A CONTEXT-DEPENDENT APPROACH FOR SPEAKER VERIFICATION USING SEQUENTIAL DECISION

Hideki Noda† Katsuya Harada† Eiji Kawaguchi† Hidefumi Sawai‡

†Kyushu Institute of Technology, Kitakyushu, 804-8550 Japan
‡Communications Research Laboratory, Kobe, 651-2401 Japan

ABSTRACT

This paper is concerned about speaker verification (SV) using the sequential probability ratio test (SPRT). In the SPRT input samples are usually assumed to be i.i.d. samples from a probability density function because an on-line probability computation is required. Feature vectors used in speech processing obviously do not satisfy the assumption and therefore the correlation between successive feature vectors has not been considered in conventional SV using the SPRT. The correlation can be modeled by the hidden Markov model (HMM) but unfortunately the HMM can not be directly applied to the SPRT because of statistical dependence of input samples. This paper proposes a method of HMM probability computation using the mean field approximation to resolve this problem, where the probability of whole input samples is nominally represented as the product of probability of each sample as if input samples were independent each other.

1 INTRODUCTION

This paper proposes an improved method of speaker verification (SV) using the statistical sequential decision called "sequential probability ratio test (SPRT)"[1]. In the SPRT a desired performance level (error rate) can be fixed and the number of samples needed for the decision is allowed to vary since further samples are required until the desired level of confidence is achieved. SV using the SPRT has at least three advantages that other SV methods do not have[2],[3]: (1) in principle verification error rate can be controlled, (2) the number of input samples for a given level of performance is less than any other method and (3) speaker-dependent adaptive processing is naturally realized.

In the SPRT making the decision procedure tractable, input samples are usually assumed to be i.i.d. (independent and identically distributed) samples from a probability density function (pdf), even though they are dependent. Feature vectors extracted every around 10 msec from speech wave are used as input samples for SV. Obviously there is a large amount of correlation between successive feature vectors but so far they are treated as independent each other in SV using the SPRT. This paper proposes a method which can consider the correlation in the SPRT-based SV.

The correlation between successive feature vectors, i.e., the context-dependence of input samples can be modeled by the well-known hidden Markov model (HMM). However the HMM can not be directly applied to the SPRT because of statistical dependence of input samples. This problem can be resolved approximately applying the mean field approximation[4] for hidden states to calculate the probability of HMM, by which the probability of whole input samples is nominally represented as the product of probability of each sample as if input samples were independent each other.

2 SPEAKER VERIFICATION USING SPRT

Let $\mathbf{y}_i, i = 1, 2, \cdots$ (i is frame number) represent a sequence of feature vectors obtained from input speech and let $p_s(\mathbf{y}_i)$ and $p_o(\mathbf{y}_i)$ be pdfs for a claimed speaker (true speaker) and all other speakers (impostors), respectively. Given $\mathbf{y}_i, i = 1, 2, \cdots, t$ and assuming that $\mathbf{y}_i$s are mutually independent, the likelihood ratio is given as

$$l_m(\mathbf{y}_1, \cdots, \mathbf{y}_t) = \prod_{i=1}^{t} \frac{p_s(\mathbf{y}_i)}{p_o(\mathbf{y}_i)}. \tag{1}$$

Using this likelihood ratio, the decision is made as

$$l_m(\mathbf{y}_1, \cdots, \mathbf{y}_t) \; > \; A \quad \text{accept as true speaker,} \tag{2}$$

$$l_m(\mathbf{y}_1, \cdots, \mathbf{y}_t) \; < \; B \quad \text{reject as impostor,} \tag{3}$$

where A and B are thresholds with $A > B$. If the value of the likelihood ratio falls between A and B, we take another feature vector and repeat the decision for $t+1$. It is known that the thresholds A and B have the following relation with false rejection rate ε_1 and false acceptance rate ε_2 [1].

