論文 / 著書情報
Article / Book Information

Title	Distance Measures for Text-Independent Speaker Recognition Based on MAR Model
Author	Chintana Griffin, Tomoko Matsui, Sadaoki Furui
Journal/Book name	IEEE ICASSP 94, Vol. , No. I, pp. 309-312
発行日 / Issue date	1994, 4
権利情報 / Copyright	(c)1994 IEEE. Personal use of this material is permitted. However, permission to reprint/republish this material for advertising or promotional purposes or for creating new collective works for resale or redistribution to servers or lists, or to reuse any copyrighted component of this work in other works must be obtained from the IEEE.

Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing,
Adelaide, Australia, pp. I-309-312 (April 1994)

DISTANCE MEASURES FOR TEXT-INDEPENDENT SPEAKER RECOGNITION BASED ON MAR MODEL

Chintana Griffin, Tomoko Matsui, and Sadaoki Furui
NTT Human Interface Laboratories
9-11, Midori-Cho 3-Chome
Musashino-Shi, Tokyo, 180 Japan

ABSTRACT

For text-independent speaker identification and verification based on the Multivariate Auto-Regression model, we consider two distance measures: the Log Likelihood Ratio (LLR) and the Symmetrized Likelihood Ratio (SLR) measure, which is a symmetric version of the Likelihood Ratio measure. The results of experiments indicate that the LLR gives better performance than the SLR for longer training data of 5 or more sentences, and the SLR measure is better for shorter training data. When 10 sentences are used for training, identification and verification rates (after likelihood normalization) are almost the same as those obtained by an HMM-based method. The optimum order of the MAR model is 2 or 3, and the optimum frame period is 16 *ms*.

1 INTRODUCTION

Various speaker models and distance/distortion measures have already been proposed for text-independent speaker recognition. Most of these models use static features, such as long-time averaged spectrum or the distribution of feature parameters. The distribution is represented by vector quantization (VQ) codebooks or HMMs. A very small number of studies on using dynamic features in text-independent speaker recognition have been reported, and this remains one of the important issues in this field.

In this paper we use Multivariate Auto-Regression (MAR) to obtain a text-independent dynamic speaker's model, following the approach taken in [1]. We consider two different distance measures: the Log Likelihood Ratio (LLR) and the Symmetrized Likelihood Ratio (SLR). The formulation of LLR distance can be found in [2]; and the formulation of SLR, which is an extension of the LLR, is given here.

2 MULTIVARIATE AUTO-REGRESSION

Parameters of MAR models can be estimated by a procedure which is a generalization of the Durbin recursion to the multivariate case (see [3] for more details). In this section, we simply give the formulas necessary for the computation of the MAR model.

Consider the multivariate auto-regression

$$\sum_{k=0}^{p} \mathbf{A}_k^p \, \mathbf{X}_{t-k} \;=\; \epsilon_t, \tag{1}$$

where $\{\mathbf{X_t}\}$ is an m-dimensional stationary process, $\{\epsilon_t\}$ is a zero mean, uncorrelated and stationary process, and p is the order of the MAR model. The coefficients $\mathbf{A}_k^p$ are $m \times m$ matrices with $\mathbf{A}_0^p \;=\; \mathbf{I}$. Analogous to the scalar case, the auto-covariances $\mathbf{\Gamma}_k \;=\; E(\mathbf{X}_t\mathbf{X}'_{t-k})$ obey the Yule-Walker relations

$$\sum_{k=0}^{p} \mathbf{A}_k^p \, \mathbf{\Gamma}_{j-k} \;=\; 0, \quad j \;=\; 1, 2, \ldots, p. \tag{2}$$

The coefficients $\mathbf{A}_k^p$, $k \;=\; 1,\ 2,\ \ldots,\ p$ can be determined from this linear equation system using the generalized Durbin recursion where the mean square error is minimized.

