SPEAKER AUTHENTICATION

Qi Li and Biing-Hwang Juang

Present by : 陳子和

Content:

- Introduction:
- Speaker Authentication :
    - Speaker Recognition and Verification
    - Verbal Information Verification
- Pattern Recognition in Speaker Authentication
    - Bayesian Decision Theory
    - Stochastic Models for Stationary Process
    - Stochastic Models for Non-Stationary Process
    - Statistical Verification
- Speaker Authentication System
    - Speaker Verification System
    - VIV System
    - Combining SV and VIV System
- Conclusion

Introduction:

- To ensure the security of a proper access to private information, passwords or personal identification numbers (PIN) have been used. To further enhance the level of security,biometric features such as signature, fingerprint, hand shape, eye iris, and voice have been considered.
- Speaker Authenticating

    1.Speaker Recognition(by characteristics)

        speaker verification (SV)

        speaker identification(SID)

    2.verbal information verification (VIV)(by verbal content)

Speaker Authentication

SV

SR

SID

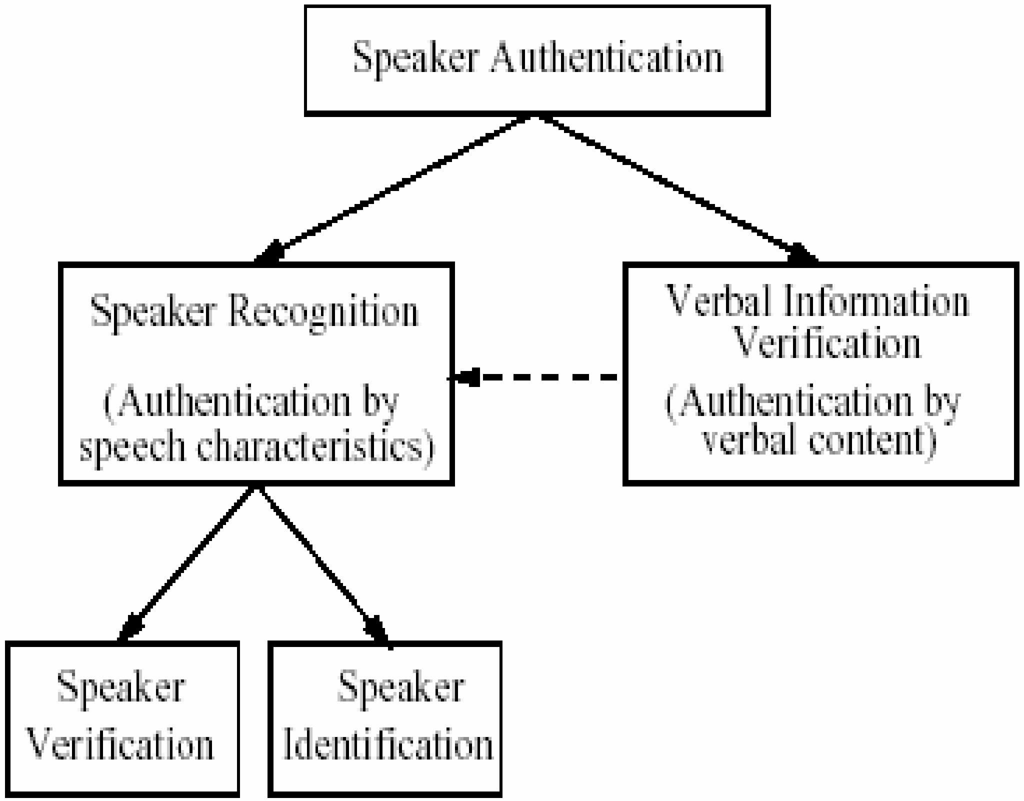

Figure 1.1 Speaker authentication approaches

Multiple-choice classification problem

Speaker Recognition and Verification

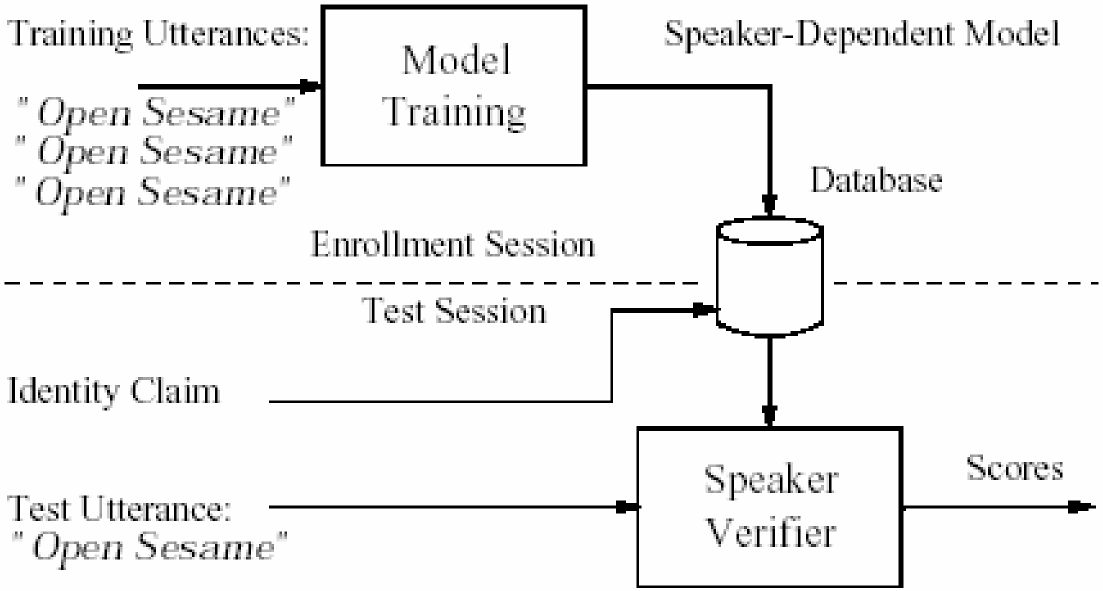

Figure 1.2 A speaker verification system

A typical SV system: enrollment and test sessions.

Speaker Recognition and Verification (cont.)

Text-dependent or text-constrained SV systems

Fixed pass-phrase system

the spoken digit string is first recognized by an ASR and the standard verification procedure then follows.

Text-prompted system (A safety concern)

the system prompts the user to utter a randomized sequence of words in the vocabulary.

Verbal Information Verification

- mismatch significantly aspects the SV performance
- Enrollment is an inconvenience to the user
- A safety concern

Verbal Information Verication (cont.)

