate statistical characterization.

We present a nonparametric pattern classification technique for making the V/UV decision in which a relatively small set of samples is used to "train" the classifier. This training is accomplished by a linear program which finds a hyperplane dividing the V/UV regions if they are separable, or minimizing the distance by which misclassification occurs if they are not. No assumptions are made about the forms of the probability distributions of the features. In addition, a method is described for constructing the training set and for selecting the features to be used.

This work was supported by NSF Grant GK-42048 and the U.S. Army Research Office - Durham under Grant DAHC04-75-0192.

Theory of Linear Separability

We will use pattern classification techniques that are described more fully in [8] and [9]. Let d be the number of features to be used and let the two classes be C_1 and C_2. Then a pattern $\underline{x} = (f_1, f_2, \ldots, f_d)$ is a point in E^d (d-dimensional Euclidean space) with f_i being the value of the ith feature. A training set will consist of n patterns $\underline{x}_i$, i = 1,...,n which will be used to "train" the classifier to identify the two classes correctly. We wish to find a discriminant function g(x) such that

$$g(\underline{x}_i) > 0 \quad \text{if } \underline{x}_i \in C_1$$
$$g(\underline{x}_i) < 0 \quad \text{if } \underline{x}_i \in C_2$$

for i = 1, ..., n. If, for the sake of simplicity, we require g to be a linear function, then the decision surface separating the two classes is a hyperplane defined by

$$g(\underline{x}) = w_1 f_1 + w_2 f_2 + \ldots + w_d f_d + w_{d+1} = 0$$

or equivalently

$$g(\underline{u}) = \underline{w}^t \underline{u} = 0$$

where $\underline{w} = (w_1, w_2, \ldots, w_d, w_{d+1})$ is the weight vector and $\underline{u} = (f_1, f_2, \ldots, f_d, 1)$ is the augmented pattern vector corresponding to the pattern $\underline{x}$. If the patterns to be classified can be separated by such a hyperplane, then the two classes are said to be linearly separable. In this case, we will be able to find a weight vector $\underline{w}$ such that

$$\underline{w}^t \underline{u}_i > 0 \quad \underline{u}_i \in C_1$$
$$\underline{w}^t \underline{u}_i < 0 \quad \underline{u}_i \in C_2 \quad i = 1, \ldots, n$$

where $\underline{u}_i$ is the augmented pattern vector corresponding to pattern $\underline{x}_i$ and we say $\underline{u}_i \in C_i$ iff $\underline{x}_i \in C_j$. If no such hyperplane exists, we wish to find the hyperplane which by some measure minimizes misclassification.

An equivalent statement of the linear separation problem is that we want a weight vector $\underline{w}$ such that

$$\underline{w}^t \underline{y}_i \geq b > 0$$

where $\underline{y}_i = \begin{cases} \underline{u}_i & \text{if } \underline{u}_i \in C_1 \\ -\underline{u}_i & \text{if } \underline{u}_i \in C_2 \end{cases}$

for i = 1, ..., n. The value b is a positive margin to avoid the solution $\underline{w}=0$ and to force the solution away from the boundary $\underline{w}^t \underline{u}_i = 0$. To handle the non-separable case, define the perceptron criterion function

$$p(\underline{w}) = \sum_{\underline{w}^t \underline{y}_i \leq b} (b - \underline{w}^t \underline{y}_i).$$

$p(\underline{w})$ is proportional to the sum of the distances from the misclassified patterns to the decision surface. We can get a useful weight vector $\underline{w}$ even if the n patterns in the training set are not linearly separable by finding $\underline{w}$ and $\underline{t}$ to minimize

$$z = \sum_{i=1}^{n} t_i$$

where we require

$$t_i \geq 0 \quad \text{and} \quad t_i \geq b - \underline{w}^t \underline{y}_i$$

for i=1, ..., n. For a fixed $\underline{w}$, the minimum value for z results in t_i being the "cost" of $\underline{w}$'s classification of the ith pattern: if $\underline{w}$ classifies $\underline{y}_i$ correctly, $t_i = 0$; otherwise $t_i = b - \underline{w}^t \underline{y}_i$. Thus $z = \sum_{i=1}^{n} t_i = p(\underline{w})$. Therefore, minimizing z over $\underline{t}$ and $\underline{w}$ will yield a separating hyperplane if one exists, or will minimize the perceptron criterion function $p(\underline{w})$ if a separating hyperplane does not exist.

