지속음 및 다층신경망을 이용한 화자증명 시스템

이태승* · 최성원* · 황병원*
*한국한공대학교

Speaker Verification System Using Continuants and Multilayer Perceptrons

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

요 약

생체정보를 활용하여 개인정보를 보호하는 기술 가운데 화자증명은 다양한 사용편의성과 구현비용 면에서 이점을 갖고 있어 폭넓은 활용이 기대된다. 화자증명은 증명성능의 신뢰성, 음성문장 사용의 유연성, 증명시스템 복잡도의 효율성 면에서 높은 수준을 달성해야 한다. 지속음은 화자 구별력이 뛰어나며 구별되는 종류가 한정적이고, MLP(multilayer perceptron)는 높은 패턴인식률과 신속한 동작성능을 갖고 있어 화자증명 시스템이 이와 같은 특성을 달성하기 위한 유력한 수단을 제공한다. 본 논문에서는 지속음과 MLP를 적용한 시스템을 구현하고 한국어 음성 데이터베이스를 이용하여 이 시스템의 성능을 측정하고 분석한다. 실험의 결과는 지속음이 세 가지 특성에 대해 우수한 효과를 가지며 MLP가 높은 신뢰성과 효율성을 달성하는 데 실질적인 도움이 됨을 확인한다.

ABSTRACT

Among the techniques to protect private information by adopting biometrics, speaker verification is expected to be widely used due to advantages in convenient usage and implementation cost. Speaker verification should achieve a high degree of the reliability in the verification score, the flexibility in speech text usage, and the efficiency in verification system complexity. Continuants have excellent speaker-discriminant power and the modest number of phonemes in the category, and multilayer perceptrons (MLPs) have superior recognition ability and fast operation speed. In consequence, the two provide viable ways for speaker verification system to obtain the above properties. This paper implements a system to which continuants and MLPs are applied, and evaluates the system using a Korean speech database. The results of the experiment prove that continuants and MLPs enable the system to acquire the three properties.

키워드
화자증명, 생체인식, 음성처리, 지속음, 다층신경망

Ⅰ. Introduction

Among the acceptable biometric-based authentication technologies, speaker recognition has many advantages due to its convenient usage and low implementation cost. Speaker recognition is a biometric recognition technique based on speech. It is classified into speaker identification and speaker verification. The former enrolls multiple speakers for system and selects one speaker out of them associated with the given speech. As compared with it, the latter selects the speaker enrolled with system previously and claimed by a customer, and decides whether the given speech of the customer is associated with the claimed speaker. The studies for speaker verification are being conducted more widely and actively because it covers speaker identification in technical aspect [1].

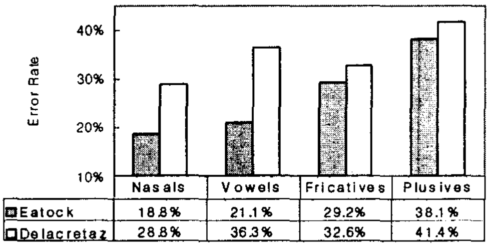
For speaker verification to be influential, it should achieve a certain degree of three properties: the reliability in the verification score of the implemented system, the flexibility in the usage of speech text, and the efficiency in the complexity of verification system. First, the reliability of verification score is the most important property to authentication system. Authentication system should give as high verification score as possible at any adverse situation. Second, the flexibility of speech text usage is required for users to access the system with little effort. To resolve the overall characteristics of a speaker, utterances with the various organ positions of the vocal tract must be provided and this burdens users with long, various and laborious uttering [2]. Hence, it must be considered that a rather short utterance might be sufficient for satisfying a high verification score. Third, the efficiency of system complexity has to be achieved for low implementation cost [3]. To prevent system from feigned accesses, many speaker verification systems are implemented in text-prompted mode [4]. Although text-prompted mode can give immunity against the improper accesses that record the speech of an enrolled speaker and present the recording to system, it requires speech recognition facility to recognize the language units out of a text. Complex and advanced speech recognition involves in increasing implementation cost of the entire system.

