One Channel Five-Way Classification Algorithm For Automatically Classifying Speech

Kyo-Sik Lee*, Kyu-Sik Park**

Abstract

In this paper, we describe the one channel five-way, V/U/M/N/S (Voice/Unvoice/Mixed /Nasal/Silent), classification algorithm for automatically classifying speech. The decision making process is viewed as a pattern recognition problem. Two aspects of the algorithm are developed: feature selection and classifier type. The feature selection procedure is studied for identifying a set of features to make V/U/M/N/S classification. The classifiers used are a vector quantization (VQ), a neural network(NN), and a decision tree method. Actual five sentences spoken by six speakers, three male and three female, are tested with proposed classifiers. From a set of measurement tests, the proposed classifiers show fairly good accuracy for V/U/M/N/S decision.

I. Introduction

In the most commonly used model of speech production, the speech signal is decomposed into a filter component and an excitation component. The excitation component is represented by one of two states: voiced-more or less periodic, produced by vibration of the vocal cords, or unvoiced-noise like, produced by forcing air past some constriction in the vocal tract. Using this model, considerable success has been achieved by employing pattern classification techniques to assign a segment of speech to one of two classes, voiced or unvoiced[1][2]. Despite the widespread use of this simplified model, the restriction of the excitation to the two classes is not adequate for the synthesis of high quality speech from analysis parameters. Experiments show that high quality synthesis requires mixed excitation for synthesis of the voiced fricatives (v(vote), th(then), z(zoo), zh(azure)). Human production of these sounds involves the vibration of the vocal cords in conjunction with a turbulent air flow at some point of constriction. Speech synthesizers driven from stored data rather than from analysis parameters commonly include a link between the unvoiced and voiced excitation paths to allow a mixed excitation in the synthesized speech. In order to allow a mixed source in an analysis-synthesis system, the excitation for a segment of speech must be identified as voiced, unvoiced, or a combination of voiced and unvoiced.

The acoustic structure of nasal consonants has long been predicted by the acoustic theory of speech production[3][4]. The presence of a side-branching resonator (the blocked oral cavity) will introduce an antiresonance in the spectrum of nasal consonants. Theoretically, the antiresonance can be used to identify the place of articulation of nasal consonants because the frequency of the antiresonance is determined by the dimension of the side-branching resonator. Such differences, however, are difficult to detect using conventional techniques of spectral analysis. The presence of the spectral zero introduces nonlinear equations to parametric methods of spectral analysis[5]. Nasal murmurs and vowel nasalization are approximated by the insertion of an additional resonator and anti-resonator into the cascade vocal tract model in the formant synthesizer The Klatt synthesizer includes an additional resonance-antiresonance pair for synthesizing nasals[6]. It is necessary to identify nasalized segments in the recorded speech in order to decide when this branch should be activated. Another purpose of nasality detection is the correction of the all-pole estimates of the formant frequencies and bandwidths; it is known that the presence of zeros in the spectra of nasal speech tends to move the formants upwards in frequency and to increase their bandwidths. Therefore, to get the synthesis of high quality speech from analysis parameters, extending the V/U/S or the V/U/M/S decision to V/U/M/N/S decision is needed.

* Dept. of Information and Telecommunication Hansei University
** Dept. of Information and Telecommunication Sangmyung University
Manuscript Received : March 17, 1998.

About the previous works on classifying algorithm, a two-channel four-way V/U/M/s classification algorithm was described in [7][8] using both the speech and the EGG (electroglottogram) signal. However, the EGG signal is usually unavailable in real situations and a designer has to design a speech system that relies only on speech input. Therefore, it is not possible to take advantage of the EGG signal as an indicator of vocal fold vibration and, as a result, the system becomes more complicated.

The main objective of this paper is to develop the one channel five-way, V/U/M/N/S classification algorithm for automatically classifying speech. Base on the pattern recognition technique, the classification algorithm is considered as feature selection and classifier type. This paper is outlined as follows.: Section II. describes proposed feature extraction method for the V/U/M/S and nasal/non-nasal decision. In section III, the methods for making five-way V/U/M/N/S classification decision are explained. Three classifier types are considered such as vector quantization(VQ), neural network(NN), and decision-tree. Section IV. discusses the measurement results of the proposed classifiers with actual sentences spoken by six speakers. Finally, the summary and the conclusion of our study are presented in section V.

II. Feature Extraction

The features considered for use in making the V/U/M/N/S classification can be divided into two categories: features for V/U/M/S decision and features for nasal/non-nasal decision.

A. Feature extraction for V/U/M/S Decision

In general, time-domain analysis techniques, such as zero crossing rate, energy, and level crossing rate, are not sufficient to achieve a successful one-channel four-way, V/U/M/S classification. Different features, such as spectral distribution[9] or the LP error signal[10] has to be included in the feature set if a reliable classifier is needed. Hence, in this paper, six spectral energy ratios and normalized autocorrelation coefficients are added to the time-domain features to form the feature set. The following parameters(measurements) are computed for each block of samples:

1) Normalized log energy SENGs - defined as

$$SENG_s = 10 * \log_{10}[10^{-5} + \frac{1}{N}\sum_{n=1}^{N} s^2(n)] \qquad (1)$$

where N is a number of samples in pitch interval.