$$A \leq \frac{1-\varepsilon_1}{\varepsilon_2} \tag{4}$$

$$B \geq \frac{\varepsilon_1}{1-\varepsilon_2} \tag{5}$$

3 MEAN FIELD APPROXIMATION IN HMM

3.1 Probability Computation in the standard HMM

In the standard HMM, probability of a sequence of observation vectors $\mathbf{y}_i, i = 1, 2, \cdots, t$ is given as

$$\begin{aligned} & p(\mathbf{y}_1, \cdots, \mathbf{y}_t) \\ &= \sum_{x_1,\cdots,x_t} p(\mathbf{y}_1,\cdots,\mathbf{y}_t|x_1,\cdots,x_t)p(x_1,\cdots,x_t) && (6) \\ &= \sum_{x_1,\cdots,x_t} \{\prod_{i=1}^{t} p(\mathbf{y}_i|x_i)\}p(x_1)p(x_2|x_1)\cdots p(x_t|x_{t-1}) && (7) \\ &= \sum_{x_1,\cdots,x_t} \prod_{i=1}^{t} p(\mathbf{y}_i|x_i)p(x_i|x_{i-1}), && (8) \end{aligned}$$

where $x_i, i = 1, 2, \cdots, t$ represents a sequence of hidden states, $p(\mathbf{y}_i|x_i)$ is an observation pdf at state x_i and in (8) $p(x_1|x_0) = p(x_1)$.

3.2 Mean Field Approximation in HMM

Suppose hereafter that $\mathbf{x}_i$ is not a scalar value but an indicator vector to represent the hidden state at i-th frame. Let the number of hidden states be M and $\mathbf{x}_i$ take one from the vector set $Q = \{\mathbf{e}_1, \ldots, \mathbf{e}_K\}$, where $\mathbf{e}_k, 1 \leq k \leq K$ is the K dimensional unit vector whose k-th component is 1 and all other components are 0. When the hidden state at i-th frame is k, $\mathbf{x}_i$ takes $\mathbf{e}_k$. Using the mean field approximation, $p(\mathbf{y}_1, \cdots, \mathbf{y}_t)$ derived by (8) can be approximated as

$$\begin{aligned} p(\mathbf{y}_1,\cdots,\mathbf{y}_t) &\simeq \sum_{\mathbf{x}_1,\cdots,\mathbf{x}_t} \prod_{i=1}^{t} p(\mathbf{y}_i|\mathbf{x}_i)p(\mathbf{x}_i|\langle\mathbf{x}_{i-1}\rangle) && (9) \\ &= \prod_{i=1}^{t} \sum_{\mathbf{x}_i} p(\mathbf{y}_i|\mathbf{x}_i)p(\mathbf{x}_i|\langle\mathbf{x}_{i-1}\rangle) && (10) \\ &= \prod_{i=1}^{t} p(\mathbf{y}_i|\langle\mathbf{x}_{i-1}\rangle), && (11) \end{aligned}$$

where $\langle\bullet\rangle$ is the mean field for $\bullet$. Given the mean fields of $\mathbf{x}_i$s and then replacing $\sum_{\mathbf{x}_1,\cdots,\mathbf{x}_t}\prod_{i=1}^{t}$ by $\prod_{i=1}^{t}\sum_{\mathbf{x}_i}$ we obtain (10) from (9). From (11) it is shown that $p(\mathbf{y}_1, \cdots, \mathbf{y}_t)$ can be computed as if $\mathbf{y}_i$s were independent each other.

Then we describe how to compute the mean fields in practice. Using the mean field approximation, a posteriori probability of a hidden state sequence given an observation sequence is decomposed as

$$\begin{aligned} & p(\mathbf{x}_1,\cdots,\mathbf{x}_t|\mathbf{y}_1,\cdots,\mathbf{y}_t) \\ &= \frac{p(\mathbf{y}_1,\cdots,\mathbf{y}_t|\mathbf{x}_1,\cdots,\mathbf{x}_t)p(\mathbf{x}_1,\cdots,\mathbf{x}_t)}{p(\mathbf{y}_1,\cdots,\mathbf{y}_t)} && (12) \\ &\simeq \frac{\prod_{i=1}^{t} p(\mathbf{y}_i|\mathbf{x}_i)p(\mathbf{x}_i|\langle\mathbf{x}_{i-1}\rangle)}{\prod_{i=1}^{t}\sum_{\mathbf{x}_i} p(\mathbf{y}_i|\mathbf{x}_i)p(\mathbf{x}_i|\langle\mathbf{x}_{i-1}\rangle)} && (13) \\ &= \prod_{i=1}^{t} p(\mathbf{x}_i|\mathbf{y}_i,\langle\mathbf{x}_{i-1}\rangle), && (14) \end{aligned}$$

where

$$p(\mathbf{x}_i|\mathbf{y}_i,\langle\mathbf{x}_{i-1}\rangle) = \frac{p(\mathbf{y}_i|\mathbf{x}_i)p(\mathbf{x}_i|\langle\mathbf{x}_{i-1}\rangle)}{\sum_{\mathbf{x}_i} p(\mathbf{y}_i|\mathbf{x}_i)p(\mathbf{x}_i|\langle\mathbf{x}_{i-1}\rangle)}. \tag{15}$$

