# IMPROVED SPEAKER VERIFICATION OVER THE CELLULAR PHONE NETWORK USING PHONEME-BALANCED AND DIGIT-SEQUENCE-PRESERVING CONNECTED DIGIT PATTERNS

Tsuneo Kato and Tohru Shimizu

KDDI R&D Laboratories Inc.
2-1-15 Ohara, Kamifukuoka-shi, Saitama, 356-8502, Japan
e-mail: tkato@kddilabs.jp

## ABSTRACT

In order to achieve high accuracy text-prompted speaker verification over the cellular phone network, a phoneme-balanced connected digit pattern for enrollment and digit-sequence-preserving connected digit patterns for verification (i.e. patterns preserving partial digit sequences of the enrollment pattern) are proposed. In addition to those, a decision procedure using multiple patterns is designed to overcome the low quality of cellular phone speech. Experimental results showed the phoneme-balanced and digit-sequence-preserving patterns reduced more than 50% of equal error rate compared to the conventional randomly-chosen and randomly-reordered digit patterns. The decision procedure reduced 60% of the error rate. Overall, the error rate obtained by the proposed method was 1% for 99% of clients and 95% of imposters.

## 1. INTRODUCTION

From the viewpoint of authentication services, high verification accuracy, less utterances for enrollment, and robustness to imposture are essential requirements of the speaker verification. Text-prompted speaker verification, in which the pass phrases are specified by the verification system, is an effective method for preventing imposture of playing recorded client speech[1, 2]. As the connected digit patterns have less phonetic contexts than arbitrary phoneme sequences, the text-prompted speaker verification using connected digit patterns [3] could achieve higher accuracy with less enrollment data. However, the conventional method could not achieve adequate accuracy for cellular phone speech because of the lack of enrollment data and the low quality of cellular phone speech.

In order to improve the accuracy with limited enrollment data, the speaker intrinsic acoustic characteristics should be efficiently contained in the limited utterances. Besides, mismatches of the characteristics between enrollment and verification should be small. Therefore, the connected digit patterns for improving the accuracy are investigated for both enrollment and verification.

Some verification errors are caused by the verification utterances degraded by the cellular phone network. In such cases, multiple verification patterns instead of a single pattern increase information on the speaker's characteristics, and improve the accuracy. Therefore, verification using multiple patterns is considered.

In this paper, a phoneme-balanced digit pattern for enrollment and digit-sequence-preserving patterns for verification are proposed. A decision procedure using multiple patterns is also proposed. In section 2, the method of creating the connected digit patterns and the design of the decision procedure using multiple patterns are described. In section 3, effectiveness of the phoneme-balanced patterns and the digit-sequence-preserving patterns are investigated. Finally the total performance of the decision procedure is evaluated.

## 2. PHONEME-BALANCED AND DIGIT-SEQUENCE-PRESERVING PATTERNS FOR SPEAKER VERIFICATION

### 2.1. Phoneme-balanced digit patterns for enrollment

In order to train accurate client models with few utterances for enrollment, the acoustic characteristics of the client should be contained in the limited training data. As the characteristics are reflected in vowels and nasals[4], we assume that the connected digit patterns for enrollment should contain as many of these phonemes as possible for obtaining accurate client models. Because less than 7-digit patterns can be easily repeated by everyone, six digits out of ten are chosen in consideration of phoneme balance.

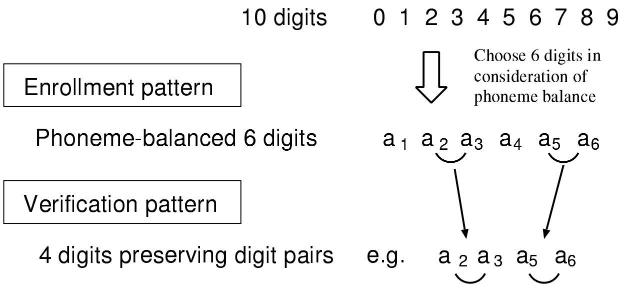
10 digits
0 1 2 3 4 5 6 7 8 9
Enrollment pattern
Choose 6 digits in consideration of phoneme balance
Phoneme-balanced 6 digits
a1 a2 a3 a4 a5 a6
Verification pattern
4 digits preserving digit pairs
e.g. a2 a3 a5 a6

Fig. 1. A 4-digit pattern preserving digit pairs

### 2.2. Digit-sequence-preserving patterns for verification

In general, the accuracy of text-prompted speaker verification is lower than that of text-dependent speaker verification, because mismatch between enrollment and verification becomes larger when the pattern for enrollment and verification is different. To reduce the mismatch of the acoustic characteristics, connected digit patterns which preserve partial digit sequences of the enrollment pattern is introduced for verification.

