MLP에 기반한 고성능 화자증명 시스템

이태승[O] 최호진
한국항공대학교[O], 한국정보통신대학교
thestaff@hitel.net[O], hjchoi@icu.ac.kr

High Performance MLP-based Speaker Verification System

Tae-Seung Lee[0] Ho-Jin Choi
Hankuk Aviation University, Information and Communication University

Abstract

Speaker verification systems based on multilayer perceptrons (MLPs) have good prospects in reliability and flexibility required as a successful authentication system. However, the poor learning speed of the error backpropagation (EBP) which is representative learning method of MLPs is the major defect to be complemented to achieve real-time user enrollments. In this paper, we implement an MLP-based speaker verification system and apply the existing two methods of the omitting patterns in instant learning (OIL) and the discriminative cohort speakers (DCS) to approach real-time enrollment. An evaluation of the system on a Korean speech database demonstrates the feasibility of the system as a speaker verification system of high performance.

1. Introduction

It is essential for an influential speaker verification system to have advantages in reliability of achieving high verification rate and flexibility of easily accessing the system. The reliability is the most important property for authentication systems by their own duty. A reliable speaker verification system should give verification scores as high as possible in any working conditions. Although authentication systems can provide reliable verification, they are not acceptable to users if access to the systems is not so flexible that users feel some difficulty. A flexible speaker verification system should present fast accessibility in verifying and enrolling users.

Among various pattern recognition engines, which are divided into parametric and nonparametric methods, for verifying identities through speech, multiplayer perceptrons (MLPs) content with the two properties in most efficient way. As nonparametric artificial neural networks, MLPs are assemblies of simple computational nodes and commonly trained by the error backpropagation (EBP) algorithm to classify their learning models. They have superior recognition performance and faster operation speed than representative parametric recognition methods, in that they utilize inter-model information and learning models of an MLP share working capability of the network [1]. In the application to speaker verification, MLPs reveal their abilities as low verification error rate and fast verification process.

However, a defect of slow learning speed drops the merits of MLPs. The standard EBP algorithm for training MLPs is notorious for long learning duration due to its dependency on local gradient [1]. In pattern recognition including speaker verification, learning patterns are required as many as possible to achieve high recognition rate. In speaker verification, abundant background speakers should be reserved to verify a claimant with strict criterion [2]. The plenty of background speakers inevitably causes longer learning duration of the EBP, one's long waiting for completion of enrollment, and ultimately poor flexibility of MLP-based speaker verification systems.

To complement the weakness of the slow EBP algorithm and relieve the burden of plentiful background speakers, Lee et al. suggested two different methods called the omitting patterns in instant learning (OIL) [3] and the discriminative cohort speakers (DCS) [4], [5]. The OIL is to exploit redundancy of pattern recognition data and achieved a substantial improvement in learning speed without losing any recognition rate. The DCS is to select the very background speakers related to enrolling speaker in order to make use of discriminant learning property of MLPs and obtained a rather effective result in enrolling speed. As the OIL works on the inside of the EBP algorithm and the DCS on the outside learning data set, they can be considered as a local and a global optimization, respectively, of speaker enrolling duration in MLP-based speaker verification systems.

In this paper, we present an implementation of MLP-based speaker verification systems and evaluate the performance of the system. The implemented system features low verification error rate and fast working speed but some tedious enrolling process. To improve the latest problem, we combine the two speedup methods of enrolling duration, obtaining better flexibility for the implemented system. This complement will lead the implemented system to show a superior performance in significant aspects.

2. Implemented MLP-based Speaker Verification System

The implemented speaker verification system isolates words from input utterance, classifies the isolated words into nine streams of Korean continuants (/a/, /e/, /ə/, /o/, /u/, /ɨ/, /i/, /l/, nasals), and learns an enrolling speaker for each continuant using MLPs. The system then calculates identity scores for customer speakers. Because the system is based on continuants, which consist of the little phoneme set, it might adapt itself easily to any of text-mode, i.e. text-dependent, text-independent or text-prompt mode. In this system, the text-dependent mode is adopted for easy implementation, in which enrolling text should be the same to verifying text.

The procedure for the system to process the speech of enrolling and verifying speakers consists of (1) analyzing and extracting features from given speech, and detecting isolated words and continuants on the features, (2) training MLPs with an enrolling speaker, and (3) evaluating identity scores of claimants and determining acceptance or rejection of them.