The generalized Durbin recursion is as follows:

$$\mathbf{A}_k^{p+1} \;=\; \mathbf{A}_k^p + \mathbf{A}_{p+1}^{p+1}\bar{\mathbf{A}}_{p-k+1}^p, \tag{3}$$

where $k \;=\; 1,\ 2,\ \ldots,\ p$.

$$\mathbf{A}_{p+1}^{p+1} \;=\; -\Delta_p \bar{\mathbf{V}}_p^{-1}, \tag{4}$$

where

$$\mathbf{V}_p \;=\; \sum_{k=0}^{p} \mathbf{A}_k^p \mathbf{\Gamma}_{-k}, \tag{5}$$

$$\Delta_p \;=\; \sum_{k=0}^{p} \mathbf{A}_k^p \mathbf{\Gamma}_{p-k+1}, \tag{6}$$

Note that the coefficient estimates $\mathbf{A}_k^p$ depend on the fitting order p.

3 DISTANCE MEASURES

A distance or distortion measure is used to quantify similarities between two speech signals. In general, we would like a distance that is nonnegative and symmetrical. In this paper, we consider two distance measures: the Log Likelihood Ratio (LLR) and Symmetrized Likelihood Ratio (SLR).

Let $\{\mathbf{X}_n\}_{n=0}^{M-1}$ be some spectral vectors of speaker $\mathbf{X}$, and $\{\mathbf{A}_k^p\}_{k=1}^{p}$ the MAR model derived from $\{\mathbf{X}_n\}_{n=0}^{M-1}$. Similarly, let $\{\mathbf{Y}_n\}_{n=0}^{N-1}$ be spectral vectors of an unknown speaker $\mathbf{Y}$, and $\{\mathbf{B}_k^p\}_{k=1}^{p}$ the MAR model derived from $\{\mathbf{Y}_n\}_{n=0}^{N-1}$.

The residual error of $\mathbf{X}_n$ filtered by the model $\{\mathbf{A}_k^p\}_{k=1}^{p}$ is

$$\mathbf{e}_n^{XA} \;=\; \mathbf{X}_n \;+\; \sum_{k=1}^{p} \mathbf{A}_k^p \mathbf{X}_{n-k}, \tag{7}$$

where $n = p, p+1, \ldots, M-1$. The residual errors $\mathbf{e}_n^{YB}$, $\mathbf{e}_n^{XB}$, $\mathbf{e}_n^{YA}$ can be obtained similarly. Then the covariance matrices of the residual errors $\mathbf{D}^{XA}$, $\mathbf{D}^{YB}$, $\mathbf{D}^{XB}$, and $\mathbf{D}^{YA}$ are obtained from $\mathbf{e}^{XA}$, $\mathbf{e}^{YB}$, $\mathbf{e}^{XB}$, and $\mathbf{e}^{YA}$, respectively.

The LLR measure is defined as follows:

$$d^{LLR}(\mathbf{X},\mathbf{Y}) = \log_{10}[\det(\mathbf{D}^{YA} * \mathbf{D}^{YB(-1)})]. \quad (8)$$

The SLR is an extension of the LLR, and is defined as follows:

$$d^{SLR}(\mathbf{X},\mathbf{Y}) = \overbrace{[\det(\mathbf{D}^{YA} * \mathbf{D}^{YB(-1)})}^{d_1^{SLR}} + \underbrace{\det(\mathbf{D}^{XB} * \mathbf{D}^{XA(-1)})]}_{d_2^{SLR}}/2. \quad (9)$$

Note also that the term d_1^{SLR} in equation (9) is the LLR without the $\log_{10}$ function, and the term d_2^{SLR} is the symmetrization part.

4 SPEAKER RECOGNITION

For speaker identification, the distance measures d^{LLR} and d^{SLR} are calculated between each of the reference speakers $\mathbf{X}$ and the unknown speaker $\mathbf{Y}$. The reference speaker that yields the smallest distance measure with respect to a particular measure is considered to be the identity of the speaker $\mathbf{Y}$.

For speaker verification, distance measures are calculated between the unknown speaker and the reference speaker whose identity has been claimed. The input speech is accepted or rejected by comparing the distance with a threshold. In our experiments, the threshold was set *a posteriori* to equalize the probability of false acceptance and false rejection.

5 DATA DESCRIPTION

The database consisted of 15 Japanese speakers: 10 males and 5 females. Each speaker uttered continuous speech (various texts). A sentence was approximately 4 *s* in duration. From the speech data, cepstral coefficients of order 16 were obtained from a 32-*ms* window every 16-*ms* frame-period (baseline condition).