Fig. 1 shows an example of 4-digit verification patterns which preserve digit pairs contained in a 6-digit enrollment pattern. The 6-digit enrollment pattern contains five digit pairs, four digit triplets and three digit quartets. The number of 4-digit verification patterns which preserve 1) two digit pairs (e.g. “2356”, “4534” from “123456”) and 2) one digit pair (i.e. placing single digits at the first and last places and placing a digit pair in the middle. e.g. “5124”, “1346”), is 202 in total, whereas the randomly chosen 4 digits out of 6 have 1,296 patterns. Though the verification patterns preserving partial digit sequences of the enrollment pattern have less patterns, higher verification accuracy can be expected by the proposed patterns.

### 2.3. Decision procedure using multiple patterns

As the quality of cellular phone speech is not good enough to achieve high accuracy by using just one verification pattern, a decision procedure using multiple patterns is introduced.

In this method, two thresholds for score $S$ (normalized log likelihood) are used for making a decision on three status, ”accept” or ”reject” or ”not decided” for each pattern. Fig. 2 shows determination of the two thresholds. $S_{acc}$ and $S_{rej}$ are determined respectively so that the false acceptance rate (FAR) / false rejection rate (FRR) become lower than the targeted values. The decision is made based on $S$ as follows,

- If $S \geq S_{acc}$, accepted.
- If $S \leq S_{rej}$, rejected.

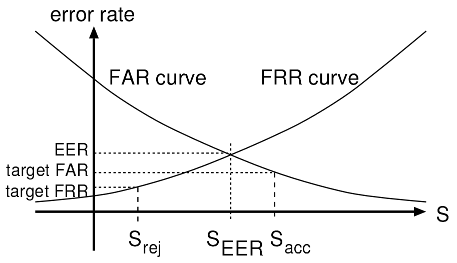

Fig. 2. Determination of the thresholds $S_{acc}$ and $S_{rej}$

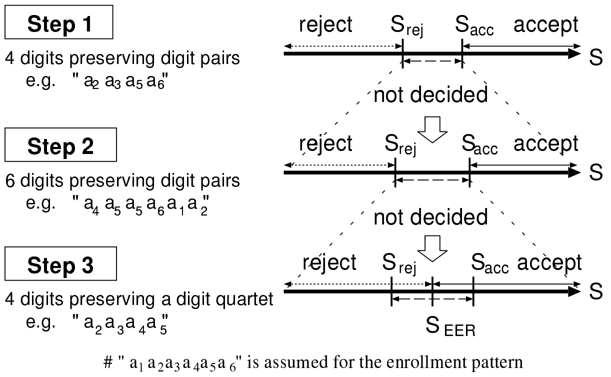

Fig. 3. A decision procedure using multiple patterns

- If $S_{rej} < S < S_{acc}$, not decided.

When the utterance is not decided, the decision is left to the next utterance of a different pattern.

Fig. 3 shows the decision procedure. In general, the patterns preserving short digit sequences have less accuracy but more patterns, whereas the patterns preserving long digit sequences have more accuracy but less patterns. Thus, the decision procedure is designed to prevent imposture by patterns which have enough patterns first, then to make a decision based on patterns which have more accuracy secondly, and to make a decision based on patterns which have the best accuracy at last. The procedure evaluated in section 3.4 follows the 3 steps:

Step 1) Make a decision based on a 4-digit pattern preserving digit pairs

Step 2) If not decided at Step 1, make a decision based on a 6-digit pattern preserving digit pairs

Step 3) If not decided at Step 2, make a decision based on a 4-digit pattern preserving a digit quartet

Table 1. Distance between two log likelihood distributions of clients and imposters for each Japanese digit

| digit | pronunciation | distance |
|---|---|---|
| 0 | zero | 2.24 |
| 1 | ichi | 1.77 |
| 2 | ni | 1.69 |
| 3 | san | 1.88 |
| 4 | yon | 1.82 |
| 5 | go | 1.57 |
| 6 | roku | 1.67 |
| 7 | nana | 2.26 |
| 8 | hachi | 1.84 |
| 9 | kyu | 2.09 |