Tele banking system

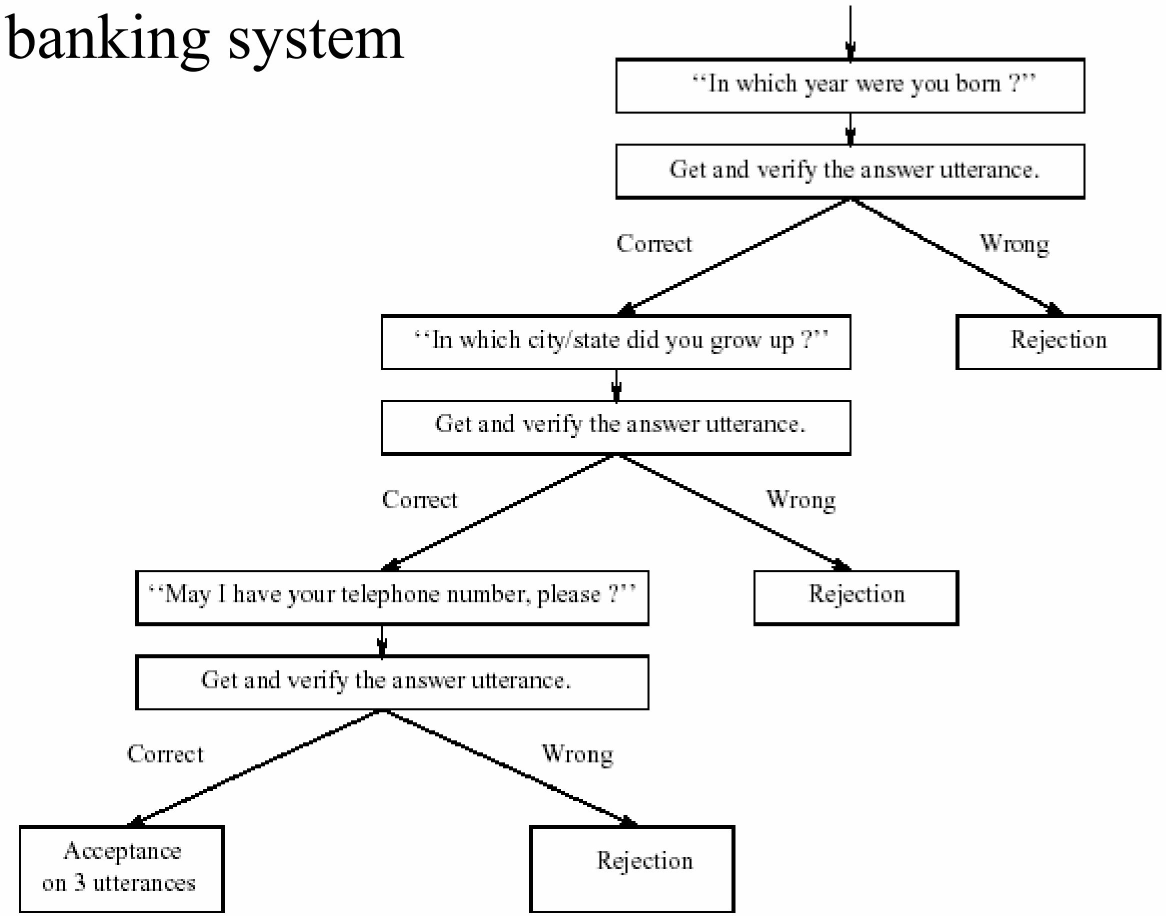

Figure 1.3 An example of verbal information verification by asking sequential questions. (Similar sequential tests can also be applied in speaker verification and other biometric or multi-modality verification.)

Pattern Recognition in Speaker Authentication

Bayesian Decision Theory

the probability of being class Ci given o P(Ci|o) is the posterior probability,P(o|Ci) is the conditional probability, P(Ci) is prior probability:

$$P(C_i|\mathbf{o}) = \frac{p(\mathbf{o}|C_i)P(C_i)}{p(\mathbf{o})}$$

$$p(\mathbf{o}) = \sum_{j=1}^{M} p(\mathbf{o}|C_j)P(C_j)$$

Let L(αi|Cj) be the loss function describing the loss incurred for taking action when the true class is Cj. The expected risk associated with taking action αi is

$$R(\alpha_i|\mathbf{o}) = \sum_{j=1}^{M} \mathcal{L}(\alpha_i|C_j)P(C_j|\mathbf{o}).$$

$$\mathcal{L}(\alpha_i|C_j) = \begin{cases} 0 & i = j \\ 1 & i \neq j. \end{cases} \qquad i, j = 1, \ldots, M$$

$$\begin{aligned} R(\alpha_i|\mathbf{o}) &= \sum_{j=1}^{M} \mathcal{L}(\alpha_i|C_j)P(C_j|\mathbf{o}) \\ &= \sum_{j \neq i} P(C_j|o) = 1 - P(C_i|\mathbf{o}). \end{aligned}$$

$$\alpha_k = \arg\max_{1 \le i \le M} P(C_i|\mathbf{o}).$$

$$\alpha_k = \arg\max_{1 \le i \le M} p(\mathbf{o}|C_i)P(C_i).$$

$$P(C_i|\mathbf{O}) = \prod_{t=1}^{T} P(C_i|\mathbf{o}_t).$$

