A LANGUAGE-INDEPENDENT PERSONAL VOICE CONTROLLER WITH EMBEDDED SPEAKER VERIFICATION

Qi Li, Augustine Tsai, and Weon-Goo Kim

Multimedia Communications Research Laboratory
Bell Labs, Lucent Technologies
600 Mountain Avenue, Murray Hill, NJ 07974, USA
{qli,atsai,wgkim}@research.bell-labs.com

ABSTRACT

In this paper, we introduce a personal voice controller for name dialing. As a personal system, all the control commands are trained based on owner-selected phrases in the owner's language. Since the owner's voice characteristics is modeled together with the command voice, the system has embedded speaker verification capability without additional processing. In other words, once the system is trained, the system can only accept the voice commands from the owner. Impostor's command can be rejected although the command may be in the owner's list. The controller can recognize the speaker's command and dial the corresponding phone number automatically. Such a system is useful for wireless handsets and portable communication devices. Its implementation requires much less memory space and computation resource compared to a speaker-independent system. Preliminary experiments showed that the proposed system has the state-of-the-art performances in both speech recognition and speaker verification.

1. INTRODUCTION

As wireless handsets and portable communication devices become more and more popular, a user-friendly voice interface with a certain level of security protection is needed. For example, in a hand-busy and eye-busy situation, it would be a great convenience to the users if they can just utter a name to dial instead of finding the number in a book and dialing it by hand. On the other hand, as a personal device, it would be ideal that the interface only accepts the commands from the owner.

The concept of a personal voice controller for voice dialing is shown in Fig. 1. There are two kinds of sessions, enrollment and test. In an enrollment session, a user needs to enroll their voice commands associated with each control action. For name dialing, a user needs to repeat a name for two or three times, then input a telephone number associated with the name. New commands or names can be added to the system while the system is in use. In a test (or application) session, the user just needs to utter a command or a name. The the system will recognize the name, verify the speaker, and dial the corresponding number directly. For a personal device, such as cellular phones, or a portable communication device, the identity claim is not necessary. For telephone, the identity can be obtained from caller ID or from user's input.

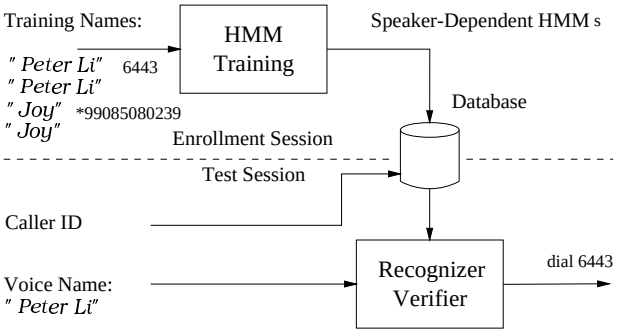

Figure 1: The concept of personal voice dialing system

While constructing the HMM models, we applied the same technique as in speaker verification (SV) [1, 2]. Therefore, the system can verify the user's voice and only execute the owner's commands. An imposter can be rejected although he or she may utter a command in the owner's list. Since speaker verification is conducted within voice recognition, there is no need for an additional effort from the user and from the system to do speaker verification. Also, since we applied an endpoint detection technique which doesn't need lexicon information, the commands can be in any language.

2. FRONT-END PROCESSOR

The front-end processor of the proposed system is shown in Fig. 2. It includes a LPC cepstral extraction, a recently developed endpoint detection algorithm [3], and cepstral mean subtraction (CMS).

The new endpoint detection algorithm [3] can provide accurate endpoint detection. It needs neither speaker independent HMM's nor the lexicon information, therefore, the system can be language-independent and users can select their commends in any language. From implementation point of view, the proposed system needs much less memory space and CPU time than the approach using phone/silence HMM's and Viterbi decoding.

