ON THE SECURITY OF HMM-BASED SPEAKER VERIFICATION SYSTEMS AGAINST IMPOSTURE USING SYNTHETIC SPEECH

Takashi Masuko†, *Takafumi Hitotsumatsu*†, *Keiichi Tokuda*††, *and Takao Kobayashi*†

†Interdisciplinary Graduate School of Science and Engineering, Tokyo Institute of Technology
††Department of Computer Science, Nagoya Institute of Technology

masuko@ip.titech.ac.jp, tokuda@ics.nitech.ac.jp, Takao.Kobayashi@ip.titech.ac.jp

ABSTRACT

For speaker verification systems, security against imposture is one of the most important problems, and many approaches to reducing false acceptance of impostors as well as false rejection of clients have been investigated. On the other hand, imposture using synthetic speech has not been considered. In this paper, we investigate imposture against speaker verification systems using synthetic speech. We use an HMM-based text-prompted speaker verification system with a false acceptance rate of 0% for human impostors as a reference system, and adopt a trainable HMM-based speech synthesis system for imposture. Experimental results show that false acceptance rates for synthetic speech reached over 70% by training the synthesis system using only 1 sentence from each customer, and current security of HMM-based speaker verification systems against synthetic speech is inadequate.

1. INTRODUCTION

Speaker verification is the technique to judge whether input speech is the same as the claimed speaker's speech. This technique will make it possible to verify the identity of persons accessing systems in various services. For such applications, security against imposture is one of the most important problems, and many approaches to reducing false acceptance of impostors as well as false rejection of clients have been investigated. For example, the text-prompted speaker verification techniques [1]-[3] are robust to the impostor with playing back recorded voice of a registered speaker.

Although most of these researches has assumed human impostors without malice, recently, imposture using converted speech has been reported [4],[5]. However, imposture using synthetic speech has not been taken into account yet, due to the facts that quality of synthetic speech was not enough, and that it was difficult to synthesize speech with arbitrary voice characteristics. Meanwhile, recent advances in speech synthesis make it possible to synthesize speech of high quality (e.g. [6]). As one of these speech synthesis systems, we have proposed an HMM-based speech synthesis system [7] which can synthesize smooth and natural speech. Moreover, we have shown that we can change voice characteristics of synthetic speech to resemble target speaker's voice characteristics by applying speaker adaptation techniques using a small amount of adaptation data [8],[9].

In this paper, from these points of view, we investigate imposture against an HMM-based text-prompted speaker verification system with the HMM-based speech synthesis system.

2. HMM-BASED SPEECH SYNTHESIS SYSTEM

This section describes an overview of the HMM-based speech synthesis system [7] briefly. The HMM-based speech synthesis system consists of two parts; training part and synthesis part. First, in the training part, mel-cepstral coefficients are obtained from speech database by mel-cepstral analysis [10]. Dynamic features, i.e., delta and delta-delta mel-cepstral coefficients, are calculated from mel-cepstral coefficients. Then phoneme HMMs are trained using mel-cepstral coefficients and their deltas and delta-deltas.

In the synthesis part, an arbitrary text to be synthesized is transformed into a phoneme sequence. Then, a sentence HMM is constructed, which represents the whole text to be synthesized, by concatenating phoneme HMMs according to this phoneme sequence. From the sentence HMM, speech parameter sequence is generated using the algorithm described in the next section. Using the MLSA (Mel Log Spectral Approximation) filter [10], speech is synthesized from the generated mel-cepstral coefficients directly.

2.1. Speech Parameter Generation from Continuous HMMs

Let $\boldsymbol{c}_t$ be the mel-cepstral coefficient vector at frame t. Then the dynamic features $\Delta\boldsymbol{c}_t$ and $\Delta^2\boldsymbol{c}_t$, i.e., delta and delta-delta mel-cepstral coefficients at frame t, respectively, are calculated by linear combination of static features as follows:

$$\Delta\boldsymbol{c}_t = \sum_{\tau=-L_1}^{L_1} w_1(\tau)\,\boldsymbol{c}_{t+\tau}, \quad (1)$$

$$\Delta^2\boldsymbol{c}_t = \sum_{\tau=-L_2}^{L_2} w_2(\tau)\,\boldsymbol{c}_{t+\tau}. \quad (2)$$

We assume that the speech parameter vector $\boldsymbol{o}_t$ at frame t consists of the static feature vector $\boldsymbol{c}_t$ and the dynamic feature vectors $\Delta\boldsymbol{c}_t, \Delta^2\boldsymbol{c}_t$, that is, $\boldsymbol{o}_t = [\boldsymbol{c}_t', \Delta\boldsymbol{c}_t', \Delta^2\boldsymbol{c}_t']'$, where $\cdot'$ denotes matrix transpose.