This can be formulated as a linear programming problem as follows:

$$\min \underline{c}^t \underline{x}$$
$$A\underline{x} \geq B$$
$$\underline{x} \geq 0$$

where

$$A = \begin{bmatrix} \underline{y}_1^t & -\underline{y}_1^t & 1 & 0 & \cdots & 0 \\ \underline{y}_2^t & -\underline{y}_2^t & 0 & 1 & 0 & \cdots 0 \\ \vdots & & & & & \\ \underline{y}_n^t & -\underline{y}_n^t & 0 & \cdots & 0 & 1 \end{bmatrix}$$

$$\underline{x} = \begin{bmatrix} \underline{w}^+ \\ \underline{w}^- \\ \underline{t} \end{bmatrix} \qquad \underline{c} = \begin{bmatrix} 0 \\ \vdots \\ 0 \\ \vdots \\ 1 \\ \vdots \\ 1 \end{bmatrix} \begin{matrix} \\ d+1 \\ \\ d+1 \\ \\ n \end{matrix}$$

$$B = [b \ \cdots \ b]^t \qquad \underline{w} = \underline{w}^+ - \underline{w}^-.$$

A basic feasible solution is $\underline{w}=0$, $\underline{t}_i = b$ for i=1,...,n. If the patterns in the training set are linearly separable, $\underline{t}=0$ and $\underline{w}$ will define a separating hyperplane. If the patterns are not linearly separable, $\underline{t}>0$ and $\underline{w}$ minimizes the perceptron criterion function.

With the addition of slack variables to convert the inequality $A\underline{x} \geq B$ to an equality, the linear program has n constraints and 2n+2(d+1) variables.

Training and Performance

The "correct" V/UV decisions for the sentences used in training and testing were made by inspection of the waveforms. In cases where it was not possible to make the V/UV decision based solely on the waveforms, an auditory comparison was made between the speech synthesized with the voiced decision and the unvoiced decision. Analysis was performed using the covariance method of linear predictive coding with 18 poles in the all-pole filter, a sampling rate of 15000 Hz, and a frame length of 250 samples. The synthesis was performed pitch synchronously, with the pitch, output energy, and predictor coefficients reset by interpolation at the beginning of each pitch period during voiced speech, and reset on the analysis frame boundary during unvoiced speech [6,10]. Since the covariance method of analysis does not insure stability, poles found outside the unit circle were folded in to a radius of .986 (corresponding to a bandwidth of about 60 Hz). Synthesis was performed with a monotone pitch of 125 Hz to isolate the V/UV decision from the pitch tracking problem. Frames in which a transition occurred so that the frame was voiced for one portion and unvoiced for another were not considered in the training or evaluation.

The features considered were:

1. RMS value of the input signal = $\left(\frac{1}{250}\sum_{i=1}^{250} s_i^2\right)^{1/2}$ (RMS);
2. Zero crossings (ZC);
3. Peak amplitude (PEAK);
4. LPC normalized minimum error = $\frac{1}{250}\sum_{i=1}^{250} (\hat{s}_i - s_i)^2 \Big/ \frac{1}{250}\sum_{i=1}^{250} s_i^2$

where $\hat{s}_i$ is the LPC approximation of sample s_i (ERRN);

5. Ratio of the energy of the signal above 4000 Hz to that below 2000 Hz (HILO). This was computed by a 256 point FFT on the Hanning windowed segment consisting of the 6 last points of the previous frame and the 250 points of the current frame.

Three sentences (A, B, and C) which together included instances of all non-vowel phonemes, and 3 male speakers (LV, DH, HJ) were considered for the training process, resulting in a set of 9 possible training sentences. From these, the goal was to obtain as small as possible a training set which would produce good separation of the voiced and unvoiced classes. The following procedure describes a method for choosing the features to be used and selecting the training set in such a manner that the effect of each addition to the set of features or the training set can be evaluated.