$$\alpha_k = \arg\max_{1 \le i \le M} \prod_{t=1}^{T} p(\mathbf{o}_t|C_i)P(C_i).$$

$$\alpha_k = \arg\max_{1 \le i \le M} \sum_{t=1}^{T} \log p(\mathbf{o}_t|C_i)P(C_i).$$

Stochastic Models for Stationary Process

Gaussian mixture model (GMM):

$$p(\mathbf{o}_t|C_j) = p(\mathbf{o}_t|\lambda_j) = \sum_{i=1}^{I} c_i \mathcal{N}(\mathbf{o}_t; \mu_i, \Sigma_i),$$

$$\mathcal{N}(\mathbf{o}_t; \mu_i, R_i) = \frac{1}{(2\pi)^{d/2}|\Sigma_i|^{1/2}} \exp\left\{-\frac{1}{2}(\mathbf{o}_t - \mu_i)^{\mathrm{T}}\Sigma_i^{-1}(\mathbf{o}_t - \mu_i)\right\},$$

Stochastic Models for Stationary Process (cont.)

EM Algorithm

Given a sequence of feature vectors, the GMM parameters can be estimated iteratively by an expectation-maximization (EM) algorithm [7]. The EM algorithm is based on an auxiliary function $Q(\cdot)$. When $Q(\cdot)$ is increases, it guarantees that $p(\mathbf{O}|\lambda)$ will be increase. The EM procedure is as follows:

1. initialize a model λ^k;
2. in an E-step, evaluate the auxiliary function $Q(\lambda^k, \lambda^{k-1})$;
3. in an M-step, optimize a new model λ^{k+1}, such that $Q(\lambda^{k+1}, \lambda^k) > Q(\lambda^k, \lambda^{k-1})$. This implies that $p(\mathbf{O}|\lambda^{k+1}) \geq p(\mathbf{O}|\lambda^k)$; and
4. repeat the above E and M steps until $Q(\lambda^{k+1}, \lambda^k) - Q(\lambda^k, \lambda^{k-1}) \leq \epsilon$, were $\epsilon > 0$ is a pre-selected small number.