2) Normalized autocorrelation coefficient at unit sample delay, C_1

$$C_1 = \frac{\sum_{n=1}^{N} s(n)s(n-1)}{\sqrt{\left[\sum_{n=1}^{N} s(n)\right]\left[\sum_{n=1}^{N-1} s(n)\right]}} \qquad (2)$$

3) Level crossing rate of speech signal, called SLC

4) Zero crossing rate of the differentiated speech signal, called SDZCR

5) Zero crossing rate of the speech signal, called SZCR

6) Ratio1 : ratio of the spectral energy of the speech frame in the 150 - 400 Hz band to that in the 3800 - 4200 Hz

7) Ratio2 : ratio of the spectral energy of the speech frame in the 150 - 400 Hz band to that in the 4200 - 4600 Hz

8) Ratio3 : ratio of the spectral energy of the speech frame in the 150 - 400 Hz band to that in the 4600 - 5000 Hz

9) Ratio4 : ratio of the spectral energy of the speech frame in the 800 - 1200 Hz band to that in the 4200 - 4600 Hz

10) Ratio5 : ratio of the spectral energy of the speech frame in the 800 - 1200 Hz band to that in the 4600 - 5000 Hz

11) Ratio6 : ratio of the spectral energy of the speech frame in the 430 - 470 Hz band to that in the 4400 - 4800 Hz

Note that all the feature values are evaluated on a frame by frame basis with a frame size of 100 data points (10 milliseconds).

Figure 1 shows the examples of spectra for the voiced, unvoiced, mixed, and silent segments. The spectrum of voiced sounds shows that most of the energy is concentrated below 1 kHz and the first formant, usually the highest peak, is located below 350Hz. For unvoiced sounds, most of the speech energy is found above 2.5 kHz and the highest peak is also found in this region. (Even though the first formant for unvoiced sound is usually located below 450 Hz, its energy level is lower than that of the third or the fourth formant). In the case of mixed sounds, the spectrum is relatively flat for the whole frequency region, The examination of the spectra of mixed sounds indicates that there are usually two peaks. One is

located below 1kHz and the other above 3 kHz. It is believed that the former is produced by the low frequency carrier component (due to a vocal fold vibration) and the latter is caused by the noise-like high frequency component (due to a turbulent airflow), both of which exist in a mixed sound.

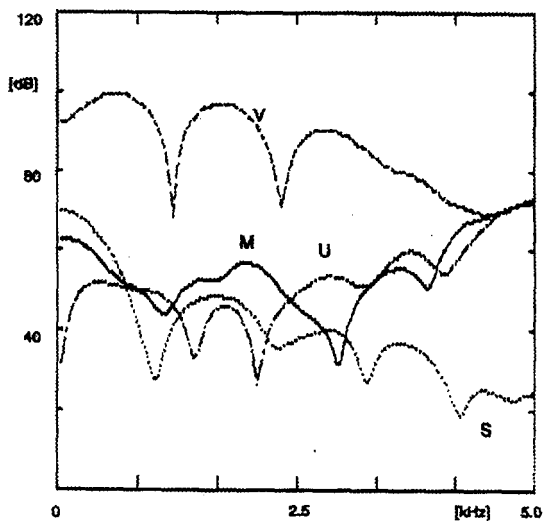


Figure 1. Spectral distribution of voiced, unvoiced,mixed, and silence.

B. Feature Extraction for Nasal/Non-nasal Decision

The search for invariant acoustic cues that indicate the presence of nasal murmurs in continuous speech has a long history. Fujimura[4] reported the spectral characteristics of nasal murmurs in intervocalic contexts. He found three essential features: first, the existence of a very low first formant in the neighborhood of 300 Hz; second, the relatively high damping factors of the formants; and third, the high density of the formants in the frequency domain. Fant[3] reported that a voiced occlusive nasal (nasal murmur) is characterized by a spectrum in which the second formant is weak or absent; a formant at approximately 250 Hz dominates the spectrum but several weaker high-frequency formants occur, and the bandwidths of nasal formants are generally larger than in vowel-like sounds.

Several recent investigations of nasals[11], and of other consonants[12] have focused upon the interactions or integration of consonantal murmur or release and vowel transition signal portions in relation to human perception, as well as algorithmic classification of phonetic contrasts. In their works on nasals, they used a method for representing spectral change at the nasal-vowel boundary. However, it is difficult to use these features in the classification of nasal/non-nasal on continuous speech directly. Our prime interest lay in classifying nasal/non-nasal in a variety of contexts such as may be encountered in free text.

The nasal sound characteristics in the spectrum[4] can be summarized as follows:

1. The existence of a very low first frequency in the neighborhood of 300 Hz.
2. The bandwidth of nasal formants are generally larger than in vowel-like sounds.
3. The second formant is weak
4. The high density of the formants is in the frequency domain.

