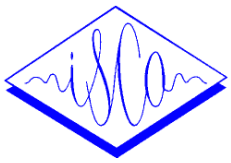

First International Conference on
Spoken Language Processing (ICSLP 90)
Kobe, Japan
November 18-22, 1990

# VOICED/UNVOICED/SILENCE CLASSIFICATION OF SPOKEN KOREAN

*Hee-Il Hahn, Minsoo Hahn*

Signal Processing Section
Elec. and Telecom. Research Inst., Korea

*ABSTRACT*

*In this paper, we presented two techniques for the automatic voiced/unvoiced/ silence classification of spoken Korean which is essential for the high quality speech synthesis and for the speech recognition system taking advantage of the acoustic-phonetic information. The database in this study is composed of five sentences spoken by 5 male and 5 female speakers. Each sentence was uttered twice by each speaker in a sound-treated room. (Almost all kinds of Korean unvoiced sounds are contained in these sentences.) One classification technique is based on the Neural Network utilizing the spectral and the time domain features such as spectral slope, energy, zero-crossing rate, and the autocorrelation coefficient at unit sample delay. The other adopts the conventional pattern classification technique, and uses almost the same features as above. Final classification accuracy of 96.2 % is achieved for both methods. Finally, the results are compared and possible future extensions are briefly discussed.*

## I. INTRODUCTION

In speech analysis, the voiced-unvoiced decision is very important, and can be used as a preprocessing for speech recognition or synthesis. There have been a variety of approaches to achieve this goal. They usually worked in conjunction with pitch analysis techniques. For example, A.M. Noll used the amplitude of the largest peak in the cepstrum for voiced-unvoiced decision [1]. Namely, he utilized the well-known quasi-periodic property of voiced sounds. But voiced sound can still become almost non-periodic when sudden changes in articulation occur. And, in that case, his algorithm usually fails at the boundaries between voiced and unvoiced sounds, because, for pitch detection, a relatively large speech segment (30-40 ms duration) is usually needed. Atal and Rabiner suggested another statistical pattern recognition approach to make three-class decision, i.e., voiced, unvoiced, and silence [2]. In their algorithm, the normal distribution of the features was basically assumed for the statistical distance measure.

In this paper, we describe two techniques for the automatic voiced/unvoiced/silence classification of spoken Korean. One is based on the Multi-Layer Perceptron (MLP), and the other uses a conventional pattern classification technique. Neural Networks have been studied in the hope of achieving human-like performance in pattern classification. While traditional statistical techniques are usually not adaptive and assume the shapes of underlying distributions, Neural net classifiers are non-parametric and require no prior knowledge about statistical informations such as mean vectors and covariance matrices. In other words, they can overcome many of the limitations imposed on most of the conventional techniques. Especially, MLP improves its performance adaptively by adjusting the connection weights in its training procedure. As a consequence, minor variations in the characteristics of processing elements could be overcome. In addition, the decision regions for any classification category can be generated in a straightforward manner by three-layer perceptron [3, 4]. Hence, we choose three-layer perceptron with two layers of hidden units.

Many studies have been reported which tried to achieve V/U/S or V/U/M/S (M for mixed sound) classification on spoken English with conventional pattern classification techniques [2, 5 - 9]. Some of them tried the classification only for the speech part, that is to say, an operator eliminated the beginning and ending silence intervals manually before processing, and the others, for the whole data. ( In the former case, the algorithm would have a critical weak point. Namely, it fails to provide the endpoint information which is inevitable for the conventional and the HMM-based speech recognition algorithms.) The reported classification accuracies usually range from 92 % to 96 %.

## II. FEATURE EXTRACTION

The speech signal is band-pass filtered (70 Hz - 4.5 kHz) and sampled at 10 kHz with the 12-bit resolution. The data are formatted into frames of 100 sample length (10 ms duration at 10 kHz sampling frequency). Since different speakers use widely varying talking levels, the signal levels, especially for female data, are scaled up such that the maximum level is about 2048. And then, five features (the first four for the MLP method and all five for the conventional one) are calculated for each frame as follows.