For a given continuous HMM λ and a state sequence $\boldsymbol{Q} = \{q_1, q_2, \ldots, q_T\}$, we obtain a sequence of mel-cepstral coefficient vectors $\boldsymbol{C} = [\boldsymbol{c}_1', \boldsymbol{c}_2', \ldots, \boldsymbol{c}_T']'$ by maximizing $P(\boldsymbol{O} \mid \boldsymbol{Q}, \lambda)$ with respect to $\boldsymbol{O} = [\boldsymbol{o}_1', \boldsymbol{o}_2', \ldots, \boldsymbol{o}_T']'$ under the constraints (1) and (2). Here we assume that the output distribution of each state is a single Gaussian distribution. Without dynamic features (i.e., $\boldsymbol{o}_t = \boldsymbol{c}_t$), it is obvious that $\boldsymbol{C}$ which maximizes $P(\boldsymbol{O} \mid \boldsymbol{Q}, \lambda)$ is equal to $\boldsymbol{M} = [\boldsymbol{\mu}_{q_1}', \boldsymbol{\mu}_{q_2}', \ldots, \boldsymbol{\mu}_{q_T}']'$, where $\boldsymbol{\mu}_{q_t}$ is the mean vector associated with state q_t. On the other hand, with dynamic features, $\boldsymbol{C}$ which maximizes $P(\boldsymbol{O} \mid \boldsymbol{Q}, \lambda)$ is

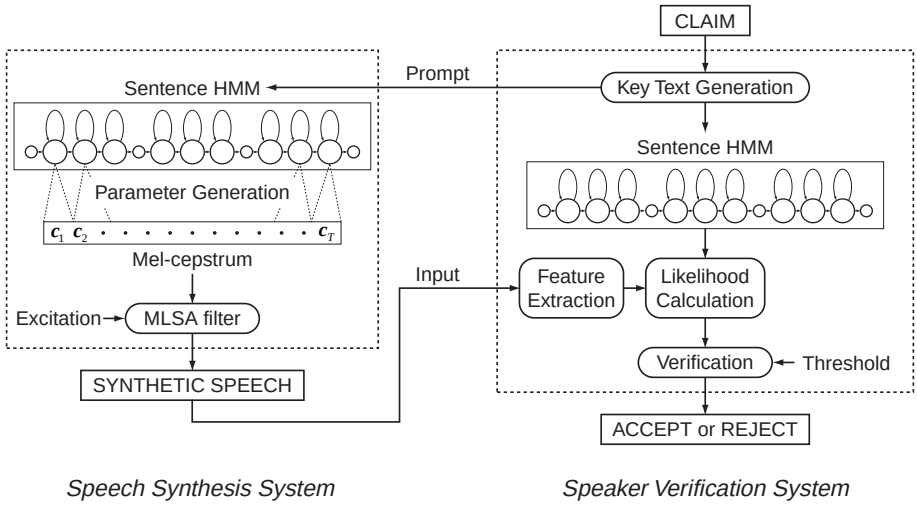

Figure 1: Imposture using the HMM-based speech synthesis system.

determined by a set of linear equations obtained by

$$\frac{\partial}{\partial \boldsymbol{C}} \log P(\boldsymbol{O} \,|\, \boldsymbol{Q}, \lambda) = \boldsymbol{0}, \tag{3}$$

which can easily be solved by a fast algorithm derived in [11],[12]. It has been shown that the obtained mel-cepstral coefficient vectors reflect not only the means of static and dynamic feature vectors but also the covariances of those, and as a result, the synthetic speech is quite smooth and natural sounding.

3. EXPERIMENTS

A block diagram of imposture using the HMM-based speech synthesis system against an HMM-based speaker verification system is shown in *Figure 1*. We focused on a text-prompted speaker verification system [1]-[3], and examined some combinations of the numbers of states of HMMs used in the speech synthesis system and the speaker verification system. In the following experiments, the text confirmation process was omitted because input speech that differs from the key text was not examined.

3.1. Speech Database

We used phonetically balanced Japanese sentences from ATR Japanese speech database. The database consists of sentence data uttered by 20 male speakers; 10 speakers were used as customers and the remainder were used as impostors. Sentence data for each speaker consists of 150 sentences. The database was divided into 3 sets, A-, B-, and C-sets, each set contained 50 sentences for each speaker. A-set was used to train speaker verification system and to determine decision thresholds of normalized log-likelihood, B-set was used to train speech synthesis system, and C-set was used as test sentences. Speech signals were sampled at 10kHz, and segmented and labeled into 48 phonemes (including silence and pause) based on phoneme labels included in the database. Both the speech synthesis system and the speaker verification system used the same phoneme set and the same phoneme sequences corresponding to the sentences.

