Voice and Unvoice Decision

Introduction

The need for deciding whether a given segment of a speech waveform should be classified as voiced speech, unvoiced speech, or silence arises in many speech analysis systems. A variety of approaches have been described in the literature for making this decision. In this project, we use a pattern recognition approach for classifying a given speech segment, which is described in [6]. The pattern recognition approach provides an effective method of combining the contributions of a number of speech measurements - which individually may not be sufficient to discriminate between classes - into a single measure of speech capable of providing reliable separation between the three classes. The method is essentially a classical hypothesis testing procedure based on the statistical decision theory. In this method, for each of the three classes, a non-Euclidean distance measure is computed from a set of measurements made on the speech segment to be classified and the segment is assigned to the class with minimum distance.

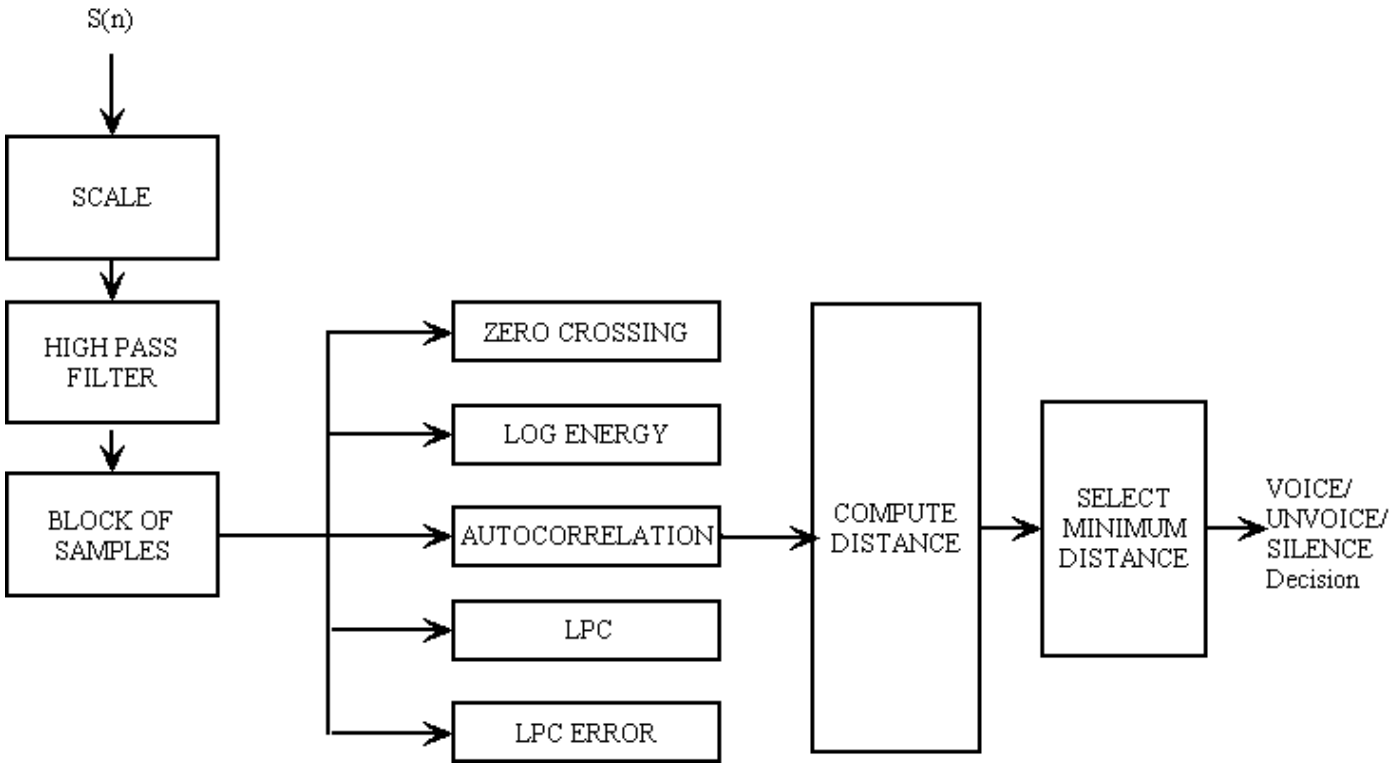
The success of a hypothesis-testing depends, to a considerable extent, upon the measurements or features which are used in the decision criterion. The basic problem is of selecting features which are simple to derive from speech and yet are highly effective in differentiating between the three classes. The following five measurements have been used in the implementation described in this report:

- Energy of the signal
- Zero-crossing rate of the signal
- Autocorrelation coefficient at unit sample delay
- First predictor coefficient
- Energy of the prediction error

The choice of these particular parameters is based on the experimental evidence that the parameters vary consistently from one class to another, which we will discuss later in this part.

