음성인식에서 특이 특징벡터의 제거에 대한 연구

이창영*

A Study on the Removal of Unusual Feature Vectors in Speech Recognition

Chang-Young Lee*

요 약

음성 인식을 위해 추출되는 특징벡터 중 일부는 드물게 나타나는 특이 패턴이다. 이들은 음성인식 시스템의 훈련에서 파라미터의 과도맞춤을 일으키며, 그 결과 새로운 입력 패턴의 인식을 저해하는 구조적 위험을 초래한다. 본 논문에서는 이러한 특이 패턴을 제거하는 하나의 방법으로서, 어느 크기 이상의 벡터를 제외시켜 음성인식 시스템의 훈련을 수행하는 방법에 대해 연구한다. 본 연구의 목적은 인식률을 저해시키지 않는 한도에서 가장 많은 특이 특징벡터를 제외시키는 것이다. 이를 위하여 우리는 하나의 절단 파라미터를 도입하고, 그 값의 변화가 FVQ(Fuzzy Vector Quantization)/HMM(Hidden Markov Model)을 사용한 화자독립 음성인식에 미치는 영향을 조사하였다. 실험 결과, 인식률을 저하시키지 않는 특이 특징벡터의 수가 3%~6% 정도임을 확인하였다.

ABSTRACT

Some of the feature vectors for speech recognition are rare and unusual. These patterns lead to overfitting for the parameters of the speech recognition system and, as a result, cause structural risks in the system that hinder the good performance in recognition. In this paper, as a method of removing these unusual patterns, we try to exclude vectors whose norms are larger than a specified cutoff value and then train the speech recognition system. The objective of this study is to exclude as many unusual feature vectors under the condition of no significant degradation in the speech recognition error rate. For this purpose, we introduce a cutoff parameter and investigate the resultant effect on the speaker-independent speech recognition of isolated words by using FVQ(Fuzzy Vector Quantization)/HMM(Hidden Markov Model). Experimental results showed that roughly 3%~6% of the feature vectors might be considered as unusual, and therefore be excluded without deteriorating the speech recognition accuracy.

키워드

Speech Recognition, Unusual Feature Vector, Structural Risk Minimization, MFCC, Hidden Markov Model
음성인식, 특이 특징벡터, 구조적 위험의 최소화, MFCC, 은닉 마코브 모델

I. Introduction

As a method of communication between man and machine, speech recognition provides a very effective interface. Speech input to a machine is about twice as fast as information entry by a skilled typist [1]. The need for and usefulness of speech-to-text transcription cannot be overestimated.

* 동서대학교 산업경영공학과(seewhy@dongseo.ac.kr)
접수일자 : 2013. 02. 01 심사(수정)일자 : 2013. 03. 25 게재확정일자 : 2013. 04. 25

The state of the art technology in the field of speech recognition has reached such a mature level of performance that permits lots of daily applications. As a result, we are now living in a world of various devices which deploy the relevant achievements [2-4].

Pattern classification proceeds largely in two stages, one for feature vector extraction from input signal and the other for pattern classification (recognition) of the feature vectors through a scoring procedure. As for the feature vectors in the field of speech recognition, mel-frequency cepstral coefficients (MFCC) were proven to be very effective [5].

Since the usual speech production shares common features over people, most of the feature vectors would agglomerate in the feature hyperspace and be categorized as being normal. Speech production in the benign circumstances (carefully articulated and spoken in a relatively noise-free environment) corresponds to this class. However, the situation is not always this case. For example, during articulation, the talker often produces sound artifacts, including lip smacks, heavy breathing, and mouth clicks and pops [6]. Some people speak in heavy dialects. These sort of speech tokens might be inferred not to produce common feature vectors. Instead, they lead to unusual patterns and hence appear as rare and unusual points in the feature vector space. These vectors in turn cause overfitting of the system parameters. Therefore, cleaning (or excluding) of these rare, unusual, and spurious patterns is advisable in the pattern classification. If appropriately removed, reduction in the computational cost might be obtained as a byproduct.