To content with the three properties, this paper implements a speaker verification system based on continuants and multilayer perceptrons (MLPs). Continuants have excellent speaker-discriminant power and the small number of classes, and MLPs have superior recognition ability and fast operation speed [2],[5],[6]. Hence, it is expected that they enable the system to have the properties and achieve high speaker verification performance.

II. Continuants and MLPs for Speaker Verification

Speaker verification system should accomplish high reliable verification score, flexible usage, and efficient system implementation. Continuants can realize the three properties in a speaker verification system and MLPs help such system acquire more reliability and efficiency. In this section, the feasibilities of continuants and MLPs to speaker verification system are briefly discussed.

Continuants have advantages as a major language unit in speaker verification. Speaker verification can be understood by a vocal track model consisting of multiple lossless tubes [11]. In view of this model, modeling of speech signal based on language information is necessary because intra-speaker variation is bigger than inter-speaker variation, i.e. speaker information come from inter-speaker variation is apt to be overwhelmed by language information come from intra-speaker variation [2]. The reliability, flexibility, and complexity of speaker verification system are determined by what language unit is selected. Of various language units, phonemes can reflect efficiently the former two properties. Phonemes are atomic language units. Thus, the characteristics of different speakers are finely discriminated within a phoneme and any word can be composed by phonemes. However, speaker verification capabilities vary as to phonemic categories mainly due to the steadiness and duration of their voicing. Eatock et al. and Delacretaz et al. have studied such variations in verification capability and their works are summarized in Fig. 1 [5], [12]. Continuants feature continuous and unconstraint voicing and include the best ones, nasals and vowels in Fig. 1. They show more verification capability than other phoneme categories, so enhance reliability. Along with it, continuants can be easily detected by speech recognition facility because of their long voicing and low number of kinds. As a result, continuants can enhance largely the implementation efficiency of speech recognition facility as well as improve the flexibility of composition into any verification word with higher verification reliability.



	Nasals	Vowels	Fricatives	Plusives
▨Eatock	18.8%	21.1%	29.2%	38.1%
□Delacretaz	28.8%	36.3%	32.6%	41.4%

Fig. 1. Error rates for the various phonemic categories reported in Eatock et al. and Delacretaz et al.

MLPs are a classifying method suitable for speaker verification due to their higher recognition rate and faster recognition speed than the existing systems based on parametric methods. MLPs learn decision boundaries to discriminate optimally between models. For speaker verification, MLPs have two models needed to be classified, i.e. enrolling speaker and background speakers. Such MLPs which have only two learning models present the similar effectiveness to the cohort speakers method developed in the existing parametric-based speaker verification, in which the cohort consists of the background speakers closest to an enrolling speaker [13]. However, the cohort speakers method based on probability density functions might derive a false recognition result according to the distribution densities of enrolling speaker and background speakers. That is, if the density of enrolling speaker is lower and the one has more variance than those of background speakers, then speakers even far from the enrolling speaker might be accepted. On the other hand, MLPs avoid such problem because it discriminates the two models on the basis of their discriminative decision boundary. Fig. 2 demonstrates such a situation when enrolling speaker is male and customer is female, and compares the cohort speakers method with MLPs. In addition to it, MLPs achieve a superior verification error rate since they need not to assume any probability distribution of underlying data [6].