As features representing the spectral properties, we employ four spectral energy parameters, all defined in relative values with respect to the energy in the first formant frequency band. Using the same bands tested by Seitz et al.[13], spectral value in the 450-700 Hz is subtracted from the values corresponding in the first formant frequency range, 150-400Hz. The same was done for the 1370-2150 Hz. The ratios are evaluated from the spectral distributions obtained by applying the Welch method[14]. We define the ratios as following:

ratio 7: The ratio of the spectral energy in the 150-400 Hz band to that in the 450-700 Hz band

ratio 8: The ratio of the spectral energy in the 150-400 Hz band to that in the 1370-2150 Hz band

ratio 9: The ratio of the spectral energy in the 150-400 Hz band to that in the 2700-3100 Hz band

ratio 10: The ratio of the spectral energy in the 4600-5000 Hz band to that in the 1370-2150 Hz band

III. Five-way V/U/M/N/S Classification Algorithm

No single feature seems to give consistently reliable performance in making the speech segmentation, so it is desirable to combine several features to obtain a good characterization of the segmentation of a speech signal. One way to incorporate a number of features is to view the segmentation decision process as a pattern classification problem. Atal and Rabiner[1] have used a statistical model to design a minimum distance classifier for V/U/S classification. 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.

In this study, methods for making the five-way V/U/M/N/S classification decision are examined. Three classifier types are considered: vector quantization(VQ), neural network(NN) and decision tree. For the V/U/M/N/S classifiers of VQ and NN methods, two decision structures are combined : multi-class and binary. A decision-tree with the V/U/M/S decision at the root has been implemented. The subtrees of the root are voiced(V), unvoiced(U), mixed(M), and silence(S) segments. In the V subtree, the speech segments are classified as one of two terminal nodes, or leaves : nasal (N) and non-nasal voiced(V). The resulting tree has five leaves as shown in figure 2: voiced(V), unvoiced(U), mixed(M), nasal(N), and silence(S). The advantage of this structure is that the feature set for V/U/M/S discrimination can be different from that for nasal/non-nasal discrimination.

A. Vector Quantization

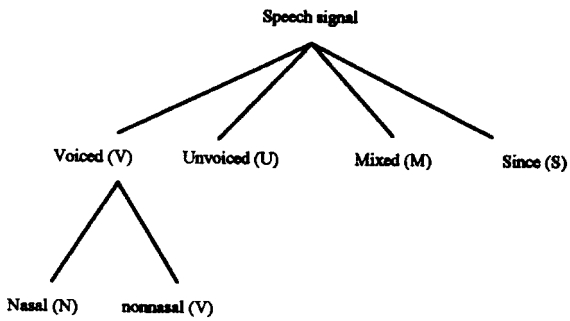
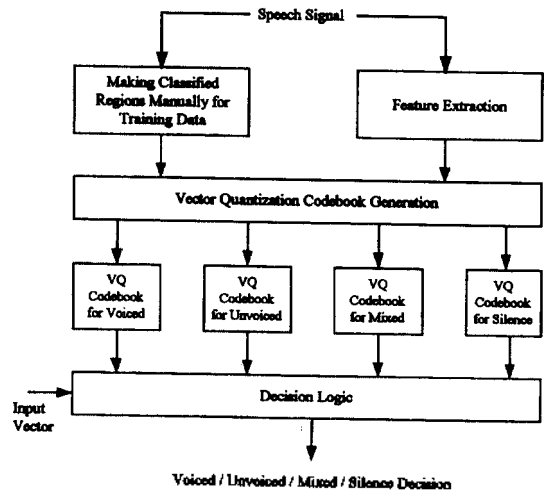


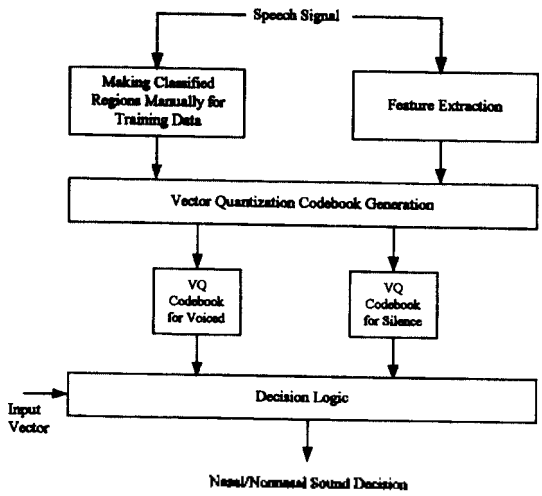
Figure 2. Decision Tree Structure in VQ(Vector Quantization) and NN(Neural Network) methods.

Vector quantization(VQ) is a process in which data to be encoded are broken into small "blocks" or vectors, which are then sequentially encoded vector by vector. The idea is to identify a set, or "codebook", of possible vectors which are representative of the information to be encoded. The VQ encoder pairs up each source vector with the closest matching vector from the code book, thus "quantizing" it. The decoding is a trivial matter of piecing together the vectors whose identity has been specified. The Pairwise Nearest Neighbor(PNN) algorithm is used, which is presented as an alternative to the Linde-Buzo-Gray(generalized Lloyd) algorithm to design a full-search VQ codebooks based on a training sequence of feature vectors is used in [15].

VQ was first used for speech coding in the 1950s and was recently revived by several researchers. The technique has been used in speech recognition and speaker recognition [16][17]. The VQ problem is part of pattern recognition problem that is concerned with classification of data into a discrete number of categories using a fidelty criterion. Typically VQ works as follows. When the input vector becomes available, the distortion between the input vector and each stored codeword is computed. The encoded output is then the binary representation of the index of the minimum distortion codeword. Since we represent the N-dimensional input vector with simply the index of the code vector, considerable data reduction is achieved. The V/U/M/S and nasal/non-nasal classification systems based on the VQ codebook approach are shown in figure 3 (a),(b).