In a general sense, structural risk minimization (SRM) is an invaluable scheme in any kind of machine learning. The SRM principle was first set out in 1974 by Vapnik and Chervonenkis [7]. Commonly, in machine learning, a generalized model must be selected from a finite data set, with the consequent problem of overfitting - the model becoming too strongly tailored to the particularities and possible random noises of the training set and thereby generalizing poorly to new data. The SRM principle addresses this problem by balancing the model's complexity against its success at fitting the training data.

In this paper, we consider a method of reducing the structural risk of overfitting in the speech recognition system by removing the rare and unusual feature vectors. This prescription would provide robustness against overfitting and reduction in the computational cost.

The organization of this paper is as follows. Section II describes experimental details in our study. After expounding various results on the efficacy of the proposed method in Section III, concluding remarks are given in section IV.

II. Experiment

Our experiments were performed on a set of phone-balanced 300 Korean words. To see the effect of vocabulary size, we divided the words into three sets as in Table 1. The sets A and B are disjoint each other and C is the union of them.

Table 1. Three sets of speech data divided for studying the effect of vocabulary size

Word Set	Number of Words
A	100
B	200
C	300

Forty people including 20 male and 20 female speakers participated in speech production. Speech utterances of them were divided into three disjoint groups as in Table 2.

Table 2. Division of the 40 people's speech production into three groups

Speaker Group	Number of People
I	28
II	6
III	6

Twenty-eight people's speech tokens of the group I were used in generating codebook of size 512, whose centroids serve for fuzzy vector quantization (FVQ) of all the speeches of 40 people. HMM parameters were updated on each iteration of training. In order to choose which values of parameters to use in the final test of speech recognition, some test speeches are necessary. The parameters that yield the best performance on the group II were stored and used for the test on the group III to obtain the final performance of the speaker-independent speech recognition system. This prescription prevents the system from falling too deep into the local minimum driven by the training samples of the group I and hence becoming less robust against the speaker-independence when applied to the group III [8].

The speech utterances were sampled at 16 kHz and quantized by 16 bits. 512 data points corresponding to 32 ms of time duration were taken to be a speech frame for short-term analysis. The next frame was obtained by shifting 256 data points, thereby overlapping the adjacent frames by 50% in order not to lose any information contents of coarticulation [9].

We discriminate MFCC feature vectors $\mathbf{v}$ according as the Euclidean norm $R = \| \mathbf{v} \|$. For this purpose, we introduce a cutoff parameter α that will serve as a criterion of separation between "normal" patterns and "spurious" (unusual) ones. We will consider a vector as spurious if its Euclidean norm is larger than $\alpha R_{\max}$, where $R_{\max}$ is the largest norm among all the feature vectors under study. The supposedly spurious vectors lead to overfitting and structural risk in the recognition system and therefore will be discarded in the processing. In our experiment, the cutoff parameter α was varied from 0.4 to 1 in steps of 0.02. The value of $\alpha = 0.7$, for example, means that the vectors of norms larger than $0.7R_{\max}$ are treated as spurious and hence discarded. It α is taken to be small, then larger fraction of the feature vectors are treated as spurious and hence only smaller fraction of feature vectors would participate in subsequent processing.

To each frame, Hanning window was applied after pre-emphasis for spectral flattening. MFCC feature vectors of order 13 were obtained and then cepstral mean subtraction (CMS) [10] were applied on utterance basis to endow robustness against various adverse effects such as system depe-ndence and noisy environment.

Codebooks of 512 clusters were generated by the Linde-Buzo-Gray clustering algorithm on the MFCC feature vectors obtained from the speeches of the group I of Table 2, with spurious vectors excluded according to the criterion explained above. As for the field of speech recognition, neural network approach might be employed [11-12] but we used hidden Markov model (HMM) in this paper. The distances between the vectors and the codebook centroids were calculated and sorted. Appropriately normalized fuzzy membership values were assigned to the nearest two clusters and a train of two doublets (cluster index / fuzzy membership) fed into HMM for speech recognition processing.