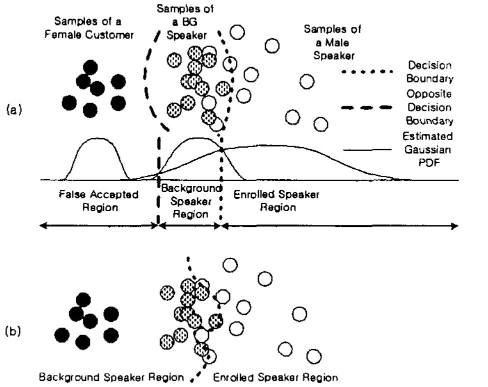


Fig. 2. A comparison of the cohort speakers method and MLPs for a specific error situation: (a) decision boundary by Gaussian probability density functions in the cohort speakers method; (b) discriminative decision boundary in MLPs for the same situation to (a)

It is finally noted that the reason that MLPs show faster recognition speed can be analyzed as all the background speakers are merged into a model. The merging enables for MLPs to have no need to calculate likelihoods for each background speaker at verifying identifications [13].

III. Implemented System

This paper implements a speaker verification system based on continuants and MLPs. Because this 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 [7]. However, the text-dependent mode is adopted in this system for easy implementation, in which enrolling text should be the same to verifying text. 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 MLPs for each continuant, and calculates identity scores of customers. The procedures performed in this system are outlined in Fig. 3 and each procedure is described in the following:

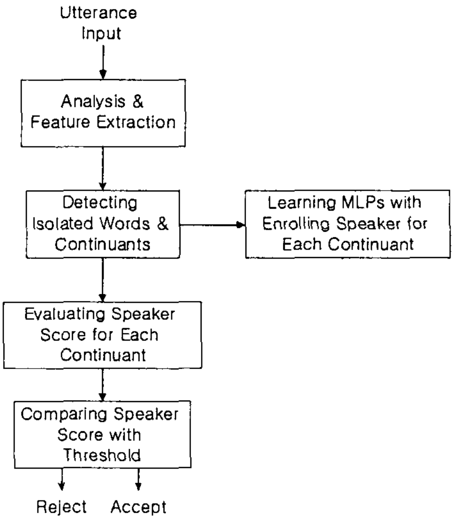

Utterance Input
Analysis & Feature Extraction
Detecting Isolated Words & Continuants
Learning MLPs with Enrolling Speaker for Each Continuant
Evaluating Speaker Score for Each Continuant
Comparing Speaker Score with Threshold
Reject
Accept

Fig. 3. The process flow of the MLP-based speaker verification system

(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 MLPs with Enrolling Speaker for Each Continuant

For each continuant, the frames detected from the isolated words are input to the corresponding MLP and the MLP learns enrolling speaker with background speakers.

(4) Evaluating Speaker Score for Each Continuant

For each continuant, all the frames detected from the isolated words are input to the corresponding MLP. 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).

Since this speaker verification system uses the continuants as speaker recognition unit, the underlying densities show mono-modal distributions [2]. So it is enough for each MLP to have two layers structure that includes one hidden layer [5],[15]. And since the number of models for the MLPs to learn is 2, the MLPs can learn the models using only one output node and two hidden nodes. Nine MLPs in total are provided for nine continuants.

Ⅳ. Performance Evaluation

To evaluate the performance of the speaker verification system, an experiment is conducted using a Korean speech database. This section records the results of the evaluation.

A. Speech Database

The speech data used in this experiment are the recording of connected four digits spoken by 40 Korean male and female speakers, in which the digits are Arabic numerals each corresponding to /goN/, /il/, /i/, /sam/, /sa/, /o/, /yug/, /cil/, /pal/, /gu/ in Korean pronunciation. Each of the speakers utters total 35 words of different digit strings four times, and the utterances are recorded in 16kHz sampling and 16bits resolution. Three of the four utterances are used to enroll speakers, and the other to verify. As background speakers for MLPs to learn enrolling speakers discriminatively, 29 Korean male and female speakers except for above 40 speakers are used.

B. Experimental Condition

In this performance evaluation, MLP learning to enroll a speaker is set up as follows [6]:

- MLPs are learned with the online mode error backpropagation (EBP) algorithm.
- Input patterns are normalized such that the elements of each pattern are into the range from -1.0 to +1.0.
- The objective of output node, i.e. training target, is +0.9 for enrolling speaker and -0.9 for background speakers to obtain faster EBP learning speed.
- Speech patterns of two models are presented in alternative manner during learning. In most cases, the numbers of patterns for the two models are not the same. Therefore, the patterns of the model having fewer patterns are repetitively presented until all the patterns of the model having more patterns are once presented, completing one learning epoch.
- Since learning might be fallen in a local minimum, the maximum number of learning epochs is limited to 1000.