$p(\mathbf{x}_i|\mathbf{y}_i,\langle\mathbf{x}_{i-1}\rangle)$ is considered as a local a posteriori probability (LAP) and hereafter we write it as $z_i(\mathbf{x}_i)$ for short. Then the LAPs for all state indicators form a vector (LAP vector), $\mathbf{z}_i = (z_i(\mathbf{x}_i = \mathbf{e}_1), \cdots, z_i(\mathbf{x}_i = \mathbf{e}_K))^T$.

Given an observation sequence, the mean field $\langle\mathbf{x}_i\rangle$ can be defined as

$$\langle\mathbf{x}_i\rangle = \sum_{\mathbf{x}_1,\cdots,\mathbf{x}_t} \mathbf{x}_i p(\mathbf{x}_1,\cdots,\mathbf{x}_t|\mathbf{y}_1,\cdots,\mathbf{y}_t). \tag{16}$$

Using the decomposition of a posteriori probability in (14), (16) is approximately computed as

$$\begin{aligned} \langle\mathbf{x}_i\rangle &\simeq \sum_{\mathbf{x}_1,\cdots,\mathbf{x}_t} \mathbf{x}_i \prod_{i=1}^{t} z_i(\mathbf{x}_i) && (17) \\ &= \sum_{\mathbf{x}_i} \mathbf{x}_i z_i(\mathbf{x}_i) && (18) \\ &= \mathbf{z}_i. && (19) \end{aligned}$$

Finally it is shown that the LAP vector $\mathbf{z}_i$ can be used as the mean field $\langle\mathbf{x}_i\rangle$.

3.3 Summary of the Proposed Probability Computation

We summarize the proposed probability computation of the HMM which can be applied for the SPRT.

(1) Compute $p(\mathbf{y}_1)$ using the initial state probability $p(\mathbf{x}_1)$, i.e.,

$$p(\mathbf{y}_1) = \sum_{\mathbf{x}_1} p(\mathbf{y}_1|\mathbf{x}_1)p(\mathbf{x}_1). \tag{20}$$

Then compute the LAP vector $\mathbf{z}_1$, whose components are derived using (15) with $i = 1$ and $p(\mathbf{x}_1|\langle\mathbf{x}_0\rangle) = p(\mathbf{x}_1)$

(2) For $t \geq 2$ compute $p(\mathbf{y}_1, \cdots, \mathbf{y}_t)$ using (10) with the LAP vector $\mathbf{z}_{i-1}$ for the mean field $\langle\mathbf{x}_{i-1}\rangle$. Let $\mathbf{A}$ be a state transition matrix whose component a_{kl} represents the transition probability from state k to l, then $p(\mathbf{x}_i|\mathbf{z}_{i-1})$ in (10) is described as

$$p(\mathbf{x}_i|\mathbf{z}_{i-1}) = \mathbf{z}_{i-1}^T\mathbf{A}\mathbf{x}_i. \tag{21}$$

Then compute $\mathbf{z}_i$ using (15) again with $\mathbf{z}_{i-1}$ for $\langle\mathbf{x}_{i-1}\rangle$.

4 PARAMETER ESTIMATION

In the following experiments an ergodic HMM is used and a single Gaussian pdf is associated as an output pdf with each hidden state in the HMM. A set of Gaussian pdfs for all states is commonly used for all speakers and only transition probabilities between states are assumed to be different from speaker to speaker. In order to compare the performance of the proposed context-dependent method with that of a conventional context-independent one using the SPRT, another experiments are carried out using a mixture of Gaussian distributions as an pdf of i.i.d. input samples[5]. This conventional approach using the Gaussian mixture can be interpreted as that a hidden state is chosen according to mixing coefficients independently from context, i.e., the previous hidden states. On the other hand, in the proposed approach the hidden states are modeled by the Markov model and therefore from the Gaussian mixture's point of view it can be said that the mixing coefficients vary with context.

Parameter estimation of the HMM is carried out by the Baum algorithm[6], which is identical to the Expectation and Maximization (EM) method[7]. However we do not straightly use the Baum algorithm to estimate the whole parameters in the HMM because we want to look at performance change over the Gaussian mixture approach. That is, the same Gaussian pdfs as those in the Gaussian mixture approach are used in the HMM and then only state transition probabilities are estimated by the Baum algorithm (in fact by an approximated version of it[1]). In addition the mixing coefficients in the Gaussian mixture are also used as the initial state probabilities in the HMM.