For the HMM, a non-ergodic left-right (or Bakis) model was adopted. The number of states that is set separately for each class (word) was made proportional to the average number of frames of the training samples in that class [13]. Initial estimation of HMM parameters $\lambda = (\pi, A, B)$ was obtained by K-means segmental clustering after the

first training. By this procedure, convergence of the parameters became so fast that enough convergence was reached mostly in several epochs of training iterations.

Backward state transitions were prohibited by suppressing the state transition probabilities a_{ij} with $i > j$ to a very small value but skipping of states was allowed. The last frame was restricted to end up with the final state associated with the word being scored within a tolerance of 3. Parameter reestimation was performed by Baum-Welch reestimation formula with scaled multiple observation sequences to avoid machine-errors caused by repetitive multiplication of small numbers. After each iteration, the event observation probabilities $b_i(j)$ were boosted above a small value.

Three features were monitored while training the HMM parameters: (1) the recognition error rate for the group II of Table 2, (2) the total probability likelihood of events summed over all the words of the training set according to the trained model, and (3) the event observation probabilities for the first state of the first word in the vocabulary list. Training was terminated when the convergences for these three features were thought to be enough. The parameter values of $\lambda = (\pi, A, B)$ that give the best result for the group II were stored and used in speech recognition test on the group III of Table 2.

We investigate the recognition error rate versus the cutoff parameter α. If α is above a certain value, then it is expected that only spurious feature vectors are excluded in the process and thus the recognition error rate does not change significantly. On the other hand, if α is decreased below a certain value towards zero, then normal (usual) and useful feature vectors begin to be excluded along with the spurious vectors, and as a result, it adversely affects and deteriorate the recognition error rate. One objective of this paper is to determine the threshold value of α below which the recognition error rate begins to decrease.

III. Results and Discussion

Figure 1 shows the distribution of MFCC feature vectors in 2-dimensional subspace spanned by the first and the second components of the MFCC feature vectors. Most of the vectors agglomerate together but some vectors reside on the outskirts of the cluster. We see that some of these supposedly spurious vectors might be better excluded in the speech recognition processing.

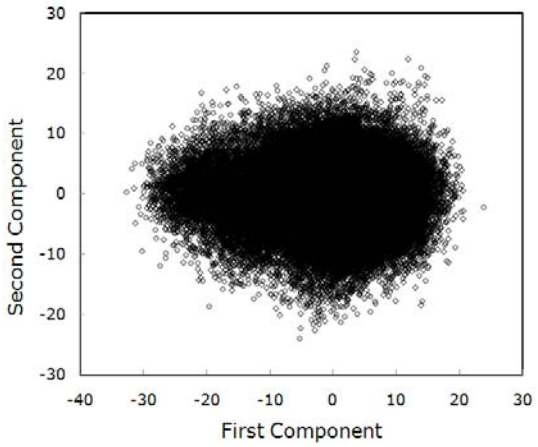

Fig. 1 Distribution of MFCC feature vectors plotted in 2-dimensional subspace spanned by the first and the second components.

Figure 2 shows the distribution of the feature vectors according as their relative norms $R/R_{\max}$. We see that most of the vectors have norms of $R < 0.5R_{\max}$.

We now exclude the feature vectors whose norms are larger than $\alpha R_{\max}$. The values of α were varied from 0.4 to 1 in steps of 0.02 and the resultant recognition error rates of speech recognition were examined.