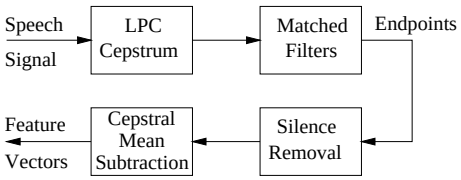

Figure 2: The front-end processor

After determining the endpoints, silence and breath signals are removed from the utterance, and CMS is performed on the speech data. Given the original observation of $\mathcal{O}$, after the endpoint detection and CMS, the feature set becomes $\mathbf{O}$ which is a subset of $\mathcal{O}$, i.e. $\mathbf{O} \subset \mathcal{O}$.

3. PERSONAL VOICE DIALER

In an enrollment session, the system collects two or three utterances for each name, then constructs one HMM for one name by the segmental K-mean algorithm [4]. It is the same as the one-word model which has been used in speaker verification [1, 2]. The models are saved with associated telephone numbers in the owner's account.

The test session is shown in Fig. 3, given an observation of utterance $\mathbf{O}$, the average frame log-likelihood score, L_i, is computed for each of the models by forced decoding as

$$L_i(\mathbf{O}, \lambda_i) = \frac{1}{N_i} \log P(\mathbf{O}|\lambda_i), \tag{1}$$

where N_i is the total number of feature vectors for the ith name, $P(\mathbf{O}|\lambda_i)$ is the accumulative likelihood score of the ith name computed by forced decoding. By comparing all the scores, the kth name is selected as the most probable candidate,

$$k = \arg \max_{1 \le i \le N} \{L_i(\mathbf{O}, \lambda_i)\}. \tag{2}$$

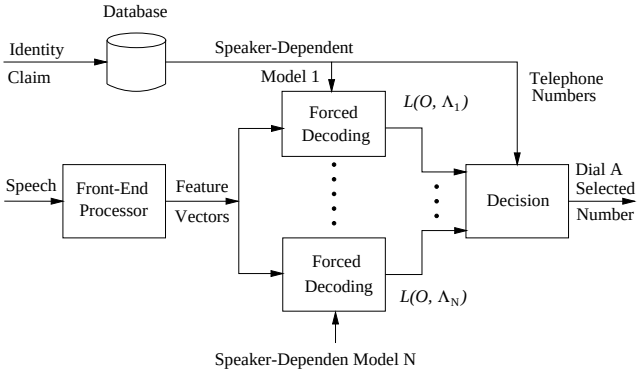

Figure 3: Personal voice dialing system

Since the spoken name may be out of the vocabulary, the selected score, L_k, is compared with a threshold, θ, to make a decision on either acceptance or rejection,

$$\begin{cases} Acceptance: & L_k \ge \theta; \\ Rejection: & L_k < \theta. \end{cases} \tag{3}$$

Actually, the threshold is also needed for speaker verification when we discuss in the next session. If a name is accepted, it means that the name is in the vocabulary and the utterance is from the owner. The phone number associated with the name is then dialed automatically.

Since only three training utterances per name are used in this system, it is difficult to estimate exact state duration distribution to improve the recognition performance. Therefore, a post-processor is presented. It modifies the above log-likelihood score as

$$\hat{L}_i(\mathbf{O}, \lambda_i) = \frac{1}{N_i} \log P(\mathbf{O}|\lambda_i) + \alpha \frac{1}{S} \sum_{j=1}^{S} \frac{1}{T_j} \log P(\mathbf{O}_j|\lambda_i), \tag{4}$$

where, α is a weighting parameter, and S is the total number of states. After forced decoding, the given observation $\mathbf{O}$ is segmented into states, $\mathbf{O} = \{\mathbf{O}_1, \mathbf{O}_2, ...\mathbf{O}_S\}$, where $\mathbf{O}_j$ is the feature vectors corresponding to state j. T_j is the total number of feature vectors of $\mathbf{O}_j$.

4. SPEAKER VERIFICATION

Since the name dialing system is developed using the same technology as we have used in SV. The system embeds SV capability without applying any additional processing. For the above system, since it does not need any text or lexicon information, the system is a language-independent system for both voice dialing and speaker verification.