A mixture of K Gaussian distributions is given as

$$p(\mathbf{y}_i) = \sum_{k=1}^{K} a_k g_k(\mathbf{y}_i; \mathbf{m}_k, \mathbf{\Sigma}_k), \quad \sum_{k=1}^{K} a_k = 1 \tag{22}$$

$$g_k(\mathbf{y}_i; \mathbf{m}_k, \mathbf{\Sigma}_k) = \frac{1}{(2\pi)^{D/2}|\mathbf{\Sigma}_k|^{1/2}} \cdot \exp\{-\frac{1}{2}(\mathbf{y}_i - \mathbf{m}_k)^T \mathbf{\Sigma}_k^{-1}(\mathbf{y}_i - \mathbf{m}_k)\}, \tag{23}$$

where $\mathbf{y}_i$ is a D dimensional feature vector at i-th frame and a_k is the mixing coefficient of the k-th Gaussian distribution $g_k(\mathbf{y}_i; \mathbf{m}_k, \mathbf{\Sigma}_k)$ with mean vector $\mathbf{m}_k$ and covariance matrix $\mathbf{\Sigma}_k$. The model parameters, $a_k, \mathbf{m}_k, \mathbf{\Sigma}_k, k = 1, \cdots, K$ are iteratively estimated by the EM method. Explicit procedures are found in [9],[5],[3]. The initial values to start the iterative procedure are obtained by clustering training samples using the VQ method[10]. The obtained Gaussian distributions are commonly used for both $p_s(\mathbf{y}_i)$ and $p_o(\mathbf{y}_i)$ in (1), also in the HMM approach as well as in the Gaussian mixture approach.

Using the derived Gaussian pdfs $g_k(\mathbf{y}_i; \mathbf{m}_k, \mathbf{\Sigma}_k)$, $k = 1, \cdots, K$, the pdf for each speaker $p_s(\mathbf{y}_i), s = 1, \cdots, S$ in the Gaussian mixture approach can be modeled as

$$p_s(\mathbf{y}_i) = \sum_{k=1}^{K} a_{s,k} g_k(\mathbf{y}_i; \mathbf{m}_k, \mathbf{\Sigma}_k), \quad \sum_{k=1}^{K} a_{s,k} = 1, \tag{24}$$

where $a_{s,k}$s are mixing coefficients for speaker s. Given training samples of speaker s, $\mathbf{y}_i, i = 1, \cdots, N_s$, $a_{s,k}$s are iteratively estimated as

$$a_{s,k}^{(p+1)} = \frac{1}{N_s}\sum_{i=1}^{N_s} \alpha_{s,k}^{(p)}(\mathbf{y}_i), \tag{25}$$

$$\alpha_{s,k}^{(p)}(\mathbf{y}_i) = \frac{a_{s,k}^{(p)} g_k(\mathbf{y}_i; \mathbf{m}_k, \mathbf{\Sigma}_k)}{\sum_{k'=1}^{K} a_{s,k'}^{(p)} g_{k'}(\mathbf{y}_i; \mathbf{m}_{k'}, \mathbf{\Sigma}_{k'})}. \tag{26}$$

The initial values, $a_{s,k}^{(0)}$s are derived by clustering the training samples $\mathbf{y}_i, i = 1, \cdots, N_s$ by

$$k = \arg\max_{k'} g_{k'}(\mathbf{y}_i; \mathbf{m}_{k'}, \mathbf{\Sigma}_{k'}), \tag{27}$$

then calculating the normalized sample frequencies assigned to each class. Once the mixing coefficients for all speakers are obtained, those for impostors of speaker s, $a_{o,k}$s are obtained by

$$a_{o,k} = \frac{1}{S-1} \sum_{\substack{t=1 \\ t \neq s}}^{S} a_{t,k}. \tag{28}$$

[1] We adopt the mean-field-based approximation method proposed by Zhang[8] instead of the original Baum algorithm, because it is reported that the approximation method has an advantage of computational simplicity and gives comparable performance of parameter estimation to the Baum algorithm.