Stochastic Models for Stationary Process (cont.)

EM Algorithm

$$\hat{c}_i = \frac{1}{T}\sum_{t=1}^{T} p(i|\mathbf{o}_t, \lambda) \tag{1.14}$$

$$\hat{\mu}_i = \frac{\sum_{t=1}^{T} p(i|\mathbf{o}_t, \lambda)\mathbf{o}_t}{\sum_{t=1}^{T} p(i|\mathbf{o}_t, \lambda)} \tag{1.15}$$

$$\hat{\Sigma}_i = \frac{\sum_{t=1}^{T} p(i|\mathbf{o}_t, \lambda)(\mathbf{o}_t - \hat{\mu}_i)(\mathbf{o}_t - \hat{\mu}_i)^{\mathrm{T}}}{\sum_{t=1}^{T} p(i|\mathbf{o}_t, \lambda)} \tag{1.16}$$

where

$$p(i|\mathbf{o}_t, \lambda) = \frac{p(\mathbf{o}_t|\lambda)c_i}{\sum_{j=1}^{J} p(\mathbf{o}_t|\lambda)c_j}. \tag{1.17}$$

Stochastic Models for Stationary Process (cont.)

When Testing: assume **the prior is the same** for all speak
Take action

$$\alpha_k = \arg\max_{1 \le i \le M} \sum_{t=1}^{T} \log p(\mathbf{o}_t|\lambda_i).$$

Stochastic Models for Non-Stationary Process

The stationary process ignored the temporal information. In other applications, such as speaker verification, the temporal information is necessary in making decisions.

A more powerful model, Hidden Markov Model (HMM) is then applied to characterize both the temporal structure and the corresponding statistical variations along the trajectory of an utterance.

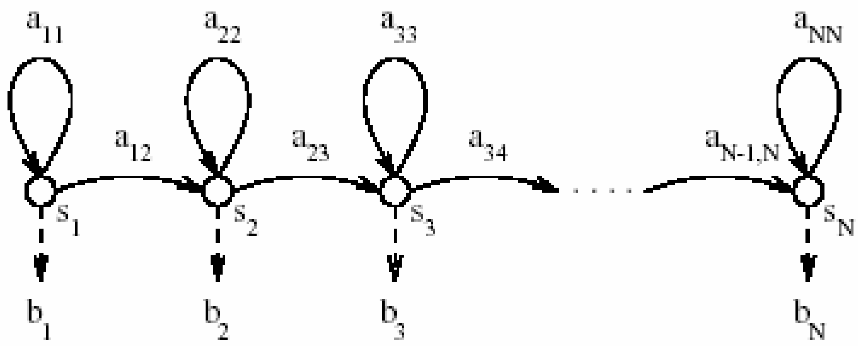

Figure 1.4 Left-to-right hidden Markov model.

Stochastic Models for Non-Stationary Process(cont.)

Speech Segmentation

Viterbi Algorithm

$$P(\mathbf{O}, s_{max}|\lambda) = \max_{\{s_t\}} \left\{ \prod_{t=1}^{T} a_{s_t s_{t+1}} b_{s_t}(\mathbf{o}_t) \right\}$$

Statistical Verification

$$R(\alpha_1|\mathbf{o}) = \mathcal{L}(\alpha_1|C_1)P(C_1|\mathbf{o}) + \mathcal{L}(\alpha_1|C_2)P(C_2|\mathbf{o}) \quad (1.21)$$
$$R(\alpha_2|\mathbf{o}) = \mathcal{L}(\alpha_2|C_1)P(C_1|\mathbf{o}) + \mathcal{L}(\alpha_2|C_2)P(C_2|\mathbf{o}) \quad (1.22)$$

