신경망 기반 화자증명 시스템에서 더욱 향상된 사용자 등록속도

이태승* · 최성원* · 황병원*
*한국한공대학교

Faster User Enrollment for Neural Speaker Verification Systems

Tae-Seung Lee* · Sung-Won Choi* · Byong-Won Hwang*
*Hankuk Aviation University
E-mail : thestaff@hitel.net

요 약

MLP(multilayer perceptron)는 화자증명에 대한 응용에 있어 우수한 특질을 지니고 있지만 동시에 느린 학습속도의 문제를 안고 있다. 편리한 사용을 위해 MLP에 기반한 화자증명 시스템에서는 신속한 화자등록이 요구되며 이 문제는 MLP의 빠른 학습속도에 전적으로 의존한다. 이러한 시스템에서 실시간 등록을 달성하기 위해 지금까지 두 가지 측면에서 연구가 시도되었으며 각기 의도한 목적을 달성하였다. 본 논문에서는 이 두 방법이 상이한 최적화 원리에서 동작한다는 가정 하에 이들을 결합하고 이를 MLP 기반 화자증명 시스템에 적용한다. 이러한 결합이 화자등록 속도를 더욱 향상시킬 수 있다는 사실은 한국어 음성 데이터베이스를 이용한 실험결과에서 입증된다.

ABSTRACT

While multilayer perceptrons (MLPs) have great possibility on the application to speaker verification, they suffer from inferior learning speed. To appeal to users, the speaker verification systems based on MLPs must achieve a reasonable enrolling speed and it is thoroughly dependent on the fast learning of MLPs. To attain real-time enrollment on the systems, the previous two studies have been devoted to the problem and each satisfied the objective. In this paper, the two studies are combined and applied to the systems, on the assumption that each method operates on different optimization principle. By conducting experiments using an MLP-based speaker verification system to which the combination is applied on real speech database, the feasibility of the combination is verified from the results of the experiments.

키워드
화자증명, 생체인식, 다층신경망, 등록속도 향상

I . Introduction

Speaker verification systems require real-time speaker enrollment as well as real-time verification to provide satisfactory performance. To use speaker verification system to be used in daily life, it is necessary to consider a very fast verification since the system must be used frequently. In addition to it, the user convenience criterion for speaker verification system requests fast enrollments of speakers. Most users want to use verification services just after enrolling themselves for the system. If they have to wait for a long time for the first usage, they may quit their enrolling process.

Unlike parametric-based speaker verification systems, the systems based on multilayer perceptrons (MLPs) more quickly conduct the computation needed to verify identities but slowly to enroll speakers [1],[2]. The structure of MLPs inspires a fast verification process even with low-computational capability. On the other hand, it is difficult to settle the optimal values of

the internally weighted connections to achieve best behaviors of MLPs. And the number of background speakers required for an MLP to learn an enrolling speaker slow down the enrolling speed further.

For the difficulty in settling the optimal values of the internally weighted connections, the previous work has attempted to reduce the number of learning steps and shorten the duration of each learning step in the error backpropagation (EBP) algorithm [3]. The EBP algorithm is widely used to learn MLPs but has somewhat poor learning speed due to its local information dependency. Nevertheless, the EBP has advantages to drive an excellent anti-overfitting ability and reveals a fairly fast learning when it is operated in online mode on pattern recognition application [4],[5]. To accommodate the fast property on pattern recognition, Lee et al. have proposed an improved EBP to exploit the redundancy of pattern recognition data and achieved a substantial improvement in learning speed without any lose of recognition performance.

For the awful number of background speakers that hinders the real-time speaker enrolling on MLP-based speaker verification systems, the method to reduce the number of background speakers required to enroll speakers has been successfully applied to the systems [6],[7]. MLPs learn the enrolling speaker by the differences to any other speakers, so background speakers should be provided sufficiently to represent the whole world speakers. However, the increasing number of background speakers means the increasing of learning data and it is not acceptable for MLP-based speaker verification systems that must enroll speakers in real-time. To relieve the burden, Lee et al. have introduced a data reduction method to select the only background speakers related to the enrolling speaker by using the discriminant learning property of MLPs and obtained a rather effective result in learning speed.