Table 1 Speaker verification errors(%) by two methods

	Proposed (HMM)				Gaussian mixture			
	closed		open		closed		open	
$\ln A, \ln B$	FR	FA	FR	FA	FR	FA	FR	FA
2.2 , -2.2	4.9	12.2	25.2	12.8	18.9	32.8	21.1	32.4
4.6 , -4.6	0.6	5.1	18.0	5.1	12.0	27.5	15.7	27.1
6.9 , -6.9	0.2	2.9	15.1	2.6	8.9	25.0	12.1	24.1
11.5 , -11.5	0.0	1.6	10.8	1.7	3.6	21.8	7.8	20.8

5 SPEAKER VERIFICATION EXPERIMENTS

Text-independent speaker verification experiments were carried out. The used telephone speech data-set consists of isolated uttered Japanese 20 words produced two repetitions by 100 male speakers in two sessions spaced three to four months apart[3]. The speech data was low-pass filtered at 4.5 kHz and digitized at 10 kHz sampling rate. The digitized speech was pre-emphasized with a first-order adaptive filter and subjected to 12th order LPC analysis with 25.6 msec Hamming window and 12.8 msec frame rate. In fact the Selective LPC analysis was applied to use the spectral information up to 4 kHz considering that the speech data is telephone speech. The twelve LPC cepstral coefficients obtained by this analysis were used as a feature vector for each time frame.

The data-set was divided into two sets and used for open and closed experiments. In open experiments training and test set are different and in closed experiments both are the same. In the following experiments the covariance matrix $\mathbf{\Sigma}_k$ of each Gaussian distribution $g_k(\mathbf{y}_i; \mathbf{m}_k, \mathbf{\Sigma}_k)$ is assumed to be diagonal. In SV experiments word utterances of each speaker are connected and used in an endless way. The number of tests per speaker is 20 for utterances of the same speaker and 20 for those of impostors, i.e., totally 2000 for both cases. In each test starting point of input is randomly selected and impostors are also randomly selected from 99 speakers excluding the relevant true speaker.

Experimental results using 16 hidden states in the HMM and 16 mixtures in the Gaussian mixture are shown in Table 1. Here $\ln A$ and $\ln B$ are log threshold values for the decisions in (2) and (3). $\ln A = 2.2$ and $\ln B = -2.2$, 4.6 & -4.6, 6.9 & -6.9, and 11.5 & -11.5 are obtained by changing the inequalities (4) and (5) to equalities and putting $\varepsilon_1 = \varepsilon_2 = 10^{-1}, 10^{-2}, 10^{-3}$, and 10^{-5}, respectively. FR and FA represent false rejection and false acceptance error, respectively. Table 1 shows that the proposed method gives much better performance than the Gaussian mixture approach except FR in open experiments.

Acknowledgments

This work was partly supported by the Tateisi Science and Technology Foundation, Japan.

References

[1] K. Fukunaga, "Introduction to statistical pattern recognition (Second Edition)," Academic Press, pp.110-119, 1990.

[2] M.A. Lund, "A robust sequential test for text-independent speaker verification," J. Acoust. Soc. Am., vol.99, no.1, pp.609-621, Jan. 1996.

[3] K. Harada, H. Noda and E. Kawaguchi, "Text-independent speaker verification using sequential decision," Technical Report of IEICE, PRMU97-81 (vol.97, no.205, pp.67-72), 1997.

[4] K. Huang, "Statistical mechanics (Second Edition)," John Wiley & Sons, 1987.

[5] D.A. Reynolds and R.C. Rose, "Robust text-independent speaker identification using Gaussian mixture speaker models," IEEE Trans. Speech & Audio Processing, vol.3, no.1, pp.72-83, Jan. 1995.

[6] L.E. Baum, "An inequality and associated maximization technique in statistical estimation for probabilistic functions of Markov processes," Inequalities, vol.3, pp. 1-8, 1972.

[7] A.P. Dempster, N.M. Laird and D.B. Rubin, "Maximum likelihood from incomplete data via the EM algorithm," J. Roy. Stat. Soc., vol.39, no.1, pp. 1-38, 1977.

[8] J. Zhang, "The mean field theory in EM procedures for Markov random fields," IEEE Trans. Signal Process., vol.40, no.10, pp. 2570-2583, Oct. 1992.

[9] G.J. McLachlan and K.E. Basford, "Mixture models - Inference and applications to clustering," Marcel Dekker, 1988.

[10] Y. Lind, A. Buzo and R.M. Gray, "An algorithm for vector quantizer design," IEEE Trans. Commun., vol.28, pp.84-95, 1980.