ISCA Archive
http://www.isca-speech.org/archive

6th International Conference on Spoken
Language Processing (ICSLP 2000)
Beijing, China
October 16-20, 2000

Gaussian Similarity Analysis and Its Application in Speaker Adaptation

Ji Wu, Zuoying Wang

(Speech Recognition Lab, Department of Electronic Engineering,
Tsinghua University, Beijing, 100084)

Abstract

A good similarity measure of random variables is crucial in many applications. The choice of distance measure directly affects quality of system design. In this paper, we present a new measure of the similarity between two random variables. The discussion here emphasizes on the case of normal distribution. Based on this Gaussian Similarity Analysis (GSA), we propose an algorithm in speaker adaptation of covariance. It is different from the traditional algorithms, which mainly focus on the adaptation of mean vector of state observation probability density. A binary decision tree is constructed offline with the similarity measure and the adaptation procedure is data-driven. It can be shown from the experiments that we can get a significant further improvement over the mean vectors adaptation.

1.Introduction

We may introduce a distance measure to describe the similarity between two points in space, but for random variables, the problem is quite complicated. It is very important to give a measure for this similarity in many applications. In this paper, we discuss how to give a general concept of this measure for random variables and then apply it to Gaussian random variables. In case of Gaussian distribution, a simple formula is given.

In recent years, speaker adaptation is one of the main research interests in the field of speech recognition, because the adaptation technique can greatly improve the performance of large-vocabulary continuous speaker-independent (SI) speech recognition system [1]. Many adaptation schemes have been proposed, such as MAP [2], MLLR [3] and MLMI [4].

In large vocabulary continuous speech recognition (LVCSR) system, speech is often modeled by continuous-density HMMs with states of multivariant Gaussian distributions. Most of the adaptation methods are only applied to update the mean vectors of Gaussian distributions. Gaussian covariance has much more parameters, so the covariance adaptation needs a great amounts of speaker-specific data, which is the situation in some adaptation schemes, like MAP [5], but it is difficult, even impossible, to have so much adaptation data in real applications.

In this paper, we propose a new algorithm, named as GSA, to estimate the similarity between two Gaussian distributed variables. Then, a method based on this similarity measure is developed in speaker adaptation. With this measure, a binary decision tree is constructed, each node of the tree is a cluster of Gaussian variables. And a group of transformations between the Gaussian variables and the corresponding cluster centers can also be derived. If we have enough adaptation data to estimate the covariance of these cluster centers for a specific speaker, then we can get all adapted covariance for him. It will be much easier because the parameters need estimation has been greatly reduced. According to different adaptation data, different nodes can be adaptively chosen to serve as clusters, so we can make the best use of these limited data.

Section 2 begins with a brief introduction of GSA. And the speaker adaptation algorithm based on GSA is described in section 3 in detail. Then the experiment results and discussion are given in section 4. Section 5 draws the conclusion and gives the further research direction in the end.

2. Gaussian Similarity Analysis

2.1 Similarity Measure of random vectors

Suppose x and h are two random vectors, then the distance between these two random vectors can be defined naturally as

$$d = \frac{1}{N}\sum_{k=1}^{N}\|h_k - x_k\|^2 = \frac{1}{N}\sum_{k=1}^{N}\|g(x_k) - x_k\|^2$$

(shown in Figure.1). Where x_k and h_k (k=1,2,…,N) are samples of x and h . When $N \longrightarrow \infty$, the above definition becomes

$$d = \overline{\|h(x) - x\|^2} = \overline{\|(g(x) - g) - (x - m)\|^2} + \|g - m\|^2 \quad (1)$$

where m and g are the means of x and h respectively,$\|\cdot\|$ means the Euclidian norm of vectors. The bar upon norm stands for the statistics average.

Let random variables

$$x = x - m, \quad y = g(x) - g \qquad (2)$$

then the distance can be written as

$$d = \overline{\|y(x) - x\|^2} + \|g - m\|^2 \qquad (3)$$

To describe the similarity, we need to seek a mapping $g(\bullet)$, which will minimize distance (1)

$$d = J_{\min}(g) = \min_{g}\left(\overline{\|g(x)-x\|^{2}} + \|g-m\|^{2}\right) \tag{4}$$

and in the mean time should satisfy restriction

$$\left|\det(\partial g(x)/\partial x)\right| p_{y}(g(x)) = p_{x}(x) \tag{5}$$

where $|det(\ \partial g(\ x\ )\ /\ \partial x\ )|$ is the absolute value of Jacobian determinant of $g(\ x\ )$. Define Lagrange functional

$$F(g) = J(g) + l\left\{\left|\det(\partial g(x)/\partial x)\right| p_{y}(g(x)) - p_{x}(x)\right\} \tag{6}$$

where l is a bounded linear functional on space $L_2(\ x\ )$ and its definition is

$$l(\ f\ ) = \int_{R^n} l(x)\ f(x)\,dx\ ,\ \forall f \in L_2(x) \tag{7}$$

$L_2(x)$ is a Hilbert space composed of all square integratable real functions on R^n.

According to formulas (5) and (6), once $p_x(x)$ and $p_y(y)$, the probability density functions (p.d.f) of two random variables are given, we can obtain the similarity $J_{\min}(g)$ between these two distributions.

The similarity described above is a useful concept for VQ of random variables. As an application of this concept, we discuss the case of Gaussian distributions.

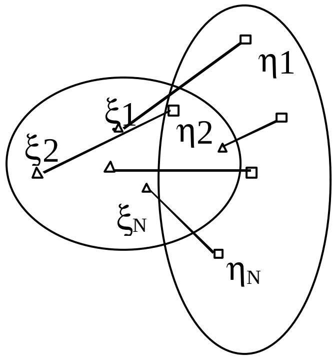


Figure.1 Correspondence between two random variables

2.2 Gaussian Similarity Analysis

If x and h assume both Gaussian distributions $N(m_x, R_x)$ and $N(m_y, R_y)$ respectively, then from (6).

$$F'(g)h = 2\overline{[g(\mathbf{x})-\mathbf{x}]h} + + \\ l\left\{\left|\det(\partial g(x)/\partial x)\right| p_y(g(x)) \begin{bmatrix} \left\langle (\partial g(x)/\partial x)^{-1}, \partial h(x)/\partial x \right\rangle_{R^n} \\ -\left\langle R_y^{-1} g(x), h(x) \right\rangle_{R^n} \end{bmatrix}\right\} = 0 \tag{8}$$

Because $L_2(x)$ is self-dual, it is easy to prove that factor $\left|\det(\partial g(x)/\partial x)\right| p_y(g(x))$ and l can be combined into one bounded linear functional on $L_2(x)$, i.e. $g = \left|\det(\partial