If an application needs a higher level of security, i.e. lower equal error rate (EER), a background model

can be applied for speaker verification, either with lexicon information [5] or without lexicon information [6]. When the lexicon information is available, a background model can be constructed by concatenating phone models corresponding with the given lexicon information [5, 1]. Based on the concept of Neyman-Pearson lemma, we have a likelihood ratio test for decision,

$$L_R(\mathbf{O}; \lambda_k; \lambda_b) = L(\mathbf{O}, \lambda_k) - L(\mathbf{O}, \lambda_b), \qquad (5)$$

where $\mathbf{O}$ is the observation sequence over the whole phrase, λ_k and λ_b are the name and background models respectively, and L is defined as in Eq. (1). The background model is a set of HMM's for phones. A finial decision on rejection or acceptance is made based on the L_R score with the threshold θ.

5. FEATURES AND DATABASE

The feature vector is composed of 12 cepstrum and 12 delta cepstrum coefficients. The cepstrum is derived from a 10th order LPC analysis over a 30 ms window. The feature vectors are updated at 10 ms intervals.

The experimental database consists of 38 speakers, 18 male and 20 female [7]. Each speaker has 15 name entries. The database evaluation is on a worst-case situation where all the names are "Call ...", e.g. "Call office", "Call Audix", "Call home", "Call mom", etc. This means that about 1/2 to 1/3 of the contents are the same. Many names are very short in about 1 second, which makes the recognition even more difficult.

The HMM models for names are left-to-right HMM's. Due to the limitation on the model size, the number of states are estimated based on 10 frames per sate. There are 4 Gaussian components associated with each state. Also, due to unreliable variance estimates from limited amount of training data, a global variance estimate is used as a common variance to all Gaussian components (e.g. [1]) in the name models.

6. EXPERIMENTAL RESULTS

We have proposed to use the common model and decoder for both name dialing and SV in one system. In order to verify that the proposed approach can give us the state-of-the-art results, we evaluate the performances of each application separately. Due to the limit space of this paper and the complexity in further evaluation, the combined system evaluation and out-of-vocabulary evaluation will be reported separately.

6.1. Language-Independent System

Three utterances of each name recorded in one session were used to train a name model. In testing, 10 name utterances from each speaker collected from 5 different sessions were used to evaluate the recognition performance. Each speaker has 15×10 names in testing. The experimental results for voice dialing using the database are listed in Table 1. These are the average recognition error rates. The second column in the table is the error rates using the likelihood score as in Eq. (1). The error rate for male and female speakers are 3.0% and 3.5% respectively. The average error rate is 3.3%. The third column is the error rates using the duration weighted scores as in Eq. (4). The error rate for male and female speakers are 2.5% and 3.0% respectively. The average error rate is 2.8%. The relation between the weighting parameter α and error rates was plotted in Fig. 4.

Table 1: **Language-Independent Voice Dialing**

Scores	L	$\hat{L}$
18 Male Speakers	3.0%	2.5%
20 Female Speakers	3.5%	3.0%
Average	3.3%	2.8%

A common pass-phrase for all speakers was used to evaluate the SV performance of the system. The pass-phrase is "Call Janice at her office phone." As an embedded SV system, no additional computation is needed. The model was trained using three utterances and the likelihood score was computed as in Eq. (1). The evaluation was separated into two groups, male and female, and tested within the same gender. For each male speaker, ten utterances from the true speaker collected from 5 different sessions and $17 \times 5 = 85$ utterances from impostors were used for testing. For each female speaker, there are 10 utterances from the true speaker while the impostor's utterances were $19 \times 5 = 95$. The equal-error rate (EER) is listed in Table 2. For male and female speakers, the EER's are 4.2% and 5.6% respectively when individual thresholds were used. The average individual EER over 38 speakers is 4.9% without using any background model and phone model. Therefore, this is a language-independent SV system.