(a)



(b)

Figure 3. (a) Block diagram of VQ(Vector Quantization) for the (a) V/U/M/S decision, (b) nasal/non-nasal decision.

In this paper, the distortion measure for the VQ procedure used is the Itakura-Saito distortion [18] measure. We conjecture that the average distortion for each subject might vary depending on the size of the codebook. The size of the codebook is to be chosen so that the maximum separation in distortion was to be obtained in V/U/M/N/S classification.

Five sentences discussed in section IV are used in this study. Each sentence was spoken by six speakers, three male and three female. Training for the VQ is performed using five sentences spoken by two male and two female. Using these training data we vary the codebook size from one to ten. We calculate the distortion for each codebook size and determined the classification errors for V/U/M/N/S classification using the training data. We found that increasing the codebook size beyond six did not reduce the number of classification errors. We therefore select our codebook size to be six. Thus, we quantize the average distortion measure to six "levels."

B. Neural Network

For the Neural Network(NN) method, the feed-forward multilayer back-propagation network was particulary well suited to the two-way V/U classification problem [19]. The principal advantages of this classification approach are its properties: 1) it focuses on correct classification of the difficult-to-classify "boundary" patterns, and 2) it makes no assumptions about the distributions of the features. The back propagation algorithm has been tested with a number of deterministic problems such as the exclusive OR problem, on problems related to speech synthesis and recognition and on problems related to visual pattern recognition. It has been found to perform well in most cases and to find good solution to the problems posed. Its principal disadvantages are computational.

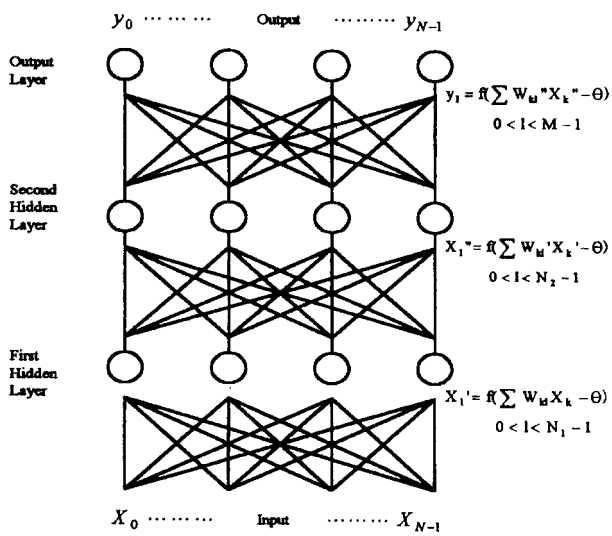


Figure 4. A three-layer perceptron with N continuous valued inputs.

We use the feedforward neural network classifier, trained with the back-propagation algorithm, for the automatic V/U/M/N/S classification. Figure 4 shows the behavior of a three-layer perceptron with N continuous valued input, M outputs and two layers of hidden units.

Eleven features for V/U/M/S and three features(SZCR:zero crossing rate, ratio7, ratio8 defined as in section II.) for nasal/non-nasal classification are used with different number of hidden units and layers and output units. Each input-output is 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 are initially set to small real pseudo-random numbers. The number of hidden units in the network will be good near the high end of what seemed reasonable for the problem; a relatively large number of hidden units will be used to safeguard against training difficulties.

The output units are meant to be binary indicators of patterns, the first output indicating V, the second output indicating U, the third output indicating M, and the third output indicating S for the V/U/M/S classification (the first output indicating nasal and the second output indicating non-nasal for nasal/non-nasal classification). Decision logic is used to select the largest value of output units.

We investigate that the requisite mapping can be carried out more directly on the input patterns. Therefore, in this paper, we decide the complexity of net using the experimental results in various numbers of layers and nodes. According to classification results depending on the network complexity, the best architectures for net are following:

for the V/U/M/S classification
- input nodes : 11
- output nodes : 4
- no. of hidden layer : 4
- no. of units for hidden layers : 11, 11, 11, and 11
- learning rate eta : 0.9
- momentum rate alpha : 0.1
- normalized system error : 0.0437
- max. no. of iteration : 2000

for the nasal/nonnasal classification

input nodes : 3
output nodes : 2
no. of hidden layer : 3
no. of units for hidden layers : 8, 11, and 5
learning rate eta : 0.9
momentum rate alpha : 0.1
normalized system error : 0.0126
max. no. of iteration : 2000

In the learning phase of training such a net, we present the pattern as input and ask that the net adjust the set of weights in all the connecting links and also all the thresholds in the nodes such that the desired outputs are obtained at the output nodes. Once this adjustment has been accomplished by the net, we present another pair of input data and desired outputs and ask that net learn that association also. In fact, we ask that the net find a single set of weights and biases that will satisfy all the (input, output) pairs presented to it. Discrepancies between actual and target output values again result in evaluation of weight changes. After complete presentation of all patterns in the training set, a new set of weights is obtained and new outputs are again evaluated in a feed forward manner. In a successful learning exercise, the system error will decrease with the number of iterations, and the procedure will converge to a stable set of weights, which will exhibit only small fluctuations in value as further learning is attempted.