Speech Measurements

A block diagram of the analysis and decision algorithm is shown in the following figure.



Prior to analysis, the speech signal is high pass filtered to remove any dc, low frequency hum, or noise components which might be presented. The formula of the high pass filter is given below.

$$H(z) = \frac{1 - 2z^{-1} + z^{-2}}{1 - e^{-aT}\cos(bT)z^{-1} + e^{-2aT}z^{-2}}$$

Then the five parameters mentioned before are computed for each block of samples. Following we state in detail the definitions of them.

1. Zero-crossing count N_Z, the number of zero crossing in the block

The zero crossing count is an indicator of the frequency at which the energy is concentrated in the signal spectrum. Voiced speech is produced as a result of excitation of the vocal tract by the periodic flow of air at the glottis and usually shows a low zero crossing count. Unvoiced speech is produced due to excitation of the vocal tract by the noise-like source at a point of constriction in the interior of the vocal tract and shows a high zero crossing count. The zero crossing count of silence is expected to be lower than for unvoiced speech, but quite comparable to that for voiced speech.

2. Log energy E_s -- defined as

$$E_s = 10\log(\varepsilon + \frac{1}{N}\sum_{n=1}^{N} S^2(n))$$

Where ε is a small positive constant added to prevent the computing of log of zero. Generally speaking, E_s for voiced data is much higher than the energy of silence. The energy of unvoiced data is usually lower than for voiced sounds but higher than for silence.

3. Normalized autocorrelation coefficient at unit sample delay, C_1 which is defined as

$$C_1 = \frac{\sum_{n=1}^{N} s(n)s(n-1)}{\sqrt{(\sum_{n=1}^{N} s^2(n))(\sum_{n=0}^{N-1} s^2(n))}}$$

This parameter is the correlation between adjacent speech samples. Due to the concentration of low frequency energy of voiced sounds, adjacent samples of voiced speech waveform are highly correlated and thus this parameter is close to 1. On the other hand, the correlation is close to zero for unvoiced speech.

4. First predictor coefficient, α_1 of a 12-pole linear predictive coding analysis using the covariance method. It can be shown that this parameter is the negative of the Fourier component of the log spectrum at unit sample delay. Since the spectra of the three classes -- voiced, unvoiced, silence -- differ considerably, so does the first LPC coefficient.

5. Normalized prediction error, E_p, expressed in DB, which is defined as

$$E_p = E_s - 10\log(10^{-6} + |\sum_{k=1}^{P} \alpha_k \phi(0,k) + \phi(0,0)|)$$

$$\phi(i,k) = \frac{1}{N}\sum_{n=1}^{N} s(n-i)s(n-k)$$

Where E_s is defined above and Ø (i, k) is the (i, k) term of the covariance matrix of the speech samples, and α_k s are the predictor coefficients. This parameter is a measure of the non-uniformity of the spectrum.

The five parameters discussed above are correlated with each other. These correlations vary between the parameters and between classes. The decision algorithm discussed in the next section will make use of it to differentiating between the classes

Decision Algorithm

As mentioned before, the five measurements are used to classify the block of the signal as either silence, unvoiced, or voiced speech. To make this decision, a classical minimum probability of error decision is used in which it is assumed that the joint probability density function of the possible values of the measurements for the ith class is a multidimensional Gaussian distribution with known mean m_i and covariance matrix W_i. i=1,2,3 corresponds to class 1 (silence), class 2 (unvoiced), and class 3 (voiced), respectively.

For the decision rule, the distribution of the measurement does not need to be necessarily exactly normal . In the case of unimodal distributions, it is sufficient that the distribution be normal in the center of its range, which is often true for physical measurements.

Let x be an L dimensional column vector (in our case L=5) representing the measurements, that is the kth component is the kth measurement. The L-dimensional Gaussian density function for x with mean vector m_i and covariance matrix W_i is given by

$$g_i(X) = (2\pi)^{-L/2} |W_i|^{-1/2} \exp(-\frac{1}{2}(X - M_i)^H W_i^{-1}(X - M_i))$$

The decision which minimizes the probability error states that the measurement vector x should be assigned to class i if

$$p_i g_i(X) \geq p_j g_j(X)$$

Where Pi is the a priori probability that x belongs to the ith class. This decision rule, by throwing away some insignificant parts and manipulations, can be further simplified: the quantity distance di^ defined as

$$\hat{d_i} = (X - M_i)^H W_i^{-1} (X - M_i))$$

is computed and the index i is chosen such that di^ is minimized.