The action α_1 corresponds to decide that the true class is class C_1 if

$$R(\alpha_1|\mathbf{o}) < R(\alpha_2|\mathbf{o}). \quad (1.23)$$

Bring (1.21) and (1.22) into(1.23) and rearranging the terms, we take action α_1 if:

$$\frac{P(C_1|\mathbf{o})}{P(C_2|\mathbf{o})} > \frac{\mathcal{L}(\alpha_1|C_2) - \mathcal{L}(\alpha_2|C_2)}{\mathcal{L}(\alpha_2|C_1) - \mathcal{L}(\alpha_1|C_1)} = \mathcal{T}_1 \quad (1.24)$$

where $\mathcal{T}_1 > 1$ is a threshold. Furthermore, if we apply the Bayes formula to replace the posterior probabilities by prior probabilities, we have

$$\frac{p(\mathbf{o}|C_1)}{p(\mathbf{o}|C_2)} > \mathcal{T}_1 \frac{P(C_2)}{P(C_1)} = \mathcal{T}_2. \quad (1.25)$$

For a sequence of observation $\mathbf{O} = \{\mathbf{o}_i\}_{i=1}^{T}$, if we assume that the distributions of the observations are independent, we have the likelihood-ratio:

$$r(\mathbf{O}) = \frac{\prod_{t=1}^{T} p(\mathbf{o}_t|C_1)}{\prod_{t=1}^{T} p(\mathbf{o}_t|C_2)} = \frac{P(\mathbf{O}|C_1)}{P(\mathbf{O}|C_2)} > \mathcal{T}_3. \quad (1.26)$$

Statistical Verification (cont.)

In practice, we compute log-likelihood ratio for verification:

$$\mathcal{R}(\mathrm{O}) = \log P(\mathrm{O}|C_1) - \log P(\mathrm{O}|C_2). \tag{1.27}$$

A decision is made as:

$$\begin{cases} \text{Acceptance:} & \mathcal{R}(\mathrm{O}) \geq \mathcal{T}; \\ \text{Rejection:} & \mathcal{R}(\mathrm{O}) < \mathcal{T}, \end{cases} \tag{1.28}$$

where $\mathcal{T}$ is a threshold value, which can be determined theoretically or experimentally.

Statistical Verification (cont.)

False rejection : rejecting the hypothesis when it is actually true.

False acceptance: accepting it when it is actually false.

Equal error rate: the error rate when the operating point is so chosen as to achieve equal error probabilities for the two types of error.

