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

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