Each of the 40 speakers is regarded as both enrolling speaker and true speaker, and when a speaker out of them is picked as true speaker the other 39 speakers are assigned to imposters. As a result for each speaker, 35-time tests are performed for true speaker and 1,560-time tests for imposter. As a whole, the experiment performed 1,400 trials for true speaker tests and 54,600 trials for imposter tests.

The experiment is conducted on an 1GHz personal computer machine. In the experiment result, the error rate designates the equal error rate, the number of learning epochs the averaged number of epochs used to enroll a speaker for a digit string word and the learning time the overall time taken to learn these patterns. Values of error rate, the number of learning epochs, and learning times are the averages for the results of three-time learning each with the same MLP learning condition to compensate for the effect of the randomly selected initial weights.

C. Evaluation Results

The learning parameters to be considered in the MLP learning using the EBP algorithm include learning rate and learning objective error energy [6]. Learning rate is the parameter to adjust the updating level of the internal weight vector in MLPs. Learning rate that is too large or small value tends to prolong learning duration and becomes a cause to increase error rate since such value makes learning oscillate around the optimal learning objective or grows the number of learning epochs until the objective is reached. Error energy gauges the difference between the desired output vector and the current output vector and learning objective error energy is the objective that MLPs must get to for the given learning data. Although error rate decreases as low learning objective error energy is taken, the number of learning epochs increases along with it, and especially for too low learning objective error energy taken it is possible for error rate to get worse. As a consequence, it needs to determine the proper learning objective error energy and learning rate in the MLP learning using the EBP.

The performance of the implemented system as to learning rates for the EBP algorithm is depicted in Fig. 4. Those values in the figure are to pursue the trajectories of the numbers of the learning epochs and the errors when learning objective error energy is fixed to 0.01. As seen in the figure, when the best learning is achieved, the point of learning rate is 0.5, the number of learning epochs 172.3, and the error rate 1.65%.

The performance of the implemented system as to learning objective error energy for the EBP algorithm is depicted in Fig. 5. Those values in the figure are to pursue the trajectories of the numbers of the learning epochs and the errors when learning rate is fixed to 0.5 as determined in Fig. 4. As seen in the figure, when the optimal learning is achieved, the point of learning objective error energy is 0.005, the number of learning epochs 301.5, and the error rate 1.59%.

The numbers of the error rates, the extracted continuants, and the frames in the continuants are presented in Table 1. These values are averaged for the digit strings with less than 1% error rate when the optimal learning parameters for the EBP learning are taken. Although the experimental database includes 35 different digit strings, such selection of the digit strings is meaningful to the speaker verification system based on continuants and MLPs since great difference in error rate is made from how many continuants are contained in the digit strings and frames in the continuants. As seen in the table, if more than 7.5 out of 9 continuants carry on about 1.6 seconds, the error rate at most 0.6% can be obtained. It is noted that the durations of the utterances to verify identities are about 1/3 compared with the utterances to enroll ones, although these durations are not seen in the table. From the results of the figures 2 and 3 and table 1, it can be found out that the utterance that is modestly short but includes various continuants is sufficient enough to obtain high speaker verification performance in the reliability, flexibility and efficiency for the speaker verification system based on continuants and MLPs.