Table 2: **Average Equal-Error Rates of Speaker Verification Using 3 Utterances for Enrollments**

Speakers	Without Background Models (Language-Independent)
18 Male Speakers	4.2%
20 Female Speakers	5.6%
Average	4.9%

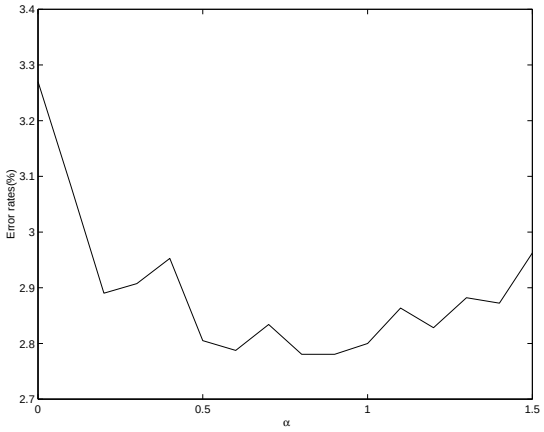

Figure 4: The relation between α and error rates.

Table 3: **Average Equal-Error Rates of Speaker Verification Using 5 Utterances for Enrollments**

Speakers	W/O BK Models (Language-Independent)	With BK Models (Language-Dependent)
18 Males	3.6%	2.0%
20 Females	4.4%	3.5%
Average	4.0%	2.8%

6.2. Further Comparison on SV

Since our previous study on SV was using 5 enrollment utterances and a background model [1, 2]. For comparison, we conducted the test under the same condition. Since the background model is concatenated phone HMM's, in this case, the system is a language-dependent system. The background model was trained on a telephone speech database from different speakers and texts. Each phone HMM has 3 states with 32 Gaussian components associated with each state. The evaluation results are shown in Table 3. The average individual EER is 4.0% and 2.8% with and without the background model respectively. The accuracy is in the same level as the state-of-the-art speaker verification system (e.g. [1, 2]). The system performance for both name dialing and SV can be further improved if model adaptation is applied. Also, the name dialing performance can be improved if more states (i.e. large model size) are allowed.

7. CONCLUSIONS

A personal voice controller for voice dialing with embedded speaker verification was introduced in this paper. The system uses the same recognition/verification procedure for both name dialing and SV, therefore, no additional computation is needed in the system for SV except the application requires a higher level of security. The proposed system is also language-independent for both name dialing and SV. In a difficult database evaluation, the name dialing accuracy is in the high nineties. In a real implementation, the performance can be even better since the system can avoid enrolling the name which is too close to the existing name in the system. The embedded SV performance was evaluated in the same method as we evaluate a real SV system. The performance is in the same level as the state-of-the-art systems designed for SV.

8. ACKNOWLEDGMENT

The authors wish to thank Mark (Rusty) Ransford, Roy Stevens, Bob Cooper, and Chin-Hui Lee for useful discussions, and Rafid Sukkar for providing the database.

9. REFERENCES

[1] S. Parthasarathy and A. E. Rosenberg, "General phrase speaker verification using sub-word background models and likelihood-ratio scoring," in *Proceedings of ICSLP-96*, (Philadelphia), October 1996.

[2] Q. Li, S. Parthasarathy, and A. E. Rosenberg, "A fast algorithm for stochastic matching with application to robust speaker verification," in *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing*, (Munich), pp. 1543–1547, April 1997.

[3] Q. Li and A. Tsai, "A matched filter approach to endpoint detection for robust speaker verification," in *Submitted to the Workshop of Automatic Identification*, (Summit, NJ), Oct. 1999.

[4] L. R. Rabiner, J. G. Wilpon, and B.-H. Juang, "A segmental k-means training procedure for connected word recognition," *AT&T Technical Journal*, vol. 65, pp. 21–31, May/June 1986.

[5] A. E. Rosenberg and S. Parthasarathy, "Speaker background models for connected digit password speaker verification," in *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing*, (Atlanta), pp. 81–84, May 1996.

[6] O. Siohan, C.-H. Lee, A. C. Surendran, and Q. Li, "Background model design for flexible and protable speaker verification systems," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing*, (Phoenix), pp. 825–828, March 1999.

[7] R. A. Sukkar and C.-H. Lee, "Vocabulary independent discriminative utterance verification for non-keyword rejection in subword based speech recognition," *IEEE Trans. Speech and Audio Process.*, vol. 4, pp. 420–429, November 1996.