(1) Log energy, E

$$E = 10 * \log_{10}\left[1 + \frac{1}{N}\sum_{n=0}^{N-1} S^2(n)\right]$$

(2) Zero-crossing count, Zc
Zc increases itself by 1 when

$$\text{sign}[s(n)s(n+1)] < 0$$

(3)Normalized autocorrelation coefficient at unit sample delay, Ac

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

(4) Spectral slope, defined as the constant 'a' in a line equation,Y=aX+b, which fits the LPC spectral shape best.

(5) Level-crossing rate, Lc
Lc increases itself by 1 when

$$\text{sign}[s(n)\text{-THS}] = \text{sign}[s(n\text{-}1)\text{-THS}]$$

where THS is one-tenth of the average magnitude of the rectified clear-cut voiced sounds.

The energy of the speech signals reflects effectively the amplitude information of speech signal. In general, the energy of the unvoiced sound is usually lower than that of the voiced sound, but higher than that of silence. The zero-crossing rate is usually high for unvoiced sounds, low for silence, and intermediate for voiced sounds. Thus, it is helpful in the silence/speech or voiced/unvoiced classification. The normalized autocorrelation coefficient is an indicator of how much adjacent samples are correlated. For voiced sound or silence, the correlation coefficient is usually high, but very low for unvoiced sounds. The spectral slopes have usually large negative values for voiced, fairly small negative values for silence, but positive values for unvoiced sound. Hence it can be a strong cue to classify voiced and silence from unvoiced. The level crossing rate has a rather large value for speech but has a small value, near zero, for silence, and mainly helps in the silence/speech classification.

It is well known that it is impossible to achieve any successful V/U/S classification with one or two features only. The distribution of the feature values from different categories almost always overlap due to the speech variability and this is why multi-features are selected.

## III. EXPERIMENTS

Data Description.

The training data set was prepared by manually categorizing speech signals into silence, unvoiced, and voiced intervals on the frame basis. It consisted of three sentences uttered by two male and two female speakers (in total, 12 sentences). The number of frames for voiced, unvoiced, and silence in the training data set are 4575, 846, and 2421, respectively. The speech data in the testing set are the remaining sentences. The total number of frames in the testing set is 57466, and the percentages of voiced, unvoiced, and silence frames are about 55%, 10%, and 35%, respectively.

V/U/S Classification using MLP.

For pattern classification, the MLP trained with back propagration, is robust and can form arbitrary decision regions, and train itself rapidly. The capabilities of MLP stem from the nonlinearities used within nodes [3, 4]. The block diagram of our current MLP is shown in Figure 1. At the input layer, four features are applied to the network. The numbers of nodes in the first and the second hidden layers are twelve and eight. The output layer has three nodes, each of which corresponds to voiced, unvoiced , or silence.

The classification is done simply by selecting the class (V, U, or S) whose output node has a maximum value. The nodes in the hidden and output layers used a sigmoid function due to its convenient mathematical property. The MLP is trained with the new back-propagation training algorithm [3, 4]. The training algorithm works in the way of iterating through the training data set until convergence or a terminating condition is met. After training, the testing set of data is applied to the MLP to evaluate the performance.

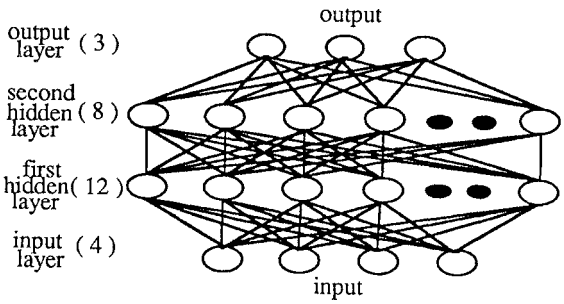


Figure 1. A three-layer perceptron with 4-input values

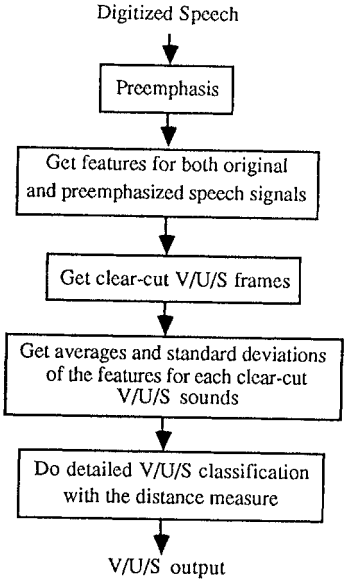


Figure 2. Block diagram of the conventional-technique -based V/U/S classification algorithm

V/U/S Classification using conventional technique.