This paper combines the two studies to get further improvement in enrolling speed on MLP-based speaker verification. The reduction method to select background speakers on qualitative criterion cuts off the useless learning data before the actual learning, so it is considered a global optimization of learning data [7]. Then the learning begins and the useless learning data out of the complete learning data set making up one learning step in MLPs are omitted from each learning step, so it is considered a local optimization [3]. When the two methods are combined, the optimality in learning data amount can be maximized and the real-time performance of MLP-based speaker verification systems might be easy to reach.

II. Discriminative Cohort Speakers Method

The prospect to reduce background speakers in MLP-based speaker verification arises from the contiguity between learning models. That is, in MLP learning, a model's learning is cooperated only with its geometrically contiguous models. When an enrolling speaker is given into background speaker crowd for its learning, an MLP's decision boundary to learn the difference between the enrolling speaker and the background speakers is affected only by the background speakers adjacent to the enrolling speaker. If a great number of background speakers are reserved to obtain very low verification error, the percentage of such background speakers does increase and the number of background speakers needed to learn decision boundary can be shortened.

The process to select the background speakers similar to an enrolling speaker in MLP-based speaker verification is implemented like this:

$$S_{Cohort} = Sel_{M_{MLP} \geq \theta, I}(Sort_{Dec}(M_{MLP}(S_{BG} \mid \mathbf{X}))),$$
$$S_{BG} = \{S_i \mid 1 \leq i \leq I\} \quad (1)$$

where, $\mathbf{X}$ is the speech of enrolling speaker, S_{BG} the background speakers set of which population is I, M_{MLP} the MLP function which evaluates likelihoods for each background speaker to given $\mathbf{X}$. $Sort_{Dec}$ represents the function which sorts given value set in descending manner, $Sel_{M_{MLP} \geq \theta, I}$ the function to select the background speakers of the maximum number of background speakers I whose M_{MLP} s exceed the preset threshold θ. In this paper, the method is called the discriminative cohort speakers (DCS).

In this paper, MLPs to calculate M_{MLP} are called MLP-I and MLPs to learn an enrolling speaker using the background speakers selected by MLP-I called MLP-II. While MLP-Is are learned before enrollments using background speakers' data, MLP-IIs are learned at the time

of enrolling speakers. It should be noted that although an MLP-II has one output node since it discriminates the current pattern input just into the enrolled speaker and the background speaker group, an MLP-I has I output nodes since it has to evaluate the likelihoods of all background speakers.

III. Omitting Patterns in Instant Learning Method

MLPs learn the representation of models by establishing decision boundaries which discriminate the model areas. If the patterns of the whole models are fully presented in iterative manner and the internal parameters of an MLP are adjusted so that all the patterns of each model are classified into its own model, the decision boundaries would be finally settled in the optimal positions.

The common MLP learning method, online mode EBP algorithm, updates the weights of an MLP using the information related to the given pattern and the current weights status like this:

$$w_{ij}(n+1) = w_{ij}(n) + \Delta w_{ij}(n) = w_{ij}(n) - \eta \frac{\partial e_p(n)}{\partial w_{ij}(n)} \quad (2)$$

$$e_p(n) = \frac{1}{2}\sum_{k=1}^{M} e_k^2(n) \quad (3)$$

$$e_k(n) = d_k(n) - y_k(n) \quad (4)$$

where, w_{ij} is the weight to link with a weighted value from the computational node j to the node i, n the weights update count, and e_p the summed error from all the output nodes for the given pattern p. e_k, d_k and y_k are the error, the learning objective and the network value of the output node k, respectively. M designates the number of output nodes and η the learning rate to determine how much portion of the weight vector change Δw_{ij} is applied to the update. The learning objective is, in general, designated 1 if the output node corresponds to the class of the current pattern, otherwise 0 or -1 corresponding to whether the activation function is binary type or bipolar type, respectively. The weight updates continue until some criterions are satisfied, for example, the summation of e_ps for all the learning patterns goes down below a certain value. After a learning is complete, the network outputs each converging to its own objective are derived from the learned weights and the decision boundaries are formed at the valleys between the high output values on each model area.