Five sentences as in section IV, spoken by six speakers, three male and three female are used in this study. For the training data of V/U/M/S classification, a total of 200 frames with 50 frames for each subject are used. To obtain the optimal training condition, the same number of data set for each subject is recommended. For the training data of nasal/non-nasal classification, a total of 400 frames with 200 frames for each subject are used.

C. Decision Tree Method

The decision tree method uses a sequence of two-way decisions and has the potential advantage that the feature sets used for each discrimination can be selected independently for each decision. A second advantage is that the use of a sequence of binary classifications can allow a more flexible division of a feature space into five regions. Compared to pattern classification techniques which use one set of discriminant functions to make the V/U/M/N/S decision, this approach allows greater flexibility in the decision surfaces which define the voiced, unvoiced, mixed, nasal, and silence speech regions.

A block diagram of the analysis and decision tree algorithm is shown in figure 5.

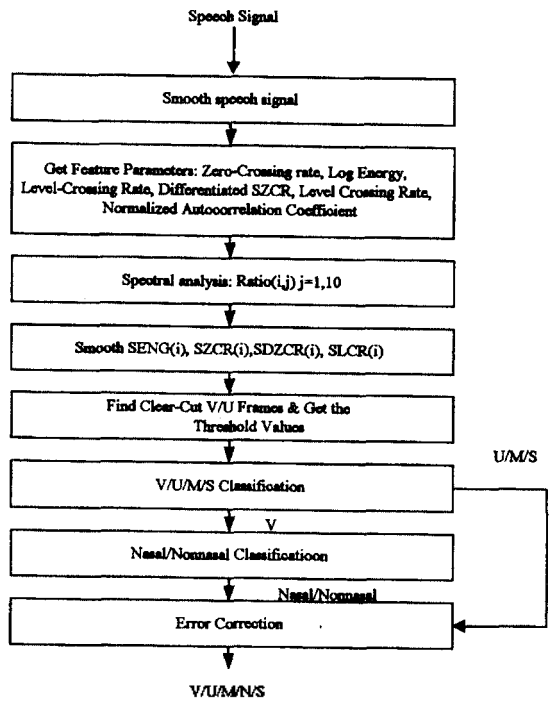


Figure 5. Block diagram of V/U/M/N/S classification using decision tree method.

The speech signal is smoothed and is formatted into blocks of 100 samples (an interval of 10ms at 10kHz sampling frequency). For each block, the features are calculated and are used for an early classification of the frames that are clear cases of voiced and unvoiced. Statistics, such as averages and standard deviations, are calculated using the features of these clear-cut frames for use directly in the tree-structure pattern classification algorithm. In that step, the remaining more difficult input speech segments are assigned to all four categories of voiced, unvoiced, mixed, or silence according to a tree-structure pattern classification technique using the features and their statistics. For the voiced frames, the nasal/nonnasal decision and the detail mixed detection are following. The last step is the error correction step, where errors such as VVVUVVV and SSSUSSS are corrected to VVVVVVV and SSSSSSS.

IV. Simulation Results

For the data base of the V/U/M/N/S classification algorithm in this study, five sentences are selected based on their phonetic contents. Each sentence was spoken by six speakers, three male and three female. These five sentences are as following:

1) We were away a year ago. (Voiced)
2) Early one morning a man and a woman ambled along a one mile lane. (Voiced and nasals.)
3) Should we chase those cowboys? (Fricatives and plosives).
4) That zany van is azure.(Voiced fricatives, i.e., mixed.)
5) We saw the ten pink fish. (Unvoiced plosives and fricatives.)

Figure 6 show the results of the V/U/M/S classification for sentence 4) spoken by a male speaker. The first plot is the spectrogram of the speech signal. The second one is the result by the manual procedure.

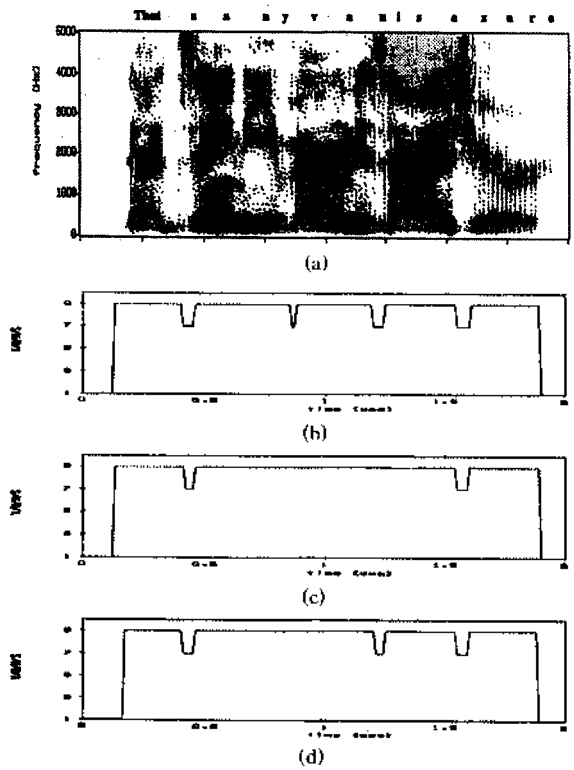


Figure 6. Comparison of V/U/M/S classification by the algorithm and manual procedures; (a) spectrogram, (b) manual classification, (c) two-channel classification algorithm, and (d) one-channel classification algorithm.

