Effects of GSM Speech on Text-Dependent Speaker Verification

GSM 語音對文句相關語者確認效能之影響

Tan-Hsu Tan, Cheng-Hsiung Wu, and Fu-Rong Jean
Department of Electrical Engineering, National Taipei University of Technology
1, Sec. 3, Chung-Hsiao E. Rd., Taipei, 106, Taiwan, Republic of China
E-mail: thtan@ntut.edu.tw

中文摘要

本論文探討 GSM 語音對文句相關語者確認效能之影響。我們依據隱藏式馬可夫模型製作語者確認系統供語者模型訓練與測試之用。為切合實際環境，我們在多種行動條件下錄製了一套 GSM 語音資料庫。最後，本論文針對各種訓練與測試環境進行了一系列的實驗，並提出較前人更為精確的結果。

關鍵詞：文句相關語者確認、隱藏式馬可夫模型、GSM 語音資料庫。

Abstract

This paper investigates the effects of GSM speech on text-dependent speaker verification performance. An HMM-based system is implemented for performance evaluation. In order to match the real-world environments, the full-rate GSM speech database over cellular network is collected under different driving speeds. Experimental results obtained from different combinations of training and test conditions are presented, which provide more accurate results than previous works that used transcoded databases.

Keywords: *text-dependent speaker verification, HMM, GSM speech database.*

Ⅰ. Introduction

With the rapidly growing of mobile phone user, the mobile commerce is becoming the most popular service in financial transactions. For sharing the immense benefits of mobile market, many securities firms attempt to offer more convenient financial services for customers via mobile cellular network. Typical example is the mobile trading system that uses voice-activated transaction interface. With the aid of automatic speech recognition (ASR) technology, such systems make stock exchange and information retrieval in hands-busy and eyes-busy situations possible. The principal problem encountered in such systems; however, is how to test and verify the identity of speakers.

Recently, several researchers have addressed the problems of speaker verification in mobile communication environments [1, 2]. Castellano, *et al*., [1] showed that full rate GSM coder and LPC10 coder significantly degrade the verification accuracy. Besacier, *et al*., [2] indicated that a low LPC order in GSM coding is responsible for the most performance degradations. In both systems; however, only transcoded speech databases were used for performance evaluation. That is, those speech data were generated by the above-mentioned coders with the process of encoding and decoding, and thus can not reflect the real effects of GSM speech in mobile environments.

To match real-world mobile environments, instead of transcoded data, the field full-rate GSM speech of 13 kbps over cellular network was collected in this research. Performance of speaker verification under different mobile conditions are evaluated and compared. The remaining part of this paper is organized as follows. Section 2 describes the implementation of speaker verification system. Experimental results are given and discussed in Section 3.

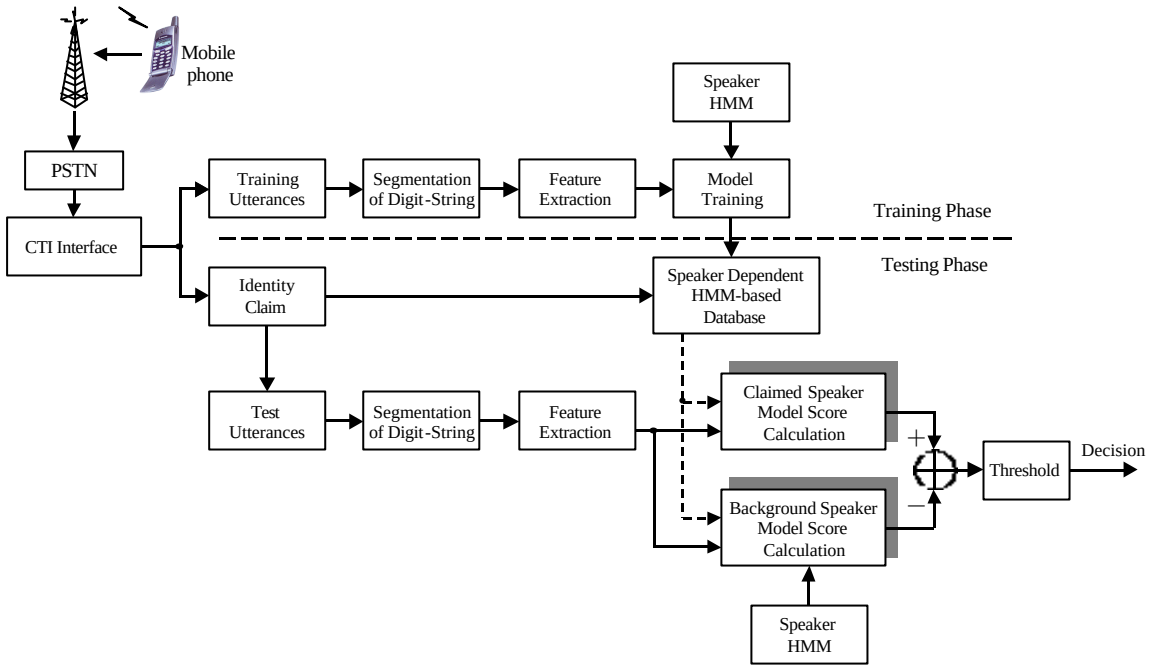

Fig. 1 HMM-based speaker verification system.