The usefulness of the given pattern in the current epoch can be determined on the criterion of the error energy objective. One epoch is the duration that all learning patterns are once presented and the evaluation of when the learning stops is carried out on the end of each epoch. In the online mode EBP, the achievement of learning in the current epoch is measured with the error energy averaged for the entire N patterns like this:

$$e_{avg}(t) = \frac{1}{N}\sum_{p=1}^{N} e_p(t) = \frac{1}{2N}\sum_{p=1}^{N}\sum_{k=1}^{M} e_k^2(t) \quad (5)$$

where, t is the epoch count. The learning continues until the average error energy $e_{avg}(t)$ is less than the learning objective e_{obj}:

$$\begin{cases} w_{ij}(n+1) = w_{ij}(n) + \Delta w_{ij}(n), & \text{if } e_{avg}(t) > e_{obj} \\ Stop, & \text{otherwise} \end{cases} \quad (6)$$

The relationship between the average error energy and the individual error energies is as follows:

$$e_{avg}(t) \le e_{obj}, \quad \text{if } e_C^2(n) \le 2\lambda e_{obj} \text{ for all } N \text{ patterns}, \quad 0 < \lambda \le 1, \quad (7)$$

where, $e_C^2(n)$ is the error energy of the output node C associated with the given pattern. This expression means that if all the $e_C^2(n)$ for the entire learning patterns are less than or equal to $2e_{obj}$, then the learning is complete, assuming that the learning is progressed sufficiently to ignore the other output values beside C. As a result, it is possible to learn only the patterns

with $e_C^2(n) > 2e_{obj}$ for completing learning. However, in actual situation the errors of the other outputs may not be ignored, so the coefficient λ is added to compensate for such errors included in $e_{avg}(t)$. In this paper, the new EBP algorithm employing (7) is called the omitting patterns in instant learning (OIL) method. MLP-IIs described in Section 2 are learned by the OIL method.

IV. Implemented System

The speaker verification system extracts isolated words from input utterances, classifies the isolated words into nine Korean continuants (/a/, /e/, /ʌ/, /o/, /u/, /ɨ/, /i/, /l/, nasals) stream, learns an enrolling speaker using MLP-I and MLP-II for each continuant, and calculates identity scores of customers. The procedures performed in this system are described in the following:

(1) Analysis & Feature Extraction [8]

The utterance input sampled in 16bit and 16kHz is divided into 30ms frames overlapped every 10ms. 16 Mel-scaled filter bank coefficients are extracted from each frame and are used to detect isolated words and continuants. To remove the effect of utterance loudness from the entire spectrum envelope, the average of the coefficients from 0 to 1kHz is subtracted from all the coefficients and the coefficients are adjusted for the average of the whole coefficients to be 0. 50 Mel-scaled filter bank coefficients that are especially linear scaled from 0 to 3kHz are extracted from each frame and are used for speaker verification. This scaling adopts another study result that more information about speakers concentrates on the second formant [9]. To remove the effect of utterance loudness from the entire spectrum envelope, the average of the coefficients from 0 to 1kHz is subtracted from all the coefficients and the coefficients are adjusted for the average of the whole coefficients to be 0.

(2) Detecting Isolated Words & Continuants

Isolated words and continuants are detected using another MLP learned to detect all the continuants and silence in speaker-independent mode.

(3) Learning MLP-II with Enrolling Speaker for Each Continuant

For each continuant, the continuants detected from the isolated words are input to corresponding MLP-I and outputs of the MLP-I are averaged. Then the background speakers to present their output averages more than the preset threshold θ are selected. MLP-IIs learn enrolling speaker with the selected background speakers for each continuant.

(4) Evaluating Speaker Score for Each Continuant

For each continuant, the all the frames detected from the isolated words are input to the corresponding MLP-II. All the outputs of the MLPs are averaged.

(5) Comparing Speaker Score with Threshold

The final reject/accept decision is made by comparing a predefined threshold with the average of the Step (4)

V. Experiments

This paper uses the implemented MLP-based speaker verification system and a speech database described in [10], and experiments on them with the same experiment condition to [10] in order to evaluate the combination of the DCS and the OIL as well as the individual methods proposed in the previous sections. In the experiments, the combination is compared with the online EBP algorithm. Also, enrolling time improvements by the individual methods and the combination are measured.

The results of all experiments are presented in Fig. 1. Experiments are conducted to evaluate the performances of the online EBP, OIL, and DCS with the OIL. In the figure, OnEBP designates the online EBP, the numbers the preset thresholds in the DCS, learning duration and number of learned patterns the averaged duration and patterns, respectively, to enroll one speaker for the entire speech database, and error rate the equal error rate. The performance of the online EBP is evaluated with the optimized learning parameters, i.e. learning rate and learning objective error energy [10]. The figures for the OIL are measured with $\lambda = 0.3$ for the same learning parameters to the online EBP. For the measurements of the DCS with the OIL, the optimal results can be taken at the threshold -0.999 because the numbers larger than this make

higher verification errors. On the basis of the online EBP algorithm, the OIL achieves a quite improvement in enrolling duration without any lose in verification error. With the OIL applied to, the DCS keeps the duration decreasing as the threshold increases. From the results, it can be known that both the two methods are effective to shorten the learning duration.