Three sets of data were used for training, test session 1, and test session 2. Each data set was collected at a different time. The data for test session 1 and training were collected 4 months apart; and test session 2 and training, 3 months apart. In each test set, each speaker uttered 5 different sentences.

6 EXPERIMENTS & RESULTS

In our experiments, we varied the number of sentences used in the training phase: 1, 2, 5, or 10 sentence(s). When 2 or more sentences were used, the auto-covariance matrices were ensemble averaged over the number of sentences. In the testing phase, only 1 sentence was used for all cases. Note that the fitting order p was 2 for the baseline condition.

Speaker identification results for each distance measure are shown in Fig. 1. Each identification rate was averaged by 150, i.e. 15 speakers, 5 test sentences, and 2 sessions per speaker. The breakdowns of d^{SLR} into two constituent parts are shown in Table 1, which will be refered to later in the discussion section.

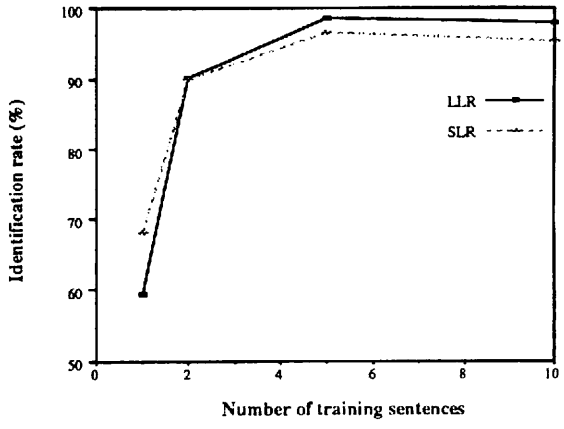

Fig. 1 : Identification rates as a function of the number of training sentences (MAR order: 2)

Table 1: Indentification rates by each constituent part of SLR (MAR order: 2)

Number of training sentences	d_1^{SLR} (d^{LLR})	d_2^{SLR}
1	59.3%	60.7%
2	90.0%	79.3%
5	98.7%	92.7%
10	98.0%	91.3%

In Fig. 1, we make the following observations:

- The identification rate of LLR and SLR increases as the number of training sentences increases;
- SLR has better rates for 1 training sentence, but LLR has better rates for 5 and 10 training sentences.

Speaker verification results for each distance measure are shown in Figure 2. There are 2 sets of results for speaker verification: one using likelihood normalization [4] and one without it. We can see that the normalization dramatically improves the rates of d^{LLR} and d^{SLR}.

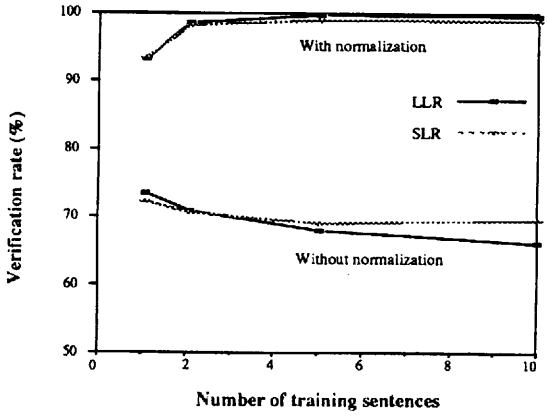

Fig. 2: Verification rates as a function of the number of training sentences (MAR order: 2)

The verification results after normalization seems to follow the same trends as those in Fig. 1.

7 DISCUSSION

7.1 LLR vs. SLR

In this section, we discuss the differences between the formulation of LLR and SLR, and how these differences affect their performance. Since the results of speaker identification and verification both follow the same trends, the discussion below applies to both cases unless otherwise stated.

From the results in the previous section, one may ask the following questions:

- Why is the performance of SLR higher than that of LLR for 1 training sentence, but lower for 5 and 10 sentences in speaker identification (Fig. 1)? One would expect SLR to consistently have better performance since it uses more information.
- Why is the performance of LLR higher than that of SLR for 5 and 10 sentences in speaker verification after likelihood normalization (see Fig. 2)?