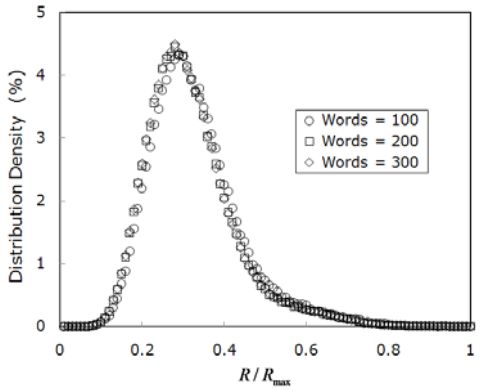

Fig. 2 Distribution of the feature vectors according as their relative norms.

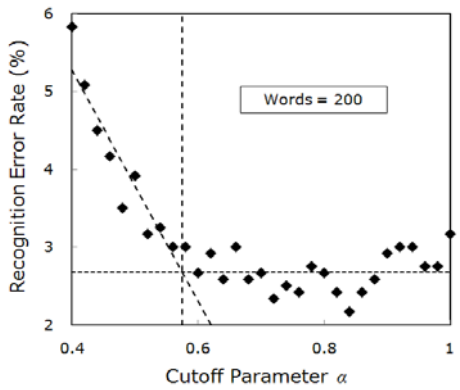

Fig. 3 Recognition error rate versus the cutoff parameter α for 200 words.

Figure 3 shows the speech recognition result as the cutoff parameter α is varied. The general behavior might be phrased in terms of two stages, one for the approximately linear decrease in the recognition error rate and the other for little change (with minor fluctuations). For the values of α close to zero, only a small fraction of useful information is included in the process and the recognition error rate becomes small inevitably. As α is increased, more useful information is included in the process and, as a result, the performance becomes better. Above a certain threshold value, however, the decrease in the recognition error rate is improved only insignificantly, meaning that the additional vectors by increase of α are actually spurious. This feature were found to pervade all the cases under our study.

By two separate curve-fittings on the two characteristic regions, the optimal cutoff parameter was located as the abscissa coordinate of the intersection of the two fitted lines. Table 3 shows the summary of the results. We see that the optimal value of α is around $\alpha^* = 0.5 \sim 0.6$. The rightmost column is the ratio of the number of feature vectors $N(\alpha)/N(1)$ for optimal values of α, i.e., α^*. $N(1)$ means the number of all the feature vectors with no vectors excluded.

Table 3. The optimal cutoff parameter α^* and the ratio of the number of feature vectors $N(\alpha^*)/N(1)$ included in the speech recognition.

# of Words	Optimal Cutoff Parameter α^*	$N(\alpha^*)/N(1)$
100	0.52	≈0.94
200	0.58	≈0.97
300	0.56	≈0.96

Figure 4 shows the ratio of the number of feature vectors $N(\alpha)/N(1)$ for the set C of Table 1 (words 300). This result is almost the same in the cases of the set B and C of Table 1. As α is

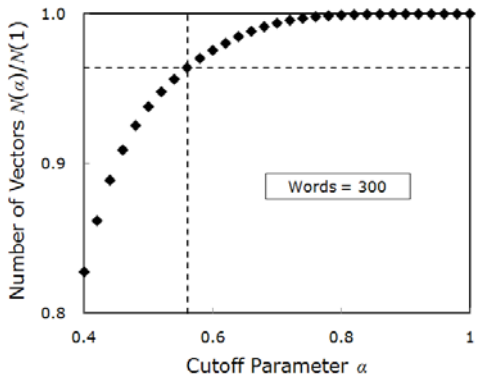

Fig. 4 The ratio of the number of feature vectors $N(\alpha)/N(1)$ for the set C of Table 1 (words 300)

increased, less vectors are excluded and more vectors are included. The vertical dotted line denotes the location of the optimal value of α above which no significant change occurs in the speech recognition performance.

For 300 words, as an example, the optimum value of α was estimated to be 0.56, above which the recognition error rate does not show significant change since additionally included feature vectors are spurious. For this optimum value of α, the ratio of the number of feature vectors is $N(0.56)/N(1) = 0.96$. This means that we might decrease 4% of the number of feature vectors without deteriorating the speech recognition performance.