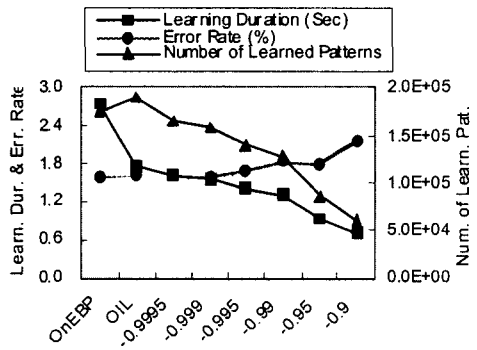

Fig. 1. Experimental results of the online EBP, OIL, and DCS with the OIL.

The final performance evaluations for each experiment are arranged in Fig. 2. With the same verification error to the online EBP, the DCS marks 14.6% improvement and the OIL 55.6%. The combination of the two methods improves enrolling duration by 75.6% over the online EBP. The higher result of the combination than the ones of the OIL and the DCS demonstrates that the two methods operate on different optimization principle and make a synergy when they are employed at a time.

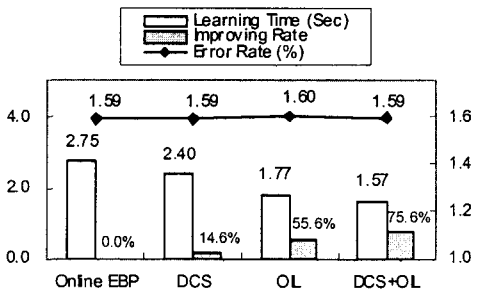

Fig. 2. Performance comparison of all methods.

VI. Conclusion

So far drift speaker enrolling problem has been studied to provide high usability to MLP-based speaker verification systems. While MLPs have great potential on the application to speaker verification, they suffer from poor learning speed. Many users may call for instant enrolling for speaker verification system, so such defect of MLPs must be amended for the attractive usage of the systems. To resolve the problem, this paper fused the existing two studies to enhance speaker enrolling speed for MLP-based speaker verification systems. By conducting several experiments on real speech database, it was acquired that the previous methods are based on distinct reduction basis and it can be concluded that the combination of the methods is very effective to shorten speaker enrolling duration for the speaker verification systems based on MLPs.

References

[1] A. E. Rosenberg, S. Parthasarathy, "Speaker Background Models for Connected Digit Password Speaker Verification," Proceedings of ICASSP, vol. 1, pp. 81-84, 1996.

[2] Y. Bengio, Neural Networks for Speech and Sequence Recognition, International Thomson Computer Press, 1995.

[3] T. S. Lee et al., "A Method on Improvement of the Online Mode Error Back-propagation Algorithm for Pattern Recognition," LNCS, vol. 2417, pp. 275284, 2002.

[4] S. Lawrence, C. L. Giles, "Overfitting and Neural Networks: Conjugate Gradient and Back propagation," Proceedings of IJCNN, vol. 1, pp. 114-119, 2000.

[5] Y. LeCun, "Generalization and Network Design Strategies," Dep. of Comp. Sc., University of Toronto, 1989.

[6] T. S. Lee et al., "Faster Speaker Enrollment for Speaker Verification Systems Based on MLPs by Using Discriminative Cohort Speakers Method," LNAI, vol. 2718, pp. 734-743, 2003.

[7] T. S. Lee et al., "A Qualitative Discriminative Cohort Speakers Method to Reduce Learning Data for MLP-Based Speaker Verification Systems," LNCS, vol. 2690, pp. 1082-1086, 2003.

[8] C. Becchetti, L. P. Ricotti, Speech Recognition, John Wiley & Sons, 1999.

[9] P. Cristea, Z. Valsan, "New Cepstrum Frequency Scale for Neural Network Speaker Verification," Proceedings of

ICECS, vol. 3, pp. 1573-1576, 1999.

[10] T. S. Lee, S. W. Choi, B. W. Hwang, "Speaker Verification System Using Continuants and Multilayer Perceptrons," To be published in Proceedings of the Fall Conference of KIMICS, 2003.