Estimation of the Means and the Covariances

In order to use the above decision algorithm, a training set of data is required to obtain the mean vector and the covariance matrix for each class. This training set is created by manually segmenting natural speech into regions of silence, unvoiced speech and voiced speech. The measurements mentioned above are made on each block of data. Let xi denotes the measurement vector for nth block for class (i=1, 2, 3) and Ni denotes the number of the blocks manually classified as class i in the training set, then

$$W_i = \frac{1}{N_i} \sum_{n=1}^{M} x_i(n)\, x_i^T(n) - M_i M_i^H$$

$$M_i = \frac{1}{N_i} \sum_{n=1}^{M} x_i(n)$$

Results

In order to use the above decision algorithm, a training set of data is required to obtain the mean vector and the covariance matrix for each class (Silence, Unvoiced and Voiced). The raw training set was preprocessed (e.g., high-pass filters and data/shift windows) as that done for the actual data set. The training set is then analyzed by manually segmenting natural speech into regions of silence, unvoiced speech, and voiced speech. The speech segments are sub-divided into 32-millisecond blocks (i.e., 256 samples in each block on speech of 8K sampling rate), and the set of 5 different measurements defined in previous section is made on each block of data.

Table 1 shows the means, standard deviations, and the normalized covariance matrices* for the three classes for a typical set of training data. The columns in Table 1 correspond to the five measurements discussed earlier. The off-diagonal terms of the covariance matrices are a measure of the correlation between the different parameters. If the measurements were all independent and uncorrelated, then all off-diagonal elements would be 0. It can be seen that the magnitudes of the off-diagonal elements vary from 0.18 to 0.94, indicating varying degree of correlations between the different parameters.

	Zero Crossings	Log Energy	First Auto-correlation	First LPC	LPC log Error
1) Silence					
Mean	9.6613	-38.1601	0.9489	0.5084	-10.8084
Covariance	1.0000	0.6760	-0.7077	-0.1904	0.7208
matrix	0.6760	1.0000	0.6933	0.2918	-0.9425
(normalized)	-0.7077	0.6933	1.000	0.3275	-0.8426
	-0.1904	0.2918	0.3275	1.0000	-0.2122
	0.7208	-0.9425	-0.8426	-0.2122	1.0000
2) Unvoiced					
Mean	10.4286	-36.7536	0.9598	0.5243	-10.9076
Covariance	1.0000	0.6059	-0.4069	0.4648	-0.4603
matrix	0.6059	1.0000	-0.1713	0.1916	-0.9337
(normalized)	-0.4069	-0.1713	1.0000	0.1990	-0.1685
	0.4648	0.1916	0.1990	1.0000	-0.2121
	-0.4603	-0.9337	-0.1685	-0.2121	1.0000
3) Voiced					
Mean	29.1853	-18.3327	0.9826	1.1977	-11.1256
Covariance	1.0000	-0.2146	-0.8393	-0.3362	0.3608
matrix	-0.2146	1.0000	0.1793	0.6564	-0.7129
(normalized)	-0.8393	0.1793	1.0000	0.3416	-0.5002
	-0.3362	0.6564	0.3416	1.0000	-0.4850
	0.3608	-0.7129	-0.5002	-0.4850	1.0000

Table1. Means and Covariance Matrices for the three classes for the training data.

The algorithm has been tested on training data and testing data respectively. The speech data in the training set consisted of utterance "The rain in the Spain spreads mainly in the greatly plain" spoken by a female speaker. The testing set was used to evaluate the performance of the algorithm. The speech data in the testing set consisted of utterance "We predict the unknown----that is linear prediction, linear prediction, vocoder" spoken by a female speaker.

The matrix of incorrection (confusion matrix) was used to evaluate how well the algorithm performs with the speech data. The algorithm was first run on the training set itself and then was used on the speech data in the test set. The confusion matrices for the two cases are presented in Table 2 and Table 3. Most of the identifications are correct, a few errors occurred at the boundaries between the different classes. Since the classification was made on the basis of consecutive 32-ms long speech segments, a segment at the boundary often included data from two classes.

Actual class	Silence	Unvoiced	Voiced
Identified as			
Silence	61	2	0
Unvoiced	1	54	1
Voiced	0	0	258
Total	62	56	259

Table 2. Matrix of incorrect identifications for the three classes for the speech data in the *training* set.