3.2. Baseline Performance of the Speaker Verification System

We used an HMM-based text-prompted speaker verification system [1]-[3] as a reference system, in which a sentence HMM corresponding to the key text is constructed by concatenating phoneme models, and the normalized log-likelihood of the input speech given the sentence HMM is compared with a threshold to decide whether to accept or reject the speaker.

Speech signals were windowed by a 25.6 ms Blackman window with a 5 ms shift, and the cepstral coefficients were calculated by 15-th order LPC analysis. The feature vector consisted of 16 cepstral coefficients including 0-th coefficient, and their deltas and delta-deltas.

For each customer, speaker dependent (SD) phoneme models were trained using 50 sentences. Speaker independent (SI) phoneme models were also trained using whole training sentences of all customers. Each phone model was a 2- or 3-state 3-mixture left-to-right model with diagonal covariance matrices. Because of limited training data, there were some SD phoneme models which had only few samples for training and remained untrained. In these cases, SI phoneme models were used.

In the verification procedure, normalized log-likelihood $L(s)$ was calculated as follows [3],

$$L(s) = \log P(\boldsymbol{O}|\lambda_s) - \log P(\boldsymbol{O}|\lambda_{all}), \tag{4}$$

where s denotes the claimed speaker, $\boldsymbol{O}$ denotes input speech, λ_s and λ_{all} denote sentence HMMs constructed by concatenating speaker s's phoneme models and SI phoneme models, respectively. Then, the normalized log-likelihood was compared with a prescribed threshold. A speaker dependent threshold is determined using training data to equalize the false acceptance rate (FAR) for impostors and the false rejection rate (FRR) for the customer. However, as shown in *Figure 2*, which shows the FAR and the FRR for training data of speaker m101 using 3-state models, there existed a region in which both the FAR and the FRR were equal to 0%. In this case, we adopted the center of the region as the threshold.

The baseline performance was examined on C-set. The system achieved FARs of 0% for both 2- and 3-state models, while FRRs reached to 7.2% and 9.8% for 2- and 3-state models, respectively.

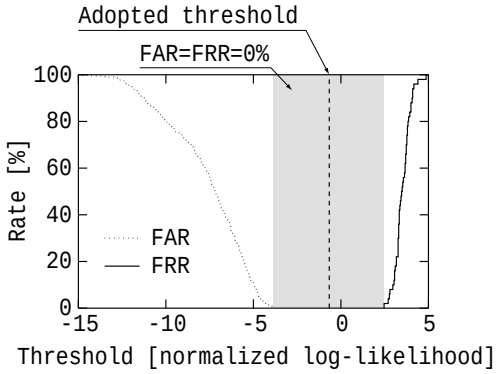

Figure 2: False acceptance and rejection rates for training sentences (speaker m101).

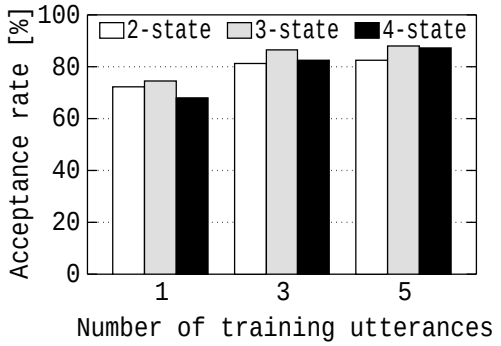

Figure 3: False acceptance rates for synthetic speech against the speaker verification system using 2-state models (speech synthesis system: 2-, 3-, or 4-state models).

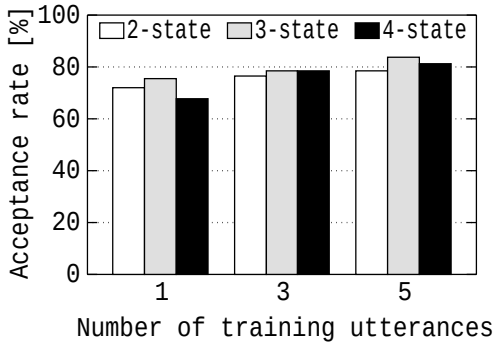

Figure 4: False acceptance rates for synthetic speech against the speaker verification system using 3-state models (speech synthesis system: 2-, 3-, or 4-state models).