Speaker Authentication System

Speaker Verification

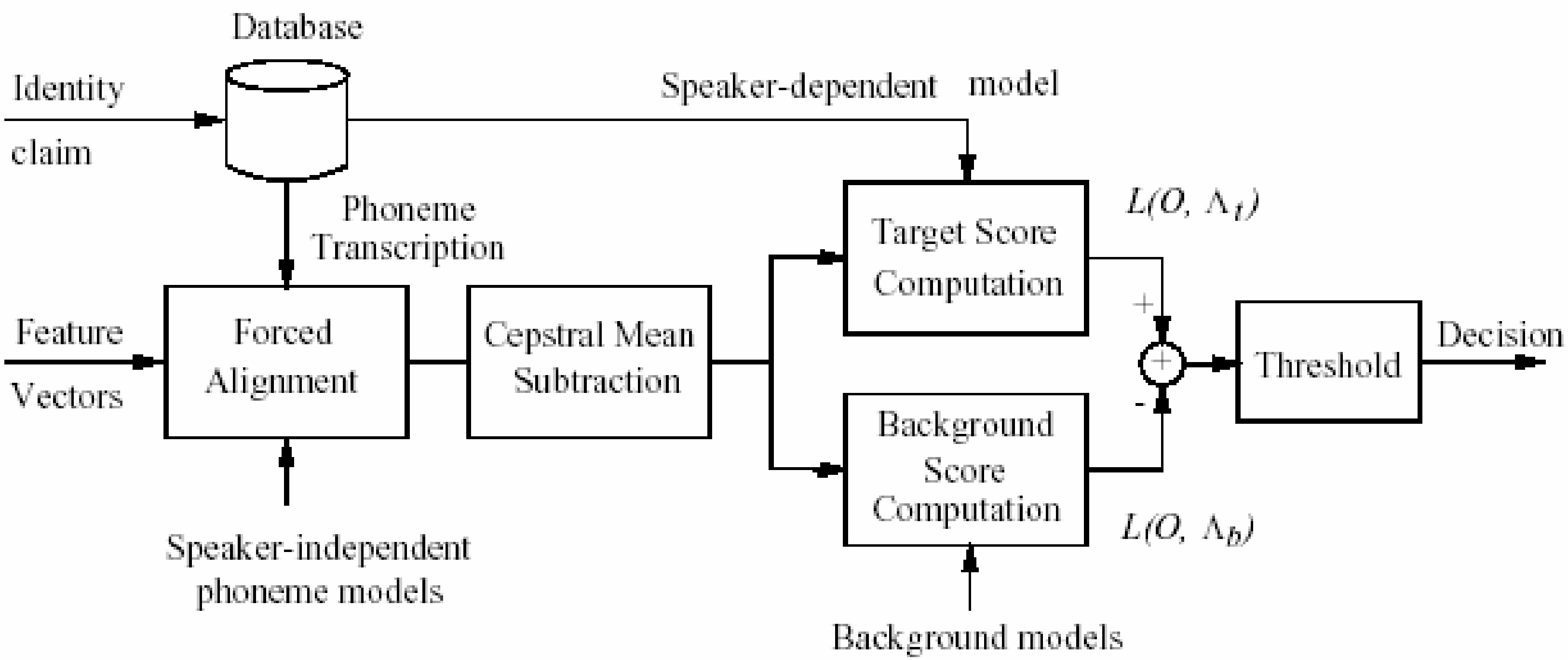

Figure 1.5 A fixed-phrase speaker verification system

Speaker Authentication System (cont.)

VIV : with UV

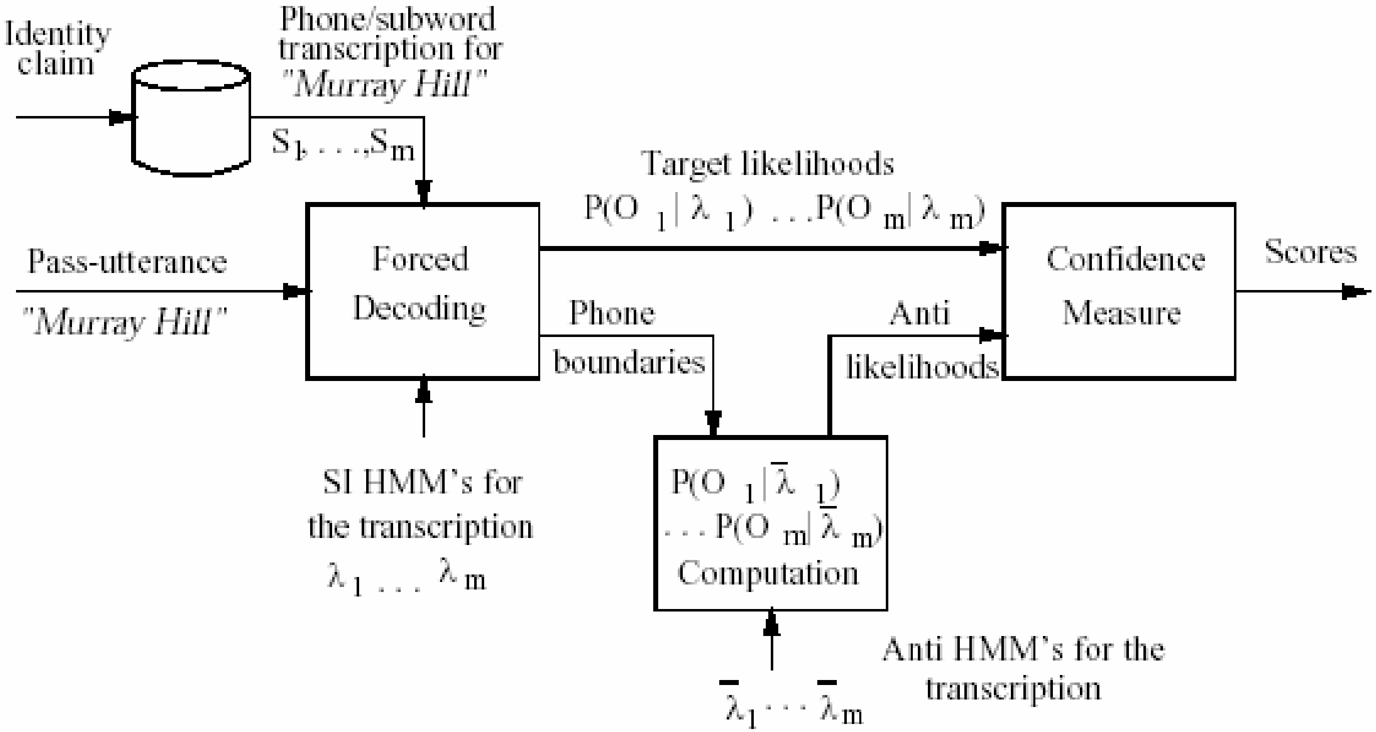

Figure 1.6 Utterance verification in VIV

Speaker Authentication System (cont.)