In Figure 2, the basic idea of the V/U/S classification algorithm based on the conventional pattern recognition technique is given. When digitized speech data are applied, the algorithm first preemphasizes the data with the equation of $s'(n) = s(n)\text{-}0.95s(n\text{-}1)$. Next step is the calculation of the features for each frame.

Based on these features, clear-cut V/U/S frames are selected. (We named the remaining unclassified frames ambiguous ones.) For these clear-cut frames, averages and standard deviations for three different sounds are calculated. Finally, the ambiguous frames are categorized into voiced, unvoiced, or silence based on the distance measure. The distance measure is defined as the sum of the taxicab distance of each feature divided by its own standard deviation. It can be easily seen that this algorithm is semi-adaptive. Because it uses input-data-dependent statistical values to achieve the ambiguous frame classification rather than fixed ones. The training data set for this algorithm is the same one used for the Neural Network-based classification algorithm.

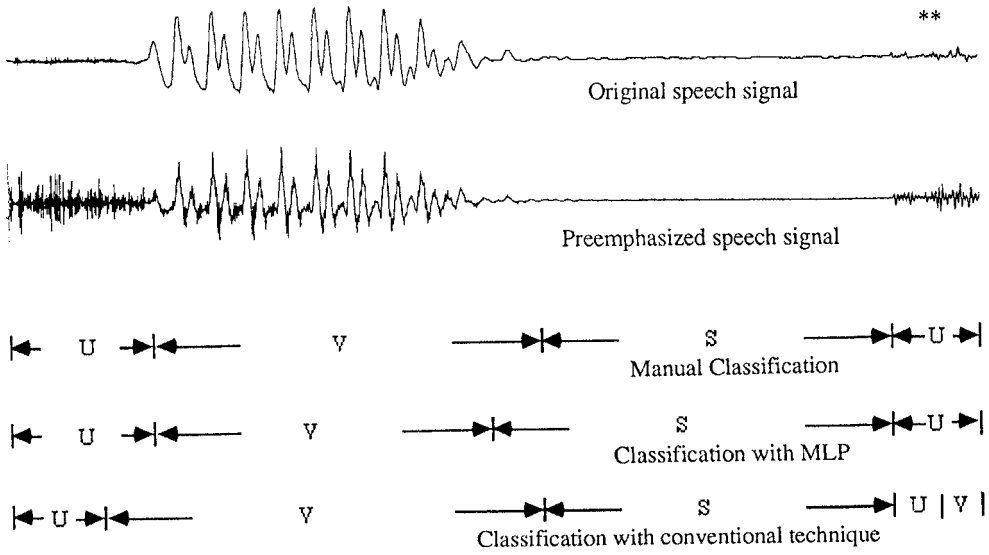


Figure 3. Comparison of V/U/S classification by the algorithms and manual procedures.

## IV. RESULTS AND DISCUSSIONS

Figure 3 shows an example of the speech waveform with the manual and the algorithmic classification results. In this figure, the symbol 'V' is used for voiced sound, 'U' for unvoiced, and 'S' for silence, respectively. The interval denoted by '**' in this figure is a part of /k/ sound. This sound is preceded by /i/ sound and followed by /eu/ sound. That is why it looks widely different from normal /k/ sound in English (It is observed that Korean unvoiced sounds are much more affected by their neighboring voiced sounds than English ones). Table 1 shows the results of the MLP classification algorithm. The recognition rate for the whole data set is 96.2 % and the difference between the recognition rate for the training data set and that for the testing data set is 0.8 %. This value is fairly small enough to confirm the robustness and the adaptability of the MLP algorithm. As expected, the male data set produces a better result than the female one. Table 2 summarizes the results of the conventional-technique-based algorithm. It may be strange that the algorithm works better for the testing data set than for the training data set. But, if it is considered that the algorithm is trained with supervision and the frequency of V-U or U-V transitions in the training data set is much higher than that in the testing dta set, no one can easily exclude the possibility of such an occurrence. And, it might be said without loss of generality that the results support the asserted semi-adaptability of the algorithm.