3.3. Imposture Using Synthetic Speech

Here we investigated whether the reference speaker verification system can reject synthetic speech generated from the HMM-based speech synthesis system.

The speech synthesis system was trained using B-set. Speech signals were windowed by a 25.6 ms Blackman window with a 5 ms shift, and the cepstral coefficients were calculated by 15-th order mel-cepstral analysis [10]. The feature vector consisted of 16 mel-cepstral coefficients including 0-th coefficient, and their deltas and delta-deltas. It is noted that the feature parameters used in the speech synthesis system were different from the speaker verification system.

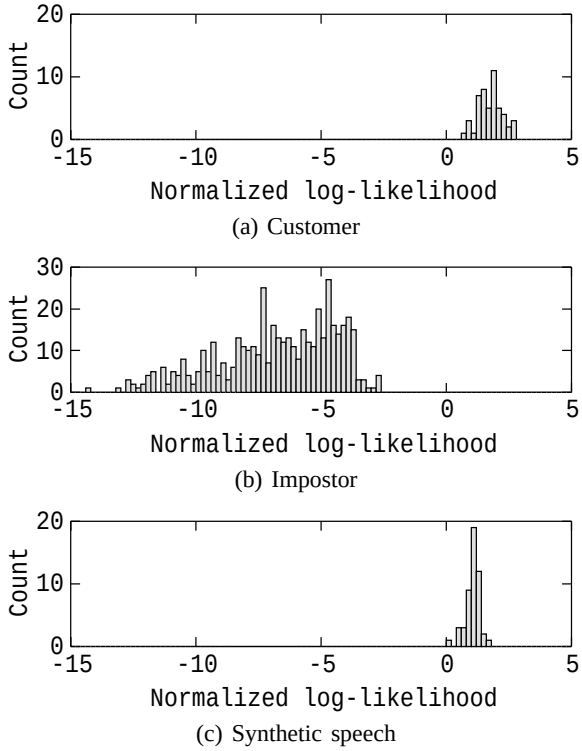

Figure 5: Histograms of normalized log-likelihood scores.

Phoneme models were trained using 1, 3, or 5 sentences uttered by customers of the speaker verification system. Phoneme models were context independent 2-, 3-, or 4-state single-mixture left-to-right models with diagonal covariance matrices. As well as the speaker verification system, SI phoneme models were used instead of untrained SD phoneme models because of insufficient training data. SI models were trained using 500 sentences in B-set uttered by 10 impostors.

In the synthesis procedure, state durations were set to means of state duration densities obtained from training data, and white noise was used as an excitation to the MLSA filter [10] whether speech is voiced or unvoiced.

Figure 3 and *4* show FARs for synthetic speech against the speaker verification systems using 2- and 3-state models, respectively. From these figures, it can be seen that synthetic speech has very high FARs for any combinations of the states of models used in speech synthesis and speaker verification systems. It is noted that FARs for synthetic speech from SI models (i.e., without training data) were 0% for all combinations.

The speaker verification system using 3-state models shows better performance than using 2-state models. However, the difference between these systems is insignificant. Furthermore, the system using 3-state models has a higher FRR for customers.

With regard to the speech synthesis system, the system using 3-state models achieves higher FARs than 2- and 4-state models. Increasing the number of states makes HMMs possible to model speech in detail resulting improvement in speech quality, but too many states cause inaccurate estimate of HMM parameters because of insufficient training data.

Figure 5 shows histograms of normalized log-likelihood for a customer, impostors, and synthetic speech, in which the claimed speaker is m101, the speaker verification system uses

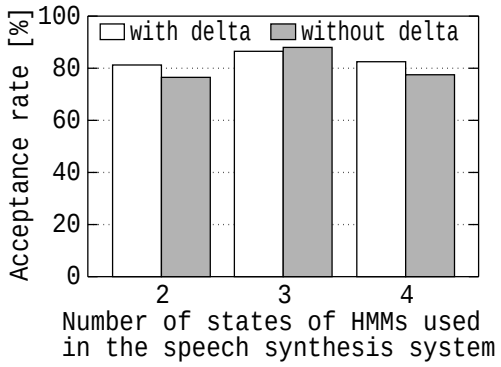

Figure 6: False acceptance rates for synthetic speech with and without delta parameters (speaker verification system: 2-state models).