To answer these questions, we shall analyze each component of d^{SLR}. Recall that d^{SLR} is the sum of d_1^{SLR} and d_2^{SLR} (see equation (9)).

First of all, we want a distance measure that is nonnegative and symmetrical. Furthermore, we would like it to attain a specific value in a certain (ideal) case; for example, if there is a perfect match between the reference and test data then the value of the distance should be unity. Both LLR and SLR are actually bounded below by unity.

Now, we shall answer the question of why SLR performs better than LLR for 1 training sentence, but worse for 5 and 10 sentences in speaker identification. Again we look at the individual identification rates in Table 1. (Note that $d^{LLR} \equiv d_1^{SLR}$.) We can see that the improvement by increasing the number of training sentences is much smaller for d_2^{SLR} than for d_1^{SLR}. The reason for this, we believe, lies in the stability of the model. From Eq. (9), only parameter **A** of d_1^{SLR} is variable while parameters **Y** and **B** are fixed in the comparison between input speech and reference models. Since **A** becomes stable as the number of training sentences increases, the performance of d_1^{SLR} improves steadily. On the other hand, in d_2^{SLR}, only parameter **B** is fixed while other parameters, **X** and **A**, are variable. Since the spectral vectors for training are directly used in d_2^{SLR}, it is probably not stabilized so much by the increase in the number of training sentences as d_1^{SLR}. Thus the sum of d_1^{SLR} and d_2^{SLR} has better identification rates than that of d^{LLR} only for the one-sentence training condition. Different weightings in the linear combination, $(1 - w) * d_1^{SLR} + w * d_2^{SLR}$, $0 \leq w \leq 1$, raise the rates to equal that of d^{LLR}.

Finally, the reason why the performance of LLR is much higher than that of SLR for 5 and 10 sentences in speaker verification after likelihood normalization is as follows. The likelihood normalization method implicitly uses results from speaker identification for normalizing the likelihood values. Since LLR has much better identification rates than SLR for 5 and 10 sentences (Fig. 1), then this is also the case for the verification rates.

7.2 Effects of the MAR order

As described in Sec. 6, we set the MAR fitting order of p at 2 in the baseline experiments. In order to check the effect of the order, additional experiments were conducted by changing the order between 1 and 4. Figure 3 indicates the speaker identification rates for LLR and SLR for the two cases where the number of training sentences was 1 and 10. Identification rates were almost stable irrespective of the order p when only one sentence was used for training. On the other hand, when 10 sentences were used for training and LLR was used as the distance measure, the identification rate increased as the MAR order increased from 1 to 3. The improvement from the order 2 to 3 was smaller than that from 1 to 2. There was no improvement in the identification rate when the order was increased from 3 to 4. These results suggest that (1) one sentence is too short for estimating an MAR model having an order larger than 1; and (2) when 10 sentences can be used for training, the optimum order of the MAR model is 2 or 3.

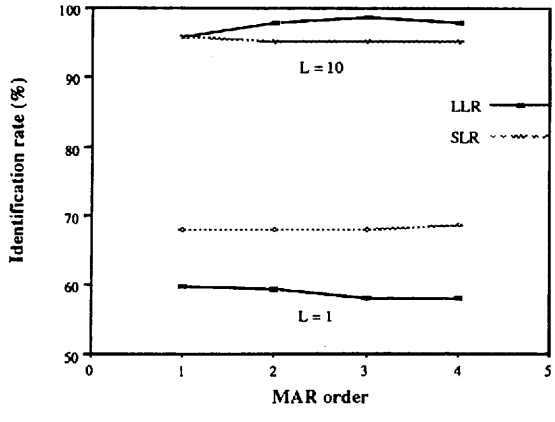

Fig. 3: Identification rates as a function of the MAR order (L: number of training sentences)

7.3 Effects of the frame period

In the experiments so far, the frame period for extracting cepstral features was set to 16 *ms*. An additional experiment was conducted to examine the effect of the frame period on the identification accuracy. In this experiment, the MAR order was fixed to 2. Therefore, the time period for which the dynamic characteristics are represented by the MAR models was changed in proportion to the frame period. Experimental results shown in Fig. 4 indicate that the optimum frame period across all the conditions of the number of training sentences is 16 *ms*. Under this condition, the MAR model predicts cepstral vectors based on the past two vectors observed 16 *ms* and 32 *ms* before, respectively. In other words, speaker-specific spectral dynamics over a 32-*ms* period is represented by the MAR model.