The third and fourth plots are the results of the V/U/M/S classification by the two-channel algorithm[8] and by the one-channel algorithm respectively. These contours can assume one of four values where value 1 is silence, value 3 is unvoiced, value 7 is mixed, and value 9 is voiced. It can be seen that the two-channel algorithm made essentially two error parts in classifying mixed intervals as voiced. The one-channel algorithm corrected one error part and left the rest of the V/U/M/S contour the same.

Figure 7 illustrates the V/U/M/N/S classification combined with the V/U/M/S and the nasal/non-nasal classifications for sentence 4) spoken by a male speaker.

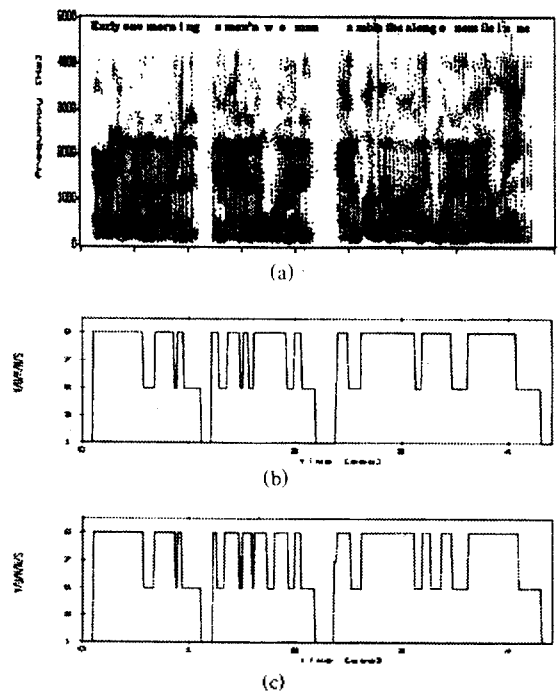


Figure 7. Comparison of V/U/M/N/S classification by the algorithm and manual procedures; (a) spectrogram, (b) manual classification, (c) one-channel five-way classification algorithm.

The first plot is the spectrogram of the speech signal. The second and third plots are the V/U/M/N/S classification by the manual procedure and by the algorithm respectively. These contours can assume one of four values where value 1 is silence, value 3 is unvoiced, value 5 is nasal, value 7 is mixed, and value 9 is voiced. This figure shows that two areas of voiced is classified as nasals and that the starting and ending of nasals make errors in classification. Although misclassification in several frames happened, the algorithm works fairly well.

In V/U/M/S classification, training for the VQ is performed using sentence 1)-5) spoken by two male and

two female speakers. The training information consist of a total of 5685 frames. Testing consist of a total 2384 frames and was performed using the same sentences spoken by to two speakers (one male and one female) not used in training. Training for the NN consist of a total of 200 frames, 50 frames for each subject. Testing consist of a total 4181 frames not used in training.

In nasal/non-nasal decision, training for the VQ is performed using sentence 1)-5) spoken by two male and two female speakers. Training consisted of a total of 4181 frames which are not classified as the silence, unvoiced, and mixed subjects in V/U/M/S classification. Testing consist of a total of 1649 frames classified as a voiced subject in V/U/M/S classification and is not used in training.

Clearly, it is desirable to use as few features as possible to perform the V/U/M/N/S classification, in order to minimize the computation time in analyzing speech. Moreover, when using statistical training methods, insufficient number of training samples can lead to what is known as the "dimensionality problem" - the phenomenon that the performance of a classifier may degrade as the number of features used is increased. As a result of the feature selection procedure, the following features were used in the classifier : (They are defined in section II.)

V/U/M/S decision : SENG, C, SZCR, SDZCR, SLCR, Ratio1, Ratio2, Ratio3, Ratio4, Ratio5, Ratio6

Nasal/nonnasal decision : SZCR, Ratio7, Ratio8, Ratio9, Ratio10

For the NN and the VQ classifiers in the nasal /non-nasal decision, we only used SZCR, ratio7 and ratio8 for the features. These features result in the improved classification on the training set.

Table 1(a) shows the V/U/M/S performance of the VQ classifier on the frames used in training. The entry in row and column of the table indicates how many classes of raw frame were classified as being in column class(e.g.. from Silence, 2 Silent frames were classified as voiced). As can be seen from the table, overall classification accuracy of 97.5% was obtained. Table 1(b) shows the V/U/M/S performance of the VQ classifier on the frames of testing not used in training. The performance has degraded somewhat (90.85% overall accuracy, 80.55% percent correct classification of mixed frames) that were available for training. Of the testing frames classified as mixed, only 80.55% (100% for training data) actually were mixed. This behavior may be partially explained by the less mixed frames in the speech used in training and in testing.

Table 1. Performance of V/U/M/S classification using the VQ classifier on the frames used in (a) training, (b) testing sentences, compared to the manual classification.

