Proceedings of the Autumn Conference in 2001
and the 30th Anniversary of the Korean Statistical Society

HMM Based Endpoint Detection for Speech Signals

Yonghyung Lee[1], Changhyuck Oh[2]

Abstract

An endpoint detection method for speech signals utilizing hidden Markov model(HMM) is proposed. It turns out that the proposed algorithm is quite satisfactory to apply isolated word speech recognition.

Key words : hidden Markov model, endpoint detection, speech signal, signal energy

1. Introduction

The problem of detecting endpoint of a speech utterance is important in speech recognition. Specially, in isolated word recognition, it is essential to detect the regions of a speech signal.

An robust endpoint detection algorithm proposed by Acero et al.[2] uses an HMM for background noises and an HMM for speech signals. The feature vectors consist of noise-normalized log energy and delta log energy. There are three states in HMM for background noise and four states in HMM for other signals. Each state is built with on Gaussian density. Kosmides et al.[3] improves the algorithm of Acero et al. by using multiple HMM. In a speech section, the signal features are described by a set of 4 HMMs. Instead of using the energy-based features, a spectral feature is developed to increase accurate. Seok and Bae[3] proposes a new feature based on a discrete wavelet transformation for endpoint detection and gives us a good result. Here a new stochastic endpoint detection algorithm is suggested and is shown to be a quite competitive alternative to existing ones.

2. Stochastic endpoint detection algorithm

Speech energy is used to derive three interim features : "shock", "slope", and "sigma". In silence status i.e. in noisy status, mean and standard deviation of 20 frames each 10ms time width are calculated. Shock has the values 1, -1, or 0 according to the difference of previous frame's energy and current one's energy. Value of slope is 1 if energy values of three consecutive frames are monotone increasing or monotone decreasing and is -1 otherwise. Sigma is 2, 1, or 0 if frame energy is greater than the mean of noisy frames plus 10, 5, 0 times of the standard deviation of the noisy frames, respectively. With this interim features, observation values for endpoint detection using these features are defined.

---

1) Graduate, Dept. of Statistics, Yeuangnam University, Kyungsan, 712-749
2) Professor, Dept. of Statistics, Yeungnam University

Observation values are noise status, speech status, speech-end-type-1 status, and speech-end-type-2 status. A discrete HMM with the three states, noisy, speech, and speech-end is modeled with the observations and the states. Initial transition probabilities and observation distribution is obtained from the empirical frequencies in the recorded words. Stepwise Viterbi search algorithm is used to find maximum likelihood estimate with observations.

Experiments are carried out to evaluate the performance of the suggested algorithm. For that purpose, 40 words from the word list in [1] are selected and are recorded to get 400 word utterances by 10 speakers (5 males and 5 females) in office environment. Time distance from endpoint positioned by hand and the detected endpoint by the suggested method is calculated. In table 1, ratio of the distances are given with Seok and Bae [4]'s results for comparison.

Table 1. Ratio of distances from the manual endpoint and automatic detected endpoint

Time distance (msec)	Proposed method	Wavelet method[3]	EZ method[3]
7.5	87.2		
15.0	93.9		
22.5	96.9		
25.0		70.8	13.1
30.0	98.8		
37.5	99.9	86.4	39.5
50.0	100.0	94.1	60.8
62.5		95.4	72.3
75.0		100.0	98.7

In Table 1, it is seen that accuracy of the given algorithm is quite high compare to wavelet method and EZ method of Seok and Bae [4].

References

[1] 은종관 (1990). 대어휘 연속음성 인식을 위한 음소인식 기술 개발, 최종연구보고서, 과학기술처.

[2] Acero, A. , Crespo, C., Torre, C. de la, and Torrecilla, J. C. (1993). *Robust HMM-based endpoint detector,* Eurospeech, pp. 1551-1554.

[3] Kosmides, E., Dermatas, E. and Kokkinakis, G. (1997). *Stochastic endpoint detection in noisy speech*, SPECOM97 workshop Clui-Napoca, Romania, pp. 109-114.

[4] Seok, J. W. and Bae, K. S. (1999). *Endpoint detection of speech signal using Wavelet transform*, 한국음향학회지, 제 18권, 제 6호, pp. 57-63.