IV. Conclusion

In this paper, an experimental method of removing unusual feature vectors was studied by introduction of a cutoff parameter. The aim is to exclude the unusual spurious feature vectors that would not give rise to signifiant adverse effect on the speech recognition performance. We introduced a cutoff parameter α and discarded vectors whose norms are larger than $\alpha R_{\max}$, where $R_{\max}$ is the largest norm.

The effect of excluding the unusual feature vectors might be stated in two respects. One is the structural risk minimization by reducing the possibility of overfitting onto the training feature vectors. The other is reduction in the calculational cost.

Speech recognition performance showed largely two stages of changes as the value of the cutoff parameter is increased from 0.4 to 1 in steps of 0.02: one is the roughly linear decrease in recognition error rate and the other is minor fluctuation for α above a certain value. Optimal value of the cutoff parameter was located by the point of intersection for curve-fittings in the two regions.

The optimal values of cutoff parameter were estimated to be around 0.5~0.6 and the corresponding number of feature vectors for processing were found to be reduced by around 3%~6%. The dependence of the results on the vocabulary size was minor.

References

[1] G. Kaplan, "Words Into Action I," IEEE Spectrum, Vol. 17, pp. 22-26, 1980.

[2] Y. Chang, S. Hung, N. Wang, & B. Lin, "CSR: A Cloud-Assisted Speech Recognition Service for Personal Mobile Device," International Conference on Parallel Processing (ICPP), pp. 305-314. 2011.

[3] 김범준, "와이브로 네트워크를 통한 음성 서비스의 측정 기반 품질 기준 수립," 한국전자통신학회논문지, 6권, 6호, pp. 823-829, 2011.

[4] 김영표, 이한영, "음성 인식률 개선 방법에 관한 연구," 한국전자통신학회논문지, 8권, 1호, pp. 77-83, 2013.

[5] J.-C. Wang, J.-F. Wang, & Y. Weng, "Chip design of MFCC extraction for speech recognition," The VLSI Journal, Vol. 32, pp. 111-131, 2002.

[6] L. Rabiner & B. Juang, "Fundamentals of Speech Recognition," Prentice Hall, pp. 143-149, 1993.

[7] V. Vapnik, "Principles of Risk Minimization for Learning Theory," Advances in Neural Information Processing Systems, Vol. 4, pp. 831-838., 1992.

[8] L. Fausett, "Fundamentals of Neural Networks," Prentice-Hall, p. 298, 1994.

[9] J. R. Deller, J. G. Proakis, & J. H. L. Hansen, "Discrete-Time Processing of Speech Signals," Macmillan, New York, pp. 143-145, 1994.

[10] W. Xu, et. al., "A Noise Robust Front-End Using Wiener Filter, Probability Model and CMS for ASR," International Conference on Natural Language Processing and Knowledge Engineering, pp. 102-105, 2005.

[11] M. D. Emmerson, & R. I. Damper, "Relations between fault tolerance and internal representations for multi-layer perceptrons," IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP), Vol. 2, pp. 281-284. 1992.
[12] 최재승, “신경회로망에 의한 음성 및 잡음 인식 시스템,” 한국전자통신학회논문지, 5권, 4호, pp. 357-362, 2010.
[13] M. Dehghan, K. Faez, M. Ahmadi, & M. Shridhar, "Unconstrained Farsi Handwritten Word Recognition Using Fuzzy Vector Quantization and Hidden Markov models," Pattern Recognition Letters, Vol. 22, pp. 209-214. 2001.

저자 소개

이창영(Chang-Young Lee)

1982년 2월 서울대학교 물리교육학과 졸업(이학사)
1984년 2월 한국과학기술원 물리학과 졸업(이학석사)
1992년 8월 뉴욕주립대학교 (버펄로) 물리학과 졸업(이학박사)
1993년~현재 동서대학교 시스템경영공학과 교수
※ 관심분야 : 음성인식, 화자인식, 신호처리