Actual class	Silence	Unvoiced	Voiced
Identified as			
Silence	14	2	0
Unvoiced	2	21	0
Voiced	0	0	139
Total	16	23	139

Table 3. Matrix of incorrect identifications for the three classes for the speech data in the *testing* set.

An example of the speech waveform showing the various voiced, unvoiced, and silence regions as determined by the algorithm is shown in Fig. 1.

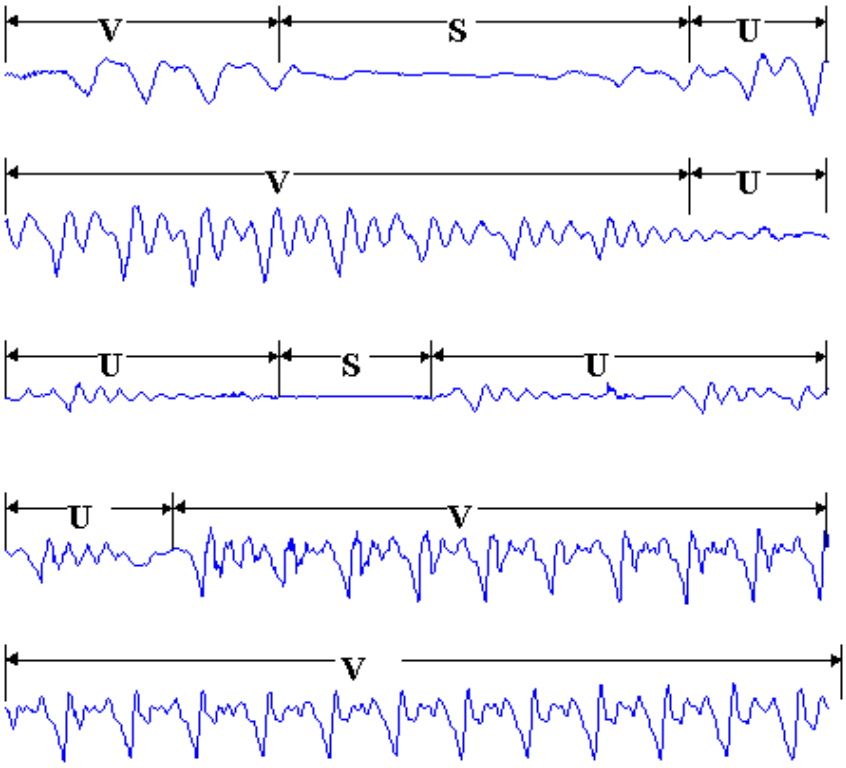


Figure 1. Comparison between actual data and V/U/S determination results.

To show the advantages of the pattern recognition algorithm based on the combination of 5 individual measurement parameters, we compare its performance with the results on the speech data in the test set using only one of the five measurement parameters at a time.

The total number of errors found for each class for each of the five parameters is shown in Table 4. For comparison, the corresponding errors where all of the five parameters are used are also shown on the last row of Table 4. It can be seen that none of the parameters by itself is capable of identifying a class with sufficiently high accuracy. The performance of the five parameters when used in combination is quite good in view of the fact that most of the errors occur at the boundaries between the different classes. Moreover, it can be found from Table 4 that the performance of any of the parameters is not equally good for discriminating between all of the three classes.

Number	Parameter Used	Errors		
		Silence	Unvoiced	Voiced
1	Zeros-crossing count	10	21	12
2	Log energy	4	13	9
3	Normalized autocorrelation coefficient	5	17	7
4	First predictor coefficient of LPC analysis	15	24	10
5	Normalized prediction error	7	18	3
6	Pattern recognition using all five parameters	1	2	1

Table 4. Total number of identification errors for the different classes with different sets of parameters. The total number of segments was 62 for the silence, 56 for the unvoiced, and 259 for the voiced class.

Discussions

A fairly general framework based on a pattern recognition approach to VUS classification has been described in which a set of measurements are made on the interval being classified, and a minimum non-Euclidean distance measure is used to select the appropriate class. Almost any set of measurements can be used so long as there is some physical basis for assuming that the measurements are capable of reliably distinguishing between these three classes.

The major limitation of the method is the necessity for training the algorithm on the specific set of measurements chosen. Strictly speaking, the training data is particular to one set of recording conditions. Thus, whenever the transmission system varies or the background noise level varies, a new set of training data is required. If the recording conditions differ considerably from one occasion to another, it may be possible to adapt the algorithm by continuously updating the training data based on some measure of the relative distances to each of the classes.