Actual class / Identified as	(a) Silence	Unvoiced	Voiced	Mixed	(b) Silence	Unvoiced	Voiced	Mixed
Silence	1029	10	37		394	7	40	1
Unvoiced	2	367	36		8	181	33	2
voiced	6	9	4052		3	16	1562	4
Mixed	1	3	68	95/96	10	18	76	29
	1029/1038 99.13 %	367/389 94.34 %	4052/4193 96.63 %	95/95 100 %	394/415 94.93 %	181/222 81.53 %	1562/1711 91.29 %	29/36 80.55%
Total	5643/5685 97.5%				2166/2384 90.85%			

Table 2(a),(b) show the performance of the nasal/non-nasal classification of the VQ classifier for training and testing. The overall performance for training is 87.2% and that for testing is 84.41%.

Table 2. Performance of nasal/ non-nasal classification using the VQ classifier on the frames in (a) training, (b) testing sentences compared to the manual classification.

Actual class / Identified as	(a) Non-nasal	Nasal	(b) Non-nasal	Nasal
Non-Nasal	2993	65	1215	25
Nasal	470	653	232	177
	2993/3463 86.42 %	653/718 91.0 %	1215/1447 83.96 %	177/202 87.62 %
Total	3646/4181 87.2%		1392/1649 84.41%	

Table 3(a) shows the performance of V/U/M/S classification of the NN classifier on the frames used in training. The number of frames for training is different from that of VQ (a total of 200 frames for the NN classifier and a total of 5685 frames for the VQ classifier). The overall performance is 97.5% which is almost the same result in table 1. Table 3(b) shows the performance for testing. An overall classification accuracy of 96.86% is obtained.

Table 4(a),(b) show the performance of nasal/non-nasal classification of the NN classifier. The number of frames used in training is 100 (4180 frames in VQ). As can be seen from the table, the overall classification accuracy is 94% for training and is 82.9% for the testing, which are slightly better than those of the VQ classifier. As on the testing frames, the overall performance is better than that

Table 3. Performance of V/U/M/S classification using the NN classifier on the frames in (a) training, (b) testing sentences, compared to the manual classification.

Actual class / Identified as	(a) Silence	(a) Unvoiced	(a) Voiced	(a) Mixed	(b) Silence	(b) Unvoiced	(b) Voiced	(b) Mixed
Silence	50			1	1375	17	25	1
Unvoiced		49		3	5	493	9	5
voiced			50		20	5	5717	9
Mixed		1		46	3	46	103	66
	50/50 100 %	49/50 98 %	50/50 100 %	46/50 92 %	1375/1403 98.00 %	493/561 87.87 %	5717/5854 97.65 %	66/81 81.48 %
Total	195/200 97.5%				7651/7899 96.86%			

for the VQ classifier. It is worth noting that, although we used less frames for training of the NN classifier, the NN classifier minimized overall misclassification.

Table 4. Performance of nasal/ non-nasal classification using the NN classifier on the frames in (a) training, (b) testing sentences, compared to the manual classification.

Actual class / Identified as	(a) Non-nasal	(a) Nasal	(b) Non-nasal	(b) Nasal
Non-Nasal	187	11	3862	80
Nasal	13	189	848	640
	187/200 93.5 %	189/200 94.5 %	3862/4710 81.99 %	640/720 88.88 %
Total	376/400	94.00 %	4502/5430	82.90%

Table 5(a),(b) show the performance of the decision tree classifier. The overall performance of V/U/M/S classification is 97.06% and that of nasal/non-nasal is 82.36%. The performance has degraded somewhat, but the classification is still fairly accurate. As mentioned before, the decision tree method allows a more flexible division of a feature space and does not need the training procedure like the VQ and the NN classifiers.

Table 5. Performance of (a) V/U/M/S classification, (b) nasal/non-nasal classification, using the statistical decision tree classifier on the frames, compared to the manual classification.

Actual class / Identified as	(b) Non-nasal	(b) Nasal
Non-Nasal	3977	95
Nasal	933	25
	3977/4910 80.99%	825/920 89.67%
Total	4802/5830	82.36%

V. Summary and Conclusion

A fairly general framework based on a pattern recognition approach to V/U/M/N/S classification has been described in which a set of measurements was made on the interval being classified, and VQ, NN, and decision tree classifiers are used to select the appropriate class. The work constitutes a demonstration that the V/U/M/N/S classification can be made with reasonable accuracy. A VQ classifier achieved 97.5% classification accuracy on training and 90.85% accuracy on testing in the V/U/M/S classification and achieved 87.02% on training and 84.41% for testing in nasal/non-nasal classification. A NN classifier achieved 97.5% accuracy on training and 96.86% on testing in the V/U/M/S classification and achieve 94% on training and 82.9% for testing in the nasal/non-nasal classification. A decision tree classifier achieved 97.06% in V/U/M/S classification and achieved 82.36% in nasal/non-nasal classification. In summary, several pattern classification approaches to making the V/U/M/N/S decision are shown to produce good results.

For the nasal/non-nasal classification, the misclassification of non-nasal to nasal occurs mostly in the areas of the nasalized vowels. So our nasal/non-nasal algorithm needs to be combined with the spectral based algorithm by Yea[20] for the identification of the nasalized vowels. This is currently under research.