Utterance Segmentation

Subword Hypothesis Testing

Confidence Measure Calculation

Sequential Utterance Verification

Speaker Authentication System (cont.)

Utterance Segmentation

$$P(\mathbf{O}|\mathbf{S}) = \max_{T_1,T_2,\ldots,T_N} P(O_1^{T_1}|S_1)P(O_{T_1+1}^{T_2}|S_2)\ldots P(O_{T_{N-1}+1}^{T_N}|S_N), \qquad (1.32)$$

where

$$\mathbf{O} = \{\mathbf{O}_1, \mathbf{O}_2, \ldots, \mathbf{O}_N\} = \{O_1^{T_1}, O_{T_1+1}^{T_2}, \ldots, O_{T_{N-1}+1}^{T_N}\}, \qquad (1.33)$$

Speaker Authentication System (cont.)

Subword Hypothesis Testing

$$r(\mathbf{O}_n) = \frac{P(\mathbf{O}_n|H_0)}{P(\mathbf{O}_n|H_1)} = \frac{P(\mathbf{O}_n|\lambda_n)}{P(\mathbf{O}_n|\bar{\lambda}_n)}, \tag{1.34}$$

$$R(\mathbf{O}_n) = \log P(\mathbf{O}_n|\lambda_n) - \log P(\mathbf{O}_n|\bar{\lambda}_n). \tag{1.35}$$

$$R_n = \frac{1}{T_n}\left[\log P(\mathbf{O}_n|\lambda_n) - \log P(\mathbf{O}_n|\bar{\lambda}_n)\right], \tag{1.36}$$

$$\begin{cases} \text{Acceptance:} & R_n \geq \mathcal{T}_n; \\ \text{Rejection:} & R_n < \mathcal{T}_n; \end{cases} \tag{1.37}$$

Speaker Authentication System (cont.)

Confidence Measure Calculation

$$\mathcal{M}(\mathbf{O}) = \mathcal{F}(R_1, R_2, ..., R_N), \qquad (1.38)$$

$$M_1 = \frac{1}{L}\sum_{n=1}^{N} l_n R_n, \qquad (1.39)$$

$$M_2 = \frac{1}{N}\sum_{n=1}^{N} R_n, \qquad (1.40)$$

$$C_n = \frac{\log P(\mathbf{O}_n|\lambda_n) - \log P(\mathbf{O}_n|\bar{\lambda}_n)}{\log P(\mathbf{O}_n|\lambda_n)} \qquad (1.41)$$

$$M = \frac{1}{N}\sum_{n=1}^{N} f(C_n), \qquad (1.42)$$

where

$$f(C_n) = \begin{cases} 1, & \text{if } C_n \geq \theta; \\ 0, & \text{otherwise}, \end{cases} \qquad (1.43)$$

Speaker Authentication System (cont.)

Sequential Utterance Verification

$$\mathcal{H}_0 = \bigcap_{i=1}^{J} H_0(i), \tag{1.44}$$

$$\mathcal{H}_1 = \bigcup_{i=1}^{J} H_1(i), \tag{1.45}$$

$$\begin{cases} \text{Acceptance:} & M(i) \geq \mathcal{T}(i); \\ \text{Rejection:} & M(i) < \mathcal{T}(i); \end{cases} \tag{1.46}$$

$$\varepsilon_r(i) = P(\ M(i) \in \mathcal{R}_1(i) \mid H_0(i)\ ), \tag{1.47}$$

and

$$\varepsilon_a(i) = P(\ M(i) \in \mathcal{R}_0(i) \mid H_1(i)\ ), \tag{1.48}$$

respectively. Furthermore, the FR error on J utterances can be evaluated as

$$\begin{aligned} E_r(J) &= P(\ \bigcup_{i=1}^{J}\{M(i) \in \mathcal{R}_1(i)\} \mid \mathcal{H}_0\ ), \\ &= 1 - \prod_{i=1}^{J}(1 - \varepsilon_r(i)), \end{aligned} \tag{1.49}$$

and the FA error on J utterances is

$$\begin{aligned} E_a(J) &= P(\ \bigcap_{i=1}^{J}\{M(i) \in \mathcal{R}_0(i)\} \mid \mathcal{H}_1\ ), \\ &= \prod_{i=1}^{J}\varepsilon_a(i). \end{aligned} \tag{1.50}$$

Speaker Authentication System (cont.)