## 3. EXPERIMENTS

### 3.1. Testset and test conditions

Test data of 81 male and 74 female speakers were collected over the cellular phone network. The performance was evaluated on the two testsets:

Testset A  Utterances ten minutes after enrollment

Testset B  Utterances two months after enrollment

For enrollment, three utterances of a 6-digit pattern were used. For verification, 4-digit patterns or 6-digit patterns were used. The enrolled client HMMs were obtained by Baum-Welch re-estimation from speaker-independent Gaussian mixture digit HMMs. Mean vectors and weights of the Gaussians were re-estimated. The parameters are 12 LPC-MEL coefficients and their derivatives. Before verification process, the utterances are recognized whether they match the prompted pattern or not. Then log likelihood is calculated by using the enrolled client HMMs and the segmentation obtained by speech recognition. The log likelihood of the client HMMs is normalized by that of the speaker-independent HMMs [3], which are also used as the initial model for training the client HMMs. All the combinations of the verification utterances and the client HMMs were examined. The performance was evaluated with a posteriori equal error rate (EER).

### 3.2. Evaluation of the phoneme-balanced digit patterns

First, Table 1 shows the distance between the log likelihood distribution of the clients and that of the imposters for each Japanese digit. The log likelihood distributions were calculated from the verification experiment using 10 randomly-chosen 6-digit enrollment patterns and randomly-reordered 4-digit verification patterns. A greater effect on verification can be expected on digits having a larger distance. In Table 1, ”7” and ”0”, which contain two vowels, have long distances, whereas ”2” and ”5”, which contain a single vowel, have short distances. Five digits of the top six in Table 1 are included in the six digits chosen in consideration of phoneme balance.

Next, Table 2 (a)-(b) show the effect of the phoneme-balanced 6-digit pattern for enrollment instead of the randomly-chosen 6-digit patterns. The phoneme-balanced 6-digit pattern reduced 10% of EER for both Testset A (from 11.0% to 9.5%) and Testset B (from 15.6% to 14.0%).

### 3.3. Evaluation of the digit-sequence-preserving patterns for verification

In the evaluation of the digit-sequence-preserving verification patterns, the following three patterns of different digit length and different preserved sequence length were examined.

(c) 4-digit patterns preserving digit pairs

(d) 6-digit patterns preserving digit pairs

(e) 4-digit patterns preserving a digit quartet

Table 2. (c)-(e) show the effect of the three patterns. For Testset A, the 4-digit patterns preserving digit pairs reduced 50% of the EER from 9.5% to 4.2%. The 6-digit patterns preserving digit pairs reduced the EER to 3.5%. The 4-digit patterns preserving a digit quartet are not good for preventing imposture because those have just three patterns. However those could reduced the EER to 2.6%.

For Testset B, though the EERs were 5% higher than Testset A, they were reduced similarly by the proposed patterns.

### 3.4. Performance of the decision procedure using multiple patterns

The decision procedure which specifies a pattern of (c) in Table 2 first, then specifies a pattern of (d) if “not decided” with the pattern (c), and specifies a pattern of (e) if “not decided” with the pattern (d), was evaluated.

Table 3 shows the proportion of clients' utterances in range $S \geq S_{acc}$ and imposters' utterances in range $S \leq S_{rej}$ at each step of the decision procedure. $n_{c,(i)}$, $n_{a,(i)}$, $n_{i,(i)}$ and $n_{r,(i)}$ represent numbers of clients, accepted clients, imposters, and rejected imposters at $i$th step. Thus, $n_{a,(i)}/n_{c,(i)}$ represents the proportions of the accepted clients to the attempting clients at $i$th step, whereas $n_{a,(i)}/n_{c,(1)}$ being those to the total clients.