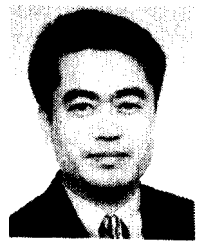
References

1. Atal, B. S. and Rabiner, L. R., "A pattern recognition approach to voiced- unvoiced- silence classification with applications to speech recognition", *IEEE Trans. on ASSP*, vol. 24, no. 3, pp. 201-212, 1976
2. Siegel, L. J., " A procedure for using pattern classification techniques to obtain a voiced/unvoiced classifier," *IEEE Trans. on ASSP*, vol. 26, no. 1, pp. 83-89
3. Fant, G., "Acoustic theory of speech production", Mouton, The Hague, 1960
4. Fujimura, O., "Analysis of nasal consonants," *Journal of Acoustical Society of America*, vol. 34, pp. 1865-1875, 1962
5. Kay, S., "Modern spectrum Estimation," Prentice-Hall Inc., Englewood Cliffs, NJ 1987
6. Klatt, D. H., "Software for a cascade/parallel formant synthesizer," *Journal of Acoustical Society of America*,

vol. 67, no. 3, pp. 971-995, 1980

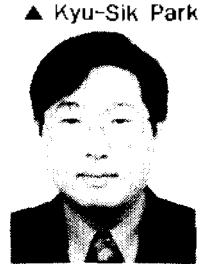
7. Krishnamurthy, A. K. and Childers, D. G., "Two channel speech analysis," *IEEE Trans. on ASSP*, vol. 34, no. 4, pp. 730-743, 1986
8. Childers, D. G., Hahn, M. and J. N. Larar, "Silent and voiced/unvoiced/mixed excitation(four-way) classification of speech," *IEEE Trans. on ASSP*, vol. 37, no. 11, pp. 1771-1774, 1989
9. Rabiner, L. R., Sambur, M. R., "Voiced-unvoiced-silence detection using the LPC distance measure," *Proc. IEEE ACASSP*, Hartford CT, pp. 323-326, 1977
10. Rabiner, L. R. and Shafer R. W., "Digital processing of speech signal," Prentice-Hall, Englewood Cliffs, NJ, 1978
11. Kurowski, K. and Blumstein, S. E., "Acosutic properties for place of articulation in nasal consonants," *Journal of Acoustical Society of America*, vol. 81, pp. 1917-1927, 1987
12. Forrest, K. and Weismer, G., "Statistical Analysis of word-initial voiceless obstruents: Preliminary data," *Journal of Acoustical Society of America*, vol. 84, pp. 115-123, 1988
13. Seitz, P. F., McCormik, M. M., and Watson, M. C., "Relational spectral features for place of articulation in nasal consonants," *Journal of Acoustical Society of America*, vol. 87, no. 1, pp.351-358, 1990
14. Welch , P. D., "The use of fast fourier transform for the estimation of power spectra: A method based on time averaging over short, modified periodograms," *IEEE Trans. Audio and Electroacoust*, vol. AU-15, pp. 70-73, 1967
15. Linde, Y. L., Buzo, A., and Gray, R. M., "An algorithm for vector quantizer design," *IEEE Trans. on Communication*, vol. 28, no. 1, pp. 84-95, 1980
16. Furui, S, "A VQ-based preprocessor using cepstral dynamic features for speaker-independent large vocabulary word recognition," *IEEE Trans. on ASSP*, vol. 36, no. 7, pp.980-987, 1988
17. Soong, F. K., and Reosenberg, A. E., "On the use of instaneous and transitional spectral information in speaker recognition," *IEEE Trans. on ASSP*, vol. 36, no. 6, pp.871-879, 1988
18. Itakura, F. and Saito, S., "Analysis-synthesis telephony based upon the maximum likelihood method," *Proc. of Int. Congress on Acoustics*, pp. 393-400, 1968
19. Bendikson, Age and Kenneth Steiglitz, "Neural networks for voiced/unvoiced speech classification," *Proc. of 1990 IEEE ICASSP*, Albuquerque, New Mexico, S10.9, pp. 521-524, 1990
20. Yea, J. J. and Childers, D. G., "Detecting speech nasalization by a spectral based algorithm," ASSP Spectrum Estimation Workshop II, Tampa, Florida, pp. 84-88, 1983

▲ Kyo-Sik Lee



Kyosik Lee received B.S. and M.S degrees in 1882, 1984 respectively from the Kyungpook University, Korea and Ph.D degree in 1992 from the Dept. of Electrical Engineering of University of Florida, Florida USA. In 1993, he served as a staff engineer in Samsung Electronics. Then he joined the Dept. of Radio Engineering in Korea Maritime University as an assistant professor in 1994. From Mar. 1995 to Aug. 1997, he served as a president in Seohan Electornics Corp. in Korea and also in Home Media Inc. in USA. From september 98, he is currently an assistant professor with the department of Information and Telecommunication in Hansei University. His research interest are digital speech processing and speech compression.

▲ Kyu-Sik Park



Kyusik Park received B.S, M.S, and Ph.D degrees in 1986, 1988, and 1994, respectively, all from the Department of Electrical Engineering of Polytechnic University, Brooklyn, NY, USA. In 1994, he joined the Semiconductor Division of Samsung Electronics as a staff engineer. He is currently an assistant professor with the Department of Information and Telecommunication in Sangmyung University, Chnungchongnam-Do, Korea. His research interests are digital signal and speech processing, and digital communication.