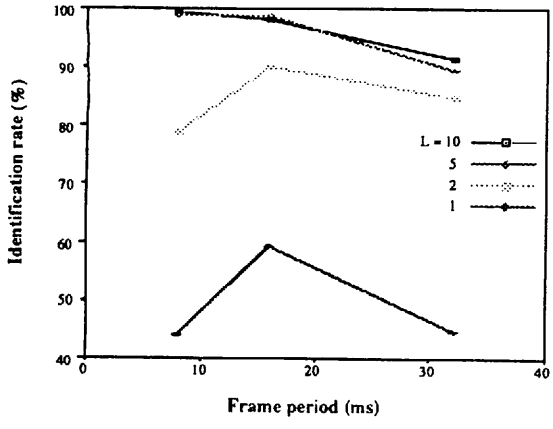

Fig. 4: Identification rate as a function of the frame period (MAR order: 2, distance measure: LLR, L: number of training sentences)

7.4 Combination with the HMM method

As described in previous sections, the MAR models represent dynamic features of the time series of cepstral vectors. We also conducted a speaker recognition experiment based on static features represented by an ergodic HMM using the same speech database [5]. Since dynamic and static information is rather independent, speaker recognition performance is expected to be improved by combining these two types of information. An experiment was conducted, in which the weighted sum of log likelihood values obtained by MAR based on the LLR distance measure and by HMM was calculated and used for making the decision. The weighting factor was set *a posteriori* at the optimum value that produced the best results. Table 2 summarizes the recognition rates. Since the verification rate after normalization by the HMM method is 100%, the combination of MAR and HMM methods was tried only for identification. The identification rate obtained by the combination method was 98.7%, which is significantly higher than either of the single methods.

Table 2 : Comparison of recognition rates (MAR order : 2, number of training utterances : 10)

	MAR (LLR)	HMM	MAR and HMM
Identification	98.0 %	93.3 %	98.7 %
Verification (unnormalized)	66.1 %	97.3 %	—
Verification (normalized)	99.7 %	100 %	—

8 SUMMARY

Based on the Multivariate Auto-Regression (MAR) model, we have analyzed the formulations and experimentally compared the performances of the Log Likelihood Ratio (LLR) and the Symmetrized Likelihood Ratio (SLR) measures for speaker identification and verification. The experimental results indicate that the LLR measure performs better than the SLR for longer training data of 5 or more sentences, while the SLR measure performs better for shorter training data of 1 sentence. Therefore, the LLR measure should be used when long training data is available, and SLR when only short training data is available. Note that only 1 sentence was used for testing for all cases.

For speaker verification, the use of the likelihood normalization dramatically improves the performance of both LLR and SLR. Since the likelihood normalization uses results from speaker identification for normalizing the likelihood in speaker verification, the rates of the verification follows the trends of the identification.

When 10 sentences are used for training, the MAR method achieves almost the same identification and verification rates as an HMM-based method. The optimum order of MAR model is 2 or 3, and the optimum frame period is 16 *ms*.

REFERENCES

[1] C. Montacié, et al.: "Cinematic Techniques for Speech Processing: Temporal Decomposition and Multivariate Linear Prediction," Trans. ICASSP, San Francisco, pp. I-153-156 (1992).

[2] S. Furui: "Digital Speech Processing, Synthesis, and Recognition," Marcel & Dekker, New York (1989).

[3] P. Whittle: "On the Fitting of Multivariate Autoregression and the Approximate Canonical Factorization of a Spectral Density Matrix," Biometrika, vol. 50, pp. 129-134 (1963).

[4] T. Matsui and S. Furui: "Speaker Recognition Using Concatenated Phoneme Models," Trans. ICASSP, Minneapolis, pp. II-391-394 (1993).

[5] T. Matsui and S. Furui: "Comparison of text-independent speaker recognition methods using VQ-distortion and discrete/continuous HMMs," Trans. ICASSP, San Francisco, pp. II-157-160 (1992).