Utterance input sampled in 16-bit and 16-kHz is divided into 30 ms frames overlapped every 10 ms. 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 entire spectrum envelope, average of the coefficients from 0-Hz to 1-kHz is subtracted from all the coefficients and the coefficients are adjusted for average of the whole coefficients to be zero. 50 Mel-scaled filter bank coefficients that are especially linear scaled from 0-Hz to 3-kHz are extracted from each frame and are used for speaker verification. This scaling adopts the study arguing that more information about speakers concentrates on the second formant rather than the first [6]. As with the extraction to detect isolated words and continuants, the same process to remove the effect of utterance loudness is applied here too.

Since the system uses the continuants as speech recognition units, the underlying densities exhibit mono-modal distribution [7]. Thus, it is good enough for each MLP to have a two-layered structure that includes one hidden layer [8], [9]. Since the MLPs need to learn only two models, i.e., one for the enrolling speaker and the other for the background speakers, they can learn the models using one output node and two hidden nodes. In total, nine MLPs are provided for the nine continuants.

3. Fast Speaker Enrolling Methods

MLPs learn the representation of models by establishing decision boundary to discriminate geometrically the model areas. If patterns of all models are fully presented in iterative manner and the internal learnable weights of an MLP are adjusted so that all patterns of each model are classified into its own model area, decision boundary can be finally settled in an optimal position.

The online mode EBP algorithm updates the weights of an MLP using the information related to given pattern and current status of weight vector [1]. The usefulness of given pattern in current epoch can be determined on the criterion of error energy objective. In the online mode EBP, achievement of learning in current epoch is measured with the error energy averaged for all N patterns like this:

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

where, e_p the summed error energy from all M output nodes for given pattern, e_k the error between network value of output node k and learning objective, and t the epoch count. Learning continues until the average error energy $e_{avg}(t)$ is less than the learning objective e_{obj}. The relationship between average error energy and error energies of individual patterns can be described as follows:

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

where, $e_C^2(n)$ is the error energy of the output node C associated with given pattern and n the update count of weight vector. This expression means that if $e_C^2(n)$ s for all learning patterns are less than or equal to $2e_{obj}$, then the learning is complete, assuming that the learning is progressed sufficiently to be able to ignore the other output values beside C. As a result, it is possible to learn only the patterns of $e_C^2(n) > 2e_{obj}$ to complete learning. In Eqn 2 the coefficient λ is inserted to determine the depth of patterns which weight vector is to be updated. When λ is near 1, the number of omitted patterns increases but the count of learning epochs increases as well. Hence it is necessary to search for a proper λ to achieve the minimum count of learning epochs and the maximum number of omitted patterns so that the shortest learning duration is obtained. The omitting weight updates of useless patterns on the criterion in Eqn. 2 is the OIL method.

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

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

$$S_{Cohort} = Sel_{M_{MLP} \ge \theta, I}(Sort_{Dec}(M_{MLP}(S_{BG} \mid \mathbf{X}))), \qquad S_{BG} = \{S_i \mid 1 \le i \le I\} \qquad (3)$$

where, $\mathbf{X}$ is the speech of enrolling speaker, S_{BG} the background speakers set which population is I, M_{MLP} the MLP function which evaluates likelihoods of the given $\mathbf{X}$ to the background speakers. $Sort_{Dec}$ stands for the function to sort given values in descending manner, $Sel_{M_{MLP} \ge \theta, I}$ for the function to select relevant background speakers whose M_{MLP} s exceed the preset threshold θ.

4. Performance Evaluation and Discussion

Performance of the implemented system is evaluated in terms of reliability and flexibility. Reliability of speaker verification systems is related to verification error rate and flexibility to working and enrolling speed. In this section, a Korean speech database and its usage are first described and measurements of the implemented system for the two properties are presented. Then, the two methods described in Section 3 are applied to the system and the improvement is reported. Finally, the possibility of further improvement in reliability of the system is discussed.