Error analysis shows that about 30 % of errors occurs at the boundaries between the different class of sounds. This type of errors are inherent to our algorithms, because they are using fixed-length frames. Namely, the frame at the boundary usually contains data of two different categories. Another major type of errors are the misclassification of the Korean-specific unvoiced sound into voiced one (about 23 % of whole errors). This kind of unvoiced sound has a very low energy but has meaningful low-frequency components. If the two types of errors are excluded, the final recognition rates of both algorithms will be improved up to 98 %.

Table 1. Classification Result with MLP

| DATA SET | TOTAL NUMBER OF FRAMES | NUMBER OF FRAMES IN ERROR | CORRECT RATE (%) |
|---|---|---|---|
| COMPLETE | 65308 | 2459 | 96.2 |
| TRAINING | 7842 | 246 | 96.9 |
| TESTING | 57466 | 2213 | 96.1 |
| MALE | 33215 | 1026 | 96.9 |
| FEMALE | 32093 | 1433 | 95.5 |

Table 2. Classification Result with Conventional Technique

| DATA SET | TOTAL NUMBER OF FRAMES | NUMBER OF FRAMES IN ERROR | CORRECT RATE (%) |
|---|---|---|---|
| COMPLETE | 65308 | 2475 | 96.2 |
| TRAINING | 7842 | 329 | 95.8 |
| TESTING | 57466 | 2146 | 96.3 |
| MALE | 33215 | 1208 | 96.4 |
| FEMALE | 32093 | 1267 | 96.1 |

## V. CONCLUSION

We described two kinds of algorithms for V/U/S classification on spoken Korean, and both algorithms produced fairly satisfactory results (96.2 % recognition rate for both algorithms). The algorithms can be used to provide the information to control the V/U/S switch in a speech synthesizer for Korean. And they may be adopted as preprocessors in speech recognition systems which utilize the acoustic-phonetic characteristics of the input data. We strongly believe that by generating one additional pattern for Korean-specific unvoiced sounds, the current algorithms can be improved a lot. (It may be more interesting to verify the production mechanism of these unvoiced sounds.) Another possible approach to improve the algorithm might be to build a male/female classifier and to design different algorithms according to the speaker's sex.

## REFERENCE


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

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

[3] R. P. Lippman, "An Introduction to Computing with Neural Nets", IEEE Trans. ASSP MAGAZINE, pp.4-22, April1987.

[4] Yoh-Han Pao, "Adaptive Pattern Recognition and Neural Networks", Addison-Wesley, 1989.

[5] L. R. Rabiner, R. W. Schafer, "Digital Processing of Speech Signals", Englewood Cliffs, N.J., Prentice Hall, 1978.

[6] F. Daaboul and J. P. Adoul, "Parametric segmentation of speech into V-U-S intervals", in Proc. Int. Conf. Acoust., Speech,Signal Processing, Hartford, CT, May 1977, pp. 327-331.

[7] L. R. Rabiner and M. R. Sambur, "Application of an LP distance measure to the voiced-unvoiced-silence detection problem", IEEE Trans. Acoust., Speech, Signal Processing, vol. ASSP-25, pp. 338-343, Aug. 1977.

[8] L. J. Siegel and A. C. Bessey, "Voiced/Unvoiced/Mixed excitation classification of speech", IEEE Trans. Acoust., Speech, Signal Processing, vol. ASSP-30, pp. 451-460, June 1982.

[9]D. G. Childers, M. Hahn, and J. N. Larar, "Silence and Voiced/Unvoiced/Mixed excitation (Four-way) classification of speech", IEEE Trans. Acoust.,Speech, Signal Processing, vol. ASSP-37, pp. 1771-1774, Nov. 1989.