Table 2. Reduction of EER by the phoneme-balanced enrollment patterns and the verification patterns preserving partial digit sequence of enrollment patterns

| | Enrollment patterns | Verification patterns | Testset A | Testset B | Patterns |
|---|---|---|---|---|---|
| (a) | randomly-chosen 6 digits | 4 digits randomly reordered | 11.0% | 15.6% | 1296 |
| (b) | phoneme-balanced 6 digits | 4 digits randomly reordered | 9.5% | 14.0% | 1296 |
| (c) | phoneme-balanced 6 digits | 4 digits preserving digit pairs | 4.2% | 9.0% | 202 |
| (d) | phoneme-balanced 6 digits | 6 digits preserving digit pairs | 3.5% | 8.3% | 1024 |
| (e) | phoneme-balanced 6 digits | 4 digits preserving digit quartet | 2.6% | 7.8% | 3 |

Table 3. Proportion of accepted clients and rejected imposters at each step of the decision procedure

| Verification patterns | | | Clients | | Imposters | |
|---|---|---|---|---|---|---|
| | | | $n_{a,(i)}/n_{c,(i)}$ | $n_{a,(i)}/n_{c,(1)}$ | $n_{r,(i)}/n_{i,(i)}$ | $n_{r,(i)}/n_{i,(1)}$ |
| Testset A | | | | | | |
| Step 1 | (c) | 4 digits preserving digit pairs | 89% | 89% | 88% | 88% |
| Step 2 | (d) | 6 digits preserving digit pairs | 70% | 8% | 43% | 5% |
| Step 3 | (e) | 4 digits preserving a digit quartet | 63% | 2% | 36% | 2% |
| | | total | | 99% | | 95% |
| Testset B | | | | | | |
| Step 1 | (c) | 4 digits preserving digit pairs | 80% | 80% | 80% | 80% |
| Step 2 | (d) | 6 digits preserving digit pairs | 50% | 10% | 32% | 6% |
| Step 3 | (e) | 4 digits preserving a digit quartet | 38% | 4% | 51% | 7% |
| | | total | | 94% | | 93% |

Thresholds $S_{acc}$ and $S_{rej}$ are set s.t. $FRR = FAR = 1.0\%$ for Testset A, and set s.t. $FAR = FAR = 3.0\%$ for Testset B.

For Testset A, the thresholds $S_{acc}$ and $S_{rej}$ were set s.t. $FAR = 1.0\%$ and $FRR = 1.0\%$ respectively. The results were as follows,

Step 1)

- 89% of clients accepted ($n_{a,(1)}/n_{c,(1)}$).
- 88% of imposters rejected ($n_{r,(1)}/n_{i,(1)}$).
- 11% of clients and 12% of imposters not decided.

Step 2)

- 70% of clients accepted ($n_{a,(2)}/n_{c,(2)}$).
- 43% of imposters rejected ($n_{r,(2)}/n_{i,(2)}$).
- The rest not decided.

Step 3)

- 63% of clients accepted in range $S \geq S_{acc}$ ($n_{a,(3)}/n_{c,(3)}$).
- 36% of imposters rejected in range $S \leq S_{rej}$ ($n_{r,(3)}/n_{i,(3)}$).
- The rest decided by $S_{EER}$.

In total, 99% of the clients were accepted and 95% of the imposters were rejected with an error rate below 1% for Testset A. The average number of utterances remains to be 1.14.

For Testset B, the two thresholds $S_{acc}$ and $S_{rej}$ were set s.t. $FAR = 3.0\%$ and $FRR = 3.0\%$. Table 3 shows 94% of the clients were accepted and 93% of the imposters were rejected with an error rate below 3%.

## 4. CONCLUSIONS

Phoneme-balanced connected digit patterns for enrollment and digit-sequence-preserving connected digit patterns for verification are proposed to improve accuracy of speaker verification over the cellular phone network. A decision procedure using multiple patterns was designed to obtain higher accuracy. In experiments, the phoneme-balanced patterns and the digit-sequence-preserving patterns reduced more than 50% of the EER. Finally, the decision procedure achieved a 1% error rate for 99% of clients and 95% of imposters.

## 5. REFERENCES

[1] A.Higgins, L.Bahler and J.Porter. "Speaker verification using randomized phrase prompting,". Digital Signal Processing 1, pages 89-106, 1991.

[2] T.Matsui and S.Furui, "Concatenated phoneme models for text-variable speaker recognition,". In Proc. ICASSP 93, pages 391–394, 1993.

[3] A.E.Rosenberg and S.Parthasarathy, "Speaker background models for connected digit password speaker verification,". In Proc. ICASSP 96, pages 81–84, 1996.

[4] Su. L.S. et al., "Identification of speakers by use of nasal coarticulation,". In J. Acoust. Soc. Am. Vol.56, No.5, pages 1876–1882, 1996.