Finally, brief concluding remarks are presented in Section 4.

Ⅱ. HMM-Based Speaker Verification System

A text-dependent speaker verification system based on continuous density Gaussian mixture, hidden Markov model (HMM) [3] was implemented for performance evaluation. The main components of the system shown in Fig. 1 are described as follows.

2.1 Speech Database

A Mandarin speech database consisting of 20 male and 20 female speakers was collected from GSM network connected to a computer telephony integrated (CTI) interface. The GSM speech coder belongs to the class of Regular Pulse Excitation-Long Term Predication-linear predictive (RPE-LTP) coders. Since most of the cellular calls are placed inside a vehicle, in this research, three in-vehicle call environments were considered: stopped cars (0 km/hr) with running engine, running cars with driving speeds of 50 km/hr (in urban area) and 90 km/hr (in freeway). Each speaker pronounced 40 7-digit strings at each condition according to Table 1, the same contents as MAT-160 database of ROC Computational Linguistics Society [4]. This resulted in a database that consists of $40 \times 40 \times 3 = 4800$ digit strings.

Table 1 The contents of GSM speech database

0424040	0637223	0830873	1182720	1640233
1642125	1674310	2123213	2036733	2865776
3019988	3261464	3489468	3576446	3582619
4309860	4779899	5081144	5243712	5326477
5492950	5523208	5939183	5962861	6337988
6377001	6393496	6412416	7329669	7362377
7380408	7688544	7992249	8198714	8532261
8631513	8640829	9085035	9560613	9738479

2.2 Hidden Markov Model

The HMM is a finite state statistical structure which has been applied in many applications such as speech recognition and channel modeling [3]. An N-state HMM is defined by the parameter set

$$\lambda = \{\pi_i, a_{ij}, b_i(x), i, j = 1,..., N\}, \quad (1)$$

where

π_i: initial state probability for state i,

a_{ij}: transition probability from state i to state j,

$b_i(x)$: state observation probability density function

(pdf).

2.3 Model Training

Firstly, the training utterances (digit strings) of each speaker are segmented into a sequence of isolated digits based on the measured values of energy and zero crossing rate. Each digit was modeled by a left-to-right HMM of 6 states as shown in Fig. 2 where each state contains 8 Gaussian mixtures. The speech features including 12 mel-frequency cepstral coefficients (MFCCs) and 12 delta MFCCs were extracted for each utterance of 30-ms Hamming-windowed frame with 10-ms frame shift. The extracted features were then used to train the HMM digit model by employing the segmental *K*-means training procedure [3]. Finally, each speaker is represented by a speaker model that comprises those derived HMM digit models.

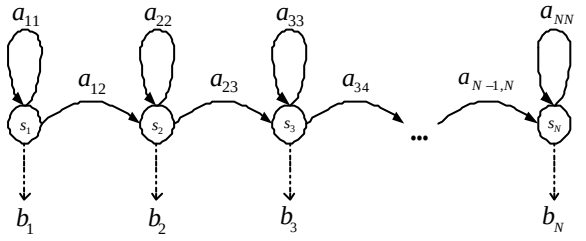

Fig.2 The left-to-right HMM.

2.4 Scoring of Verification

After constructing speaker model, the speaker background model of each specified speaker was then obtained based on similarity measure [5] as follows. The similarity between speaker *i* and speaker *j* is defined as

$$d(\lambda_i, \lambda_j) = \log \frac{P(O_i \mid \lambda_i)}{P(O_i \mid \lambda_j)} + \log \frac{P(O_j \mid \lambda_j)}{P(O_j \mid \lambda_i)} \qquad (2)$$

where

λ_i: the sequence of digit HMMs of speaker *i*,

λ_j: the sequence of digit HMMs of speaker *j*,

O_i: the training utterances of speaker *i*,

O_j: the training utterances of speaker *j*.

Note that the most similar speakers will receive the smallest similarity score. Next, a cohort speaker set is constructed depending on the ranking of measured similarity scores, and finally the first *M* (between 10 and 15) speakers are selected to represent the speaker background model of speaker *i*, which is denoted as $\lambda_{\hat{i}}$.

In the testing phase, the normalized log likelihood score [6] used for verification is calculated against a specified speaker model and its corresponding speaker background model:

$$\log L(O) = \log P(O / \lambda_i) - \log P(O / \lambda_{\hat{i}}) \qquad (3)$$

where

O: the testing utterances,

$P(O / \lambda_i)$: the likelihood related to speaker *i*,

$P(O / \lambda_{\hat{i}})$: the likelihood related to speaker background model of speaker *i*.

The speaker *i* will be accepted if its score is larger than a predetermined threshold.

III. Experimental Results

A series of experiments were conducted to evaluate the effects of GSM speech on the performance of speaker verification by using the system described in Section 2. In all experiments, three test environments were differentiated: (1) stopped cars (0 km/hr) with running engine, (2) running cars of 50 km/hr, and (3) running cars of 90 km/hr. For each mobile environment, utterances of 10, 20, and 30 digit strings are randomly chosen from each true speaker for training, which are denoted by T10, T20, and T30, respectively. All test results are presented in terms of equal-error-rate (EER).