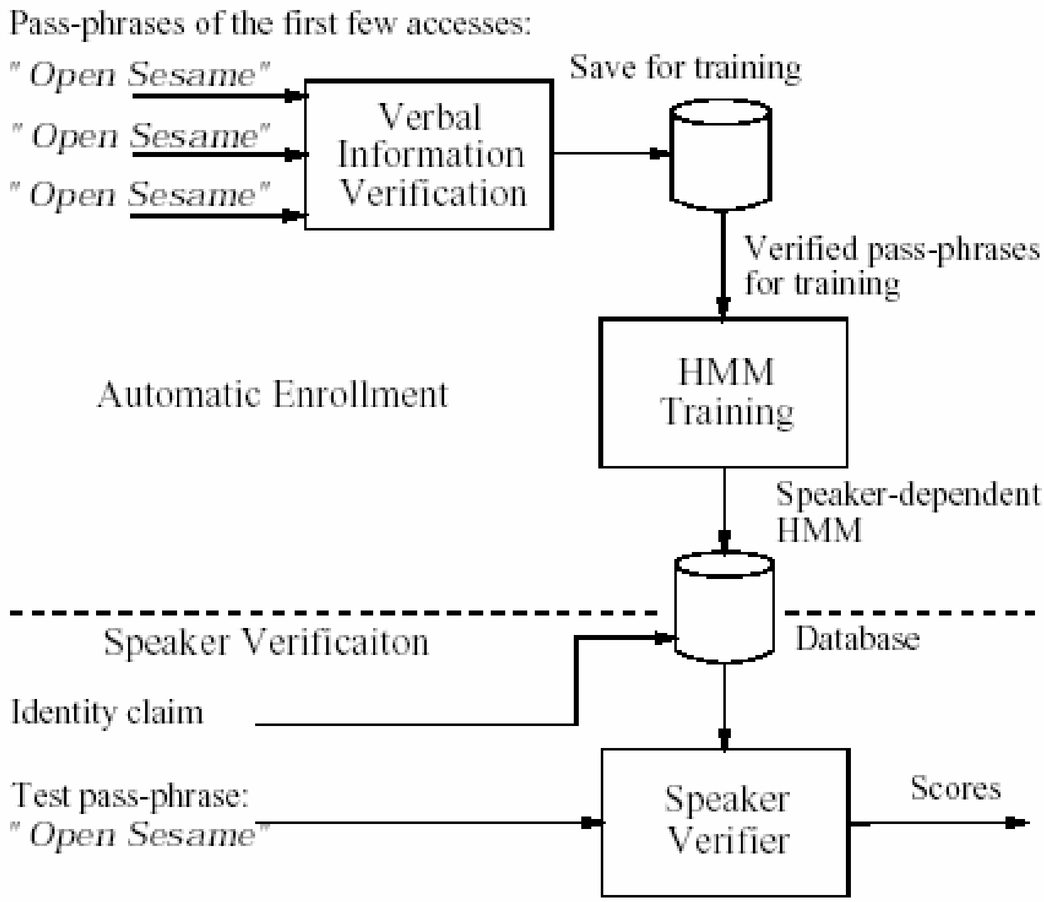

Figure 1.7 The proposed system by combining VIV with speaker verification

Speaker Authentication System (cont.)

Experimental Results

Figure 1.7 Experimental Results without Adaptation in Average Equal-Error Rates

Algorithms	Individual Thresholds	Pooled Thresholds
SV (Baseline)	3.03 %	4.96 %
VIV+SV(proposed)	1.59 %	2.89 %

Figure 1.7 Experimental Results with Adaptation in Average Equal-Error Rates

Algorithms	Individual Thresholds	Pooled Thresholds
SV (Baseline)	2.15 %	3.12 %
VIV+SV(proposed)	1.20 %	1.83 %

Conclusion

Depend on Bayesian decision theory and hypothesis testing, the hypothesis testing may be conducted at phrase, word, phoneme, or subword levels.

On extension to the Bayesian theory to authentication is the sequential verification procedure, which can also be applied to speaker verification to achieve even lower equal error rates.

Currently, the fixed phrase SV system is more attractive to real applications due to its good performances. it is easy to remember and convenient to use.

since VIV is to verify the verbal content instead of the voice characteristics, it is users' responsibility to protect their personal information from impostors

To improve the user convenience and system performance, the VIV and SV is combined to construct a convenient speaker authentication system.

The combined system is convenient to users since they can start to use the system without going through a formal enrollment session and waiting for model training. On the other hand, since the training data could be collected from different channels in different VIV sessions, the acoustic mismatch problem is mitigated, potentially leading to a better system performance in test sessions.

The SD HMM's can be updated to cover different acoustic environments while the system is in use to further improve the system performance.

VIV can also be used to ensure training data for SV.