1. Evaluate separability using 1 or 2 features.

Pairwise scatter plots of the features for 1 sentence and 1 speaker (sentence A, speaker LV) indicated that: (a) no single feature was sufficient to make the V/UV decision; (b) RMS and PEAK were highly correlated, so only one of them should be considered; (c) the combination of HILO and either RMS or PEAK gave reasonably good separation (figure 1); (d) HILO and ERRN gave fair separation; (e) no other pairs gave good separation. The HILO-RMS combination on the other training sentences proved to be insufficient (figure 2). This demonstrated the need for at least 3 features. The above observations could

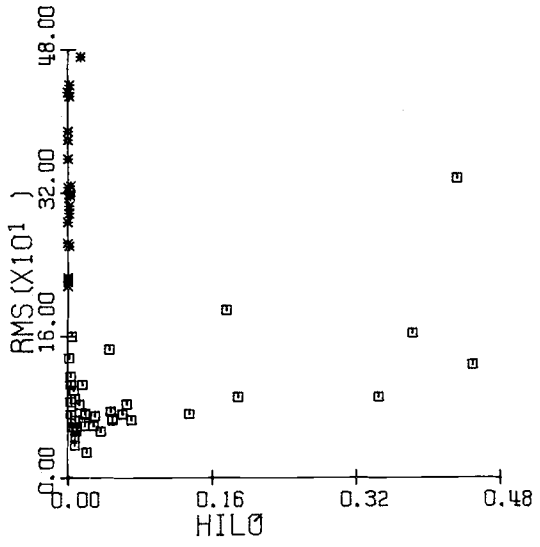

Fig. 1. Portion of plot of voiced (*) and unvoiced (□) frames - sentence A , speaker LV.

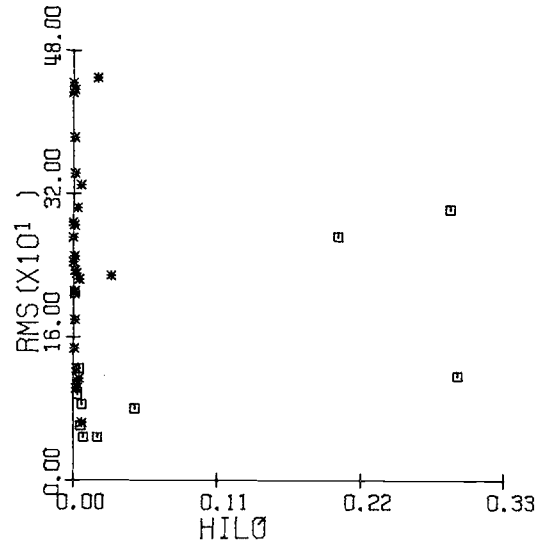

Fig. 2. Portion of plot of voiced (*) and unvoiced (□) frames - sentence B , speaker LV.

also be made by running the linear program with the various pairs of features, and a training set consisting of one training sentence, using the resulting cost z as the measure of separability.

2. Construct a training set using 3 features, 1 speaker.

Because of (c) and (d) above, ERRN was added to HILO and RMS to form a 3-feature classifier. The training set consisted of all patterns (frames) by a single speaker (LV) which were within a fixed distance of the hyperplane for the HILO-RMS case of step (1). These <u>boundary patterns</u> are the ones most likely to be misclassified. Since the value of the discriminant function g is a measure of the distance from the hyperplane, the training set can be constructed by choosing all patterns x_i for which $|g(x_i)|$ for the HILO-RMS case is less than some distance δ. The resulting weight vector correctly classified most frames of the training sentences spoken by two of the speakers (LV and HJ) but misclassified several frames by the third (DH).

3. Augment the training set.

Using the same distance δ and the hyperplane of the HILO-RMS case, the boundary patterns for the speaker for which the previous training set was inadequate were added to the training set. The resulting weight vector performed extremely well on all 3 speakers.

4. Test additional features.

The patterns in the training set of step (3) were not linearly separable. The addition of a fourth feature, ZC, did not significantly improve the separability on the training set.