In the first experiment, tests with 30 true speakers (15 male and 15 female) and 10 impostors (5 male and 5 female) are performed with 15-speaker background model. The utterances of impostors and those of true speakers other than training are used for test. Thus, the total number of test trials for the cases of T10, T20, and T30 are respectively equal to 30 (true speakers) $\times$ 40 (total speakers) $\times$ 30 (test strings) = 36000, $30 \times 40 \times 20$ = 24000, and $30 \times 40 \times 10$ = 12000. The results obtained from different combinations of training and test conditions are given in Table 2. It can be seen from the results that the performance is improved with the increase of training data in all test conditions. For example, an improvement of 38.4% was obtained when the training data is increased from T10 to T30 in the matched case of 0 km/hr. In addition, the mismatched conditions present significant performance degradations as compared to the matched conditions, especially in the case of 0 km/hr and 90 km/hr

where the degree of mismatch between training and test conditions is the highest.

In the second and third experiments, a total of 20 male and 20 female speakers were respectively tested, 15 of them are true speakers and the others are impostors. 10-speaker background model is utilized in both experiments. As expected, the experimental results illustrated in Table 3 for 20 male speakers and Table 4 for 20 female are consistent with that of Table 2. That is, the improved performance can be obtained with the increase of training data and matched condition yields the best result. Furthermore, it is of interest to note that the performance of male is much better than that of female, with at most 34.3% in the matched case of 0 km/hr.

Ⅳ. Conclusions

We have investigated the effects of GSM speech on a text-dependent speaker verification system based on hidden Markov model. A field GSM speech database was built from different mobile environments. Experimental results demonstrate that verification performance decreases with the increase of degree of mismatch between training and test conditions. Best performance was obtained in case of matched condition for male speakers. It must also be mentioned that this study gives more accurate results than previous works that used transcoded databases. Consequently, this investigation provides a useful baseline for performance evaluation of speaker verification in mobile ASR-based trading systems. In order to improve verification performance, the future work will focus on the development of compensation techniques for background noise and channel variations.

Acknowledgment

This research was supported by the National Science Council, Republic of China, under the contract NSC-89-2218-E-027-002.

References

[1] P.J. Castellano, S. Sridharan, and S. Boland, “Effects of Speech Coding on Speaker Verification,” *Electronics Letters*, Vol. 32, No. 6, pp. 517-518, March 1996.

[2] L. Besacier, *et al*., “ GSM Speech Coding and Speaker Recognition,” in *Proc*. *ICASSP*, pp. II-1085-II-1088, 2000.

[3] L.R. Rabiner and B.H. Juang, *Fundamentals of Speech Recognition*, Prentice Hall, 1993.

[4] http://rocling.iis.sinica.edu.tw/ROCLING/

[5] A. E. Rosenberg and S. Parthasarathy, “Speaker Background Models for Connected Digit Password Speaker Verification,” in *Proc*. *ICASSP*, pp. 81-84, 1996.

[6] A. Higgins, L. Bahler, and J. Porter, “Speaker Verification Using Randomized Phrase Prompting,” *Digital Signal Processing*, pp. 89-106, 1991.

Table 2 Experimental results (EERs) with 30 true speakers and 10 impostors.

Training data / Test data	0 km/hr			50 km/hr			90 km/hr		
	T10	T20	T30	T10	T20	T30	T10	T20	T30
0 km/hr	0.1395	0.0947	0.0860	0.1597	0.1173	0.1011	0.1630	0.1372	0.1114
50 km/hr	0.1483	0.1154	0.0942	0.1367	0.1082	0.0889	0.1571	0.1335	0.1090
90 km/hr	0.1486	0.1162	0.0984	0.1561	0.1148	0.0978	0.1470	0.1151	0.0948

Table 3 Experimental results (EERs) with male speakers.

Training data / Test data	0 km/hr			50 km/hr			90 km/hr		
	T10	T20	T30	T10	T20	T30	T10	T20	T30
0 km/hr	0.1104	0.0759	0.0654	0.1437	0.1083	0.0980	0.1508	0.1173	0.1054
50 km/hr	0.1283	0.0953	0.0732	0.1336	0.0892	0.0797	0.1435	0.1035	0.0990
90 km/hr	0.1302	0.1022	0.0794	0.1391	0.1014	0.0935	0.1343	0.0928	0.0829

Table 4 Experimental results (EERs) with female speakers.

Training data / Test data	0 km/hr			50 km/hr			90 km/hr		
	T10	T20	T30	T10	T20	T30	T10	T20	T30
0 km/hr	0.1453	0.1157	0.0902	0.1787	0.1398	0.1171	0.1830	0.1427	0.1249
50 km/hr	0.1526	0.1273	0.1004	0.1526	0.1203	0.0972	0.1791	0.1405	0.1297
90 km/hr	0.1603	0.1309	0.1036	0.1760	0.1337	0.1058	0.1650	0.1256	0.1131