The speech data used in this evaluation are the recorded voice of connected four digits, spoken by 40 Korean male and female speakers. The digits are ten Arabic numerals pronounced in Korean as /goN/, /iI/, /i/, /sam/, /sa/, /o/, /yug/, /cil/, /pal/, /gu/, each corresponding to a digit from 0 to 9. The average duration of each 4-digit string is about 1 to 1.5 second. Each speaker utters 35 words of different 4-digit strings four times, when the utterance is recorded in 16-bit resolution and 16-kHz sampling. Three of the four utterance samples are used to enroll the speaker, and the last utterance is used for verification. In order to learn the enrolling speakers discriminatively, additional 29 male and female speakers are participated as background speakers for MLPs other than the above 40 speakers.

Each of the 40 speakers can be treated as both the enrolling speaker and the test speaker. When one of them is picked as the test speaker, then the other 39 speakers are used as imposters. As a result, 35 tests using the 35 words are performed for a true speaker and 1,365 (35 * 39) tests for the imposters. In total, we performed 1,400 (35 * 40) trials of test for true speaker and 54,600 (35 * 40 * 39) trials for imposters.

In the results of the evaluation, EER stands for equal error rate, and the number of learning epochs for average number of epochs used to enroll a speaker for an isolated word. These values are calculated by taking the average of values obtained from three trials of learning, each trial being set to the same MLP conditions. The evaluation is conducted on a 1 GHz personal computer machine.

The best performance of the implemented system achieved with the online EBP algorithm when the learning rate of 0.5 and the learning objective error energy of 0.005 are selected is summarized in Table 1. In the table, the enrolling frames and verifying frames designate the extracted average frames of continuants from enrolling three 4-digit strings and verifying one, respectively, and the durations mean processing time to enroll and verify a speaker. As seen in the table, the EER is good in consideration of the length of enrolling (about 3 seconds) and verifying (about 1 second) utterances when compared with the existing parametric speaker verifications [2], [10]. Especially,

the verifying duration, which is measured in milli-second, is excellent because most parametric speaker verifications take long verifying durations by computing matrices [10], [11]. However, the enrolling duration is very slower than a single Gaussian model though faster than the Gaussian mixture model [11]. The MLPs used in this paper correspond to the single Gaussian model since mono-modal distribution of speech generation probability is assumed.

Table 1. The best performance of the implemented system with the online EBP algorithm

EER (%)	Number of Enrolling Frames	Number of Verifying Frames	Enrolling Duration (sec)	Verifying Duration (millisec)
1.59	164.2	53.5	2.7	0.86

To shorten the enrolling duration, the OIL and DCS methods are applied to the implemented system. The applications are evaluated in sequence of the OIL and the DCS combined with the OIL, and the performances are compared with that of the online EBP algorithm. The results of all evaluations are presented in Fig. 1. In the figure, OnEBP designates the online EBP and the numbers on the bottom the preset threshold θ s in the DCS. The figures for the OIL performance are measured with the learning rate of 1, the learning objective error energy of 0.005, and λ of 0.3. In the measurements of the DCS combined with the OIL, the optimal result can be taken at $\theta = -0.999$ because the numbers beyond the point make higher verification errors. Comparing with the online EBP algorithm, the OIL achieves a quite improvement in enrolling duration without making verification error worse. With the OIL applied, the DCS keeps the learning duration decreasing as the threshold increases. From the results, it can be known that the combination of the two methods is effective to shorten the enrolling duration over the individual methods.

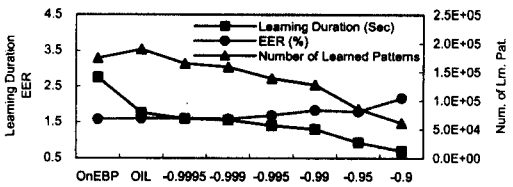

Fig. 1. Evaluation results of the online EBP, the OIL, and the DCS with the OIL

Keeping up the same level of verification error as the online EBP algorithm, the DCS marks the improvement of 14.6 % and the OIL 55.6 % over the online EBP. The combination of the two methods further improves enrolling duration by 75.6% over the online EBP. The better result of the combination to those of the OIL and the DCS demonstrates that the two methods operate on different optimization principles and make a synergy when they are employed together. As a result of the improvement, the enrolling duration of the implemented system achieves 1.5 seconds from 2.7.

The reliability of the implemented systems could be further improved when looking after the tendency of verification errors according to the numbers of background speakers and continuants. In figure (a) of Fig. 2, the declining rate of the EERs is near-linear as the number of background speakers increases from 6 through 10, 16, 20, and 26 to 29. Therefore if more background speakers are given, it can be expected for EER to be more lowered. In figure (b) of Fig. 2, EERs are plotted according to the numbers of continuants and extracted frames of the continuants included in enrolling 4-digit strings. As seen in the figure, if more than seven continuants are included in all enrolling utterances, i.e. the enrolling utterances contain continuants evenly, lower EER than 1.59 % can be achieved. It is noted that the amount of enrolling frames hardly affects EER when more than some 220 enrolling frames are given. The EERs as to the number of continuants are recorded in the first three columns of Table 2.

It is also worth evaluating EERs when the sex of verifying speaker is the same as and different from that of enrolling speaker. It has been reported that large part of verification error is occurred with parametric speaker verifications when the sexes are different [12]. Such result is inferred from that the amount of training data is insufficient, hence parameters are inaccurately estimated. In the system implemented using MLPs, the opposite results are presented with the same situation, i.e. superior EER is obtained for different sexes. The results can be inferred from the learning characteristic of MLPs to establish decision boundary to discriminate adjacent models.

Table 2. Verification errers to various conditions

	> 6 Continuants	> 7 Continuants	> 8 Continuants	Same Sexes	Different Sexes
EER (%)	1.15	0.89	0.81	1.84	1.01

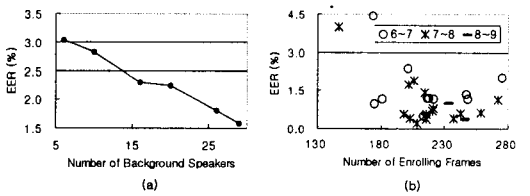

Fig. 2. Verification error tendency according to the numbers of (a) background speakers and (b) continuants

Putting the results of the evaluation together, MLPs show good properties in the application to speaker verification in terms of reliability and flexibility required in a successful authentication system. Although the major weak point of MLPs, slow learning speed, courses enrolling duration of the system to be prolonged and refused by users, it can be complemented by adopting the combination of the existing methods for improving enrolling speed.

References

1. Bengio, Y.: Neural Networks for Speech and Sequence Recognition. International Thomson Computer Press, London Boston (1995)
2. Rosenberg, A. E., Parthasarathy, S.: Speaker Background Models for Connected Digit Password Speaker Verification. ICASSP 1 (1996) 81-84
3. Lee, T., Choi, H., Kwag, Y., Hwang, B.: A Method on Improvement of the Online Mode Error Backpropagation Algorithm for Pattern Recognition. LNAI 2417 (2002) 275-284
4. Lee, T., Choi, S., Choi, W., Park, H., Lim, S., Hwang, B.: Faster Speaker Enrollment for Speaker Verification Systems Based on MLPs by Using Discriminative Cohort Speakers Method. LNAI 2718 (2003) 734-743
5. Lee, T., Choi, S., Choi, W., Park, H., Lim, S., Hwang, B.: A Qualitative Discriminative Cohort Speakers Method to Reduce Learning Data for MLP-Based Speaker Verification Systems. LNCS 2690 (2003) 1082-1086
6. Cristea, P., Valsan, Z.: New Cepstrum Frequency Scale for Neural Network Speaker Verification. ICECS 3 (1999) 1573-1576
7. Savic, M., Sorensen, J.: Phoneme Based Speaker Verification. ICASSP 2 (1992) 165-168
8. Delacretaz, D. P., Hennebert, J.: Text-Prompted Speaker Verification Experiments with Phoneme Specific MLPs. ICASSP 2 (1998) 777-780
9. Lippmann, R. P.: An Introduction to Computing with Neural Nets. IEEE ASSP Magazine 4 (1987) 4-22
10. Zhang, Y., Zhu, X., Zhang, D.: Speaker Verification by Removing Common Information. Electronics Letters 35 (1999) 2009-2011
11. Zilca, R. D.: Text-independent Speaker Verification Using Utterance Level Scoring and Covariance Modeling. IEEE Trans. on Speech and Audio Processing 10 (2002) 363-370
12. Parris, E. S., Carey, M. J.: Language Independent Gender Identification. ICASSP 2 (1996) 685-688