The training set described in step (3) used 3 features and contained 103 patterns. 53 of the patterns were chosen in step (2) from sentences by speaker LV, and 50 were added in step (3) from sentences by speaker DH. The resulting linear program had 103 constraints and 214 variables. Using the revised simplex algorithm, it required maintaining a (104,104) matrix and ran in approximately 7.5 seconds on an IBM 360/91. The weight vector which minimized the perceptron criterion function was found after 131 iterations.

Testing was performed on 5 sentences spoken by each of the 3 speakers. We define a type 1 error to be the classification of a voiced frame as unvoiced, and a type 2 error to be the classification of an unvoiced frame as voiced. The following table summarizes the results of the testing. The rows represent the 3 speakers, and the columns the 5 sentences. An entry i,j means that i type 1 errors and j type 2 errors were committed. If the table entry includes values in parentheses, frames of that sentence were used in the training set from which the weight vector was computed, and the values in parentheses indicate the number of (type 1, type 2) misclassifications on members of the training set. The average number of frames in each sentence was 170.

	A	B	C	D	E
LV	1,1 (1,0)	4,1 (2,1)	0,2 (0,1)	0,0	0,0
DH	3,3 (2,2)	1,0 (0,0)	2,1	1,1	0,0
HJ	0,0	1,0	0,1	0,0	0,0

The misclassification rate over the sentences tested was approximately 0.9 %. No audible misclassifications were made.

Conclusions

We have presented a pattern classification technique for making the V/UV decision based on several parameters. The method uses a linear program to find the decision surface which minimizes the misclassification distance over a training set. A procedure was presented for selecting a training set and for evaluating and choosing features to be used in the V/UV decision making process. The training set found sufficient to obtain a good classifier for 3 male speakers, including one not used in the training set, was relatively small (3 features, 103 patterns). In limited testing, the classifier performed with a misclassification rate of less than 1 %.

Further testing is necessary to determine the sensitivity of the technique to different classes of speakers (e.g., women, children) and to recording conditions. More features may be examined to determine the combination which works best over the broadest range of speakers and recording conditions. Another area of further research involves extension of the model to allow mixed voiced-unvoiced excitation. Since the discriminant function is a measure of the distance to the decision hyperplane, this function may be useful in indicating the degree of voicing. Such a measure could be used to determine a mixed voiced-unvoiced excitation, as found in voiced fricatives. This possibility is being examined.

References

1. J.D. Markel, "Formant trajectory estimation from a linear least - squares inverse filter formulation," Speech Comm. Res. Lab., Santa Barbara, Cal., Monograph 7, Aug. 1971.

2. _____, "Application of a digital inverse filter for automatic formant and F0 analysis," IEEE Trans. Audio and Elect., vol. AU-21, no. 3, pp. 154-160, June 1973.

3. A.M. Noll, "Cepstrum pitch determination," J. Acoust. Soc. Amer., vol. 41, pp. 293-309, Feb. 1967.

4. M.M. Sondhi, "New methods of pitch extraction," IEEE Trans. Audio and Elect., vol. AU-16, no. 2, pp. 262-266, June 1968.

5. M.J. Ross, et al., "Average magnitude difference function pitch extractor," IEEE Trans. Acoust., Speech, Signal Processing, vol. ASSP-22, no. 5, pp. 353-362, Oct. 1974.

6. B.S. Atal and S.L. Hanauer, "Speech analysis and synthesis by linear prediction of the speech wave," J. Acoust. Soc. Amer., vol. 50, pp. 637-655, Aug. 1971.

7. B.S. Atal and L.R. Rabiner, "A pattern - recognition approach to voiced - unvoiced - silence classification with applications to speech recognition," unpublished memorandum, Bell Telephone Laboratories, Murray Hill, N.J., 1975.

8. R.O. Duda and P.E. Hart, Pattern Classification and Scene Analysis, John Wiley and Sons, N.Y., ch. 5, 1973.

9. N.J. Nilsson, Learning Machines, McGraw Hill, N.Y., 1965.

10. J.D. Markel and A.H. Gray, "A linear prediction vocoder simulation based upon the autocorrelation method," IEEE Trans. Acoust., Speech, Signal Processing, vol. ASSP-22, no. 2, pp. 124-134, Apr. 1974.