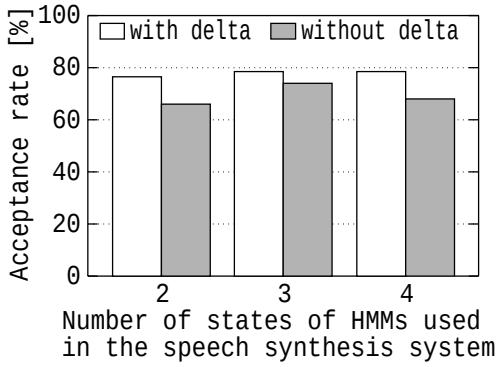

Figure 7: False acceptance rates for synthetic speech with and without delta parameters (speaker verification system: 3-state models).

3-state models, and the speech synthesis system uses 3-state models trained using 5 sentences. From these figures, it can be seen that the distribution of synthetic speech overlaps with the customer's distribution, and it is difficult to discriminate synthetic speech from customer's speech effectively by adjusting the decision threshold.

Finally, we investigated effect of dynamic features on synthetic speech. *Figure 6* and *7* shows FARs for synthetic speech with/without delta parameters. The speech synthesis system was trained using 3 sentences. Without dynamic features, subjective quality of synthetic speech is much lower than with dynamic features [7]. However, it is still sufficient for imposture against the reference speaker verification system.

4. CONCLUSION

In this paper, we have investigated imposture using synthetic speech. In the experiments, conditions we used might be unrealistic. For example, both the speaker verification system and the speech synthesis system used the same phoneme sets and the same phoneme sequences corresponding to key texts. However, from the facts that the false acceptance rates for synthetic speech reached over 70% by training the speech synthesis system with only 1 sentence from each customer, and that quality of synthetic speech will be improved with the advance of technique of speech synthesis, current security of HMM-based speaker verification systems against synthetic speech is inadequate even though these disadvantages are taken into account. To put speaker verification systems into practice, it is required to develop techniques to discriminate synthetic speech from natural speech, for example, utilizing pitch contours or phoneme durations in addition to spectral parameters. It is also necessary to investigate in another conditions, such as different likelihood normalization techniques, and speaker verification systems with different frameworks.

5. ACKNOWLEDGMENT

This work was supported in part by the Ministry of Education, Science, Sports and Culture of Japan, Grant-in-Aid for Scientific Research (B) (2) 1055125, 1998, and Grant-in-Aid for Encouragement of Young Scientists, 11750311, 1999.

6. REFERENCE

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

[2] T. Matsui and S. Furui, "Speaker adaptation of tied-mixture-based phoneme models for text-prompted speaker recognition," Proc. ICASSP-94, pp.125–128, Apr. 1994.

[3] T. Matsui and S. Furui, "Likelihood normalization for speaker verification using a phoneme- and speaker-independent model," Speech Communication, vol.17, no.1–2, pp.109–116, Aug. 1995.

[4] D. Genoud, G. Chollet, "Speech pre-processing against intentional imposture in speaker recognition," Proc. ICSLP-98, pp.105–108, Dec. 1998.

[5] B.L. Pellom and J.H.L. Hansen, "An experimental study of speaker verification sensitivity to computer voice-altered imposters," Proc. ICASSP-99, Mar. 1999.

[6] A. Black and N. Campbell, "Optimizing selection of units from speech database for concatenative synthesis," Proc. EUROSPEECH-95, pp.581–584, Sept. 1995.

[7] T. Masuko, K. Tokuda, T. Kobayashi and S. Imai, "Speech synthesis using HMMs with dynamic features," Proc. ICASSP-96, pp.389–392, May 1996.

[8] T. Masuko, K. Tokuda, T. Kobayashi, and S. Imai, "Voice characteristics conversion for HMM-based speech synthesis system," Proc. ICASSP-97, pp.1611–1614, 1997.

[9] M. Tamura, T. Masuko, K. Tokuda, and T. Kobayashi, "Speaker adaptation for HMM-based speech synthesis system using MLLR," Proc. The Third ESCA/COSCOSDA International Workshop on Speech Synthesis, pp.273–276, Nov. 1998.

[10] T. Fukada, K. Tokuda, T. Kobayashi and S. Imai, "An adaptive algorithm for mel-cepstral analysis of speech," Proc. ICASSP-92, pp.137–140, Mar. 1992.

[11] K. Tokuda, T. Kobayashi and S. Imai, "Speech parameter generation From HMM using dynamic features," Proc. ICASSP-95, pp.660–663, May 1995.

[12] K. Tokuda, T. Masuko, T. Yamada, T. Kobayashi and S. Imai, "An algorithm for speech parameter generation from continuous mixture HMMs with dynamic features," Proc. EUROSPEECH-95, pp.757–760, Sept. 1995.