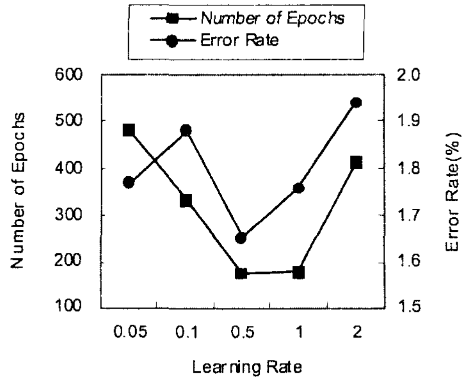


Fig. 4. The performance points of the system with the different learning rates

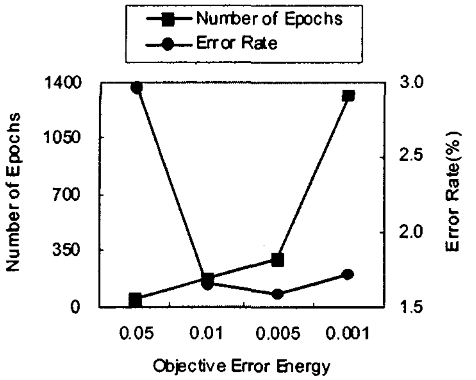


Fig. 5. The performance points of the system with the different learning objective error energy

Table 1. Performance evaluation results for the digit strings presenting less than 1% error rate

EER(%)	Number of Continuants	Number of Frames (1 frame = 10ms)
0.58	7.5	168.3

Ⅴ. Conclusion

The results of the experiment brought out that continuants and MLPs can show the credibility, elasticity and practicality for the application to speaker verification. To be effectual, speaker verification should achieve a high degree of the credibility in the verification score, the elasticity in speech text usage, and the practicality in verification system complexity. Continuants have excellent speaker-discriminant power and the little number of their phonemic classes, and MLPs have high recognition ability and fast operation speed. In consequence, the two provide feasible means for speaker verification system to obtain the above properties. This paper implemented the speaker verification system employing continuants and MLP, and evaluated the system using the Korean continuously spoken four-digit speech database. The results of the experiment ascertained that continuants MLPs enabled the system to acquire the three properties. Nevertheless, it was observed that the speaker enrolling speed of MLPs was slower up to 3000 times than the verifying speed of them. In the future work, it needs to shorten the enrolling duration because the duration has an influence on the elasticity of speaker verification system.

References

[1] Q. Li, et al., "Recent Advancements in Automatic Speaker Authentication," IEEE RA Magazine, vol. 6, pp. 24-34, 1999.

[2] M. Savic, J. Sorensen, "Phoneme Based Speaker Verification," Proceedings of ICASSP, vol. 2, pp. 165-168, 1992.

[3] A, Lodi, M. Toma, R. Guerrieri, "Very Low Complexity Prompted Speaker Verification System Based on HMM-Modeling," Proceedings of ICASSP, vol. 4, pp. IV-3912 -IV-3915, 2002.

[4] ChiWei Che, Qiguang Lin, Dong-Suk Yuk, "An HMM Approach to Text-Prompted Speaker Verification," Proceedings of ICASSP, vol. 2, pp. 673-676, 1996.

[5] D. P. Delacretaz, J. Hennebert, "Text-Prompted Speaker Verification Experiments with Phoneme Specific MLPs," Proceedings of ICASSP, vol. 2, pp. 777-780, 1998.

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

[7] S. Furui, "An Overview of Speaker Recognition Technology," Automatic Speech and Speaker Recognition. Kluwer Academic Publishers, pp. 31-56, 1996.

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

[9] P. Cristeal Z. Valsan, "New Cepstrum Frequency Scale for Neural Network Speaker Verification," Proceedings of ICECS, vol. 3, pp. 1573-1576, 1999.

[10] R. P. Lippmann, "An Introduction to Computing with Neural Nets," IEEE ASSP Magazine, vol. 4, pp. 4-22, 1987.

[11] D. O'Shaughnessy, Speech Communications: Human and Machine, IEEE Press, 2000.

[12] J. P. Eatock, J. S. Mason, "A Quantitative Assessment of the Relative Speaker Discriminating Properties of Phonemes," Proceedings of ICASSP, vol. 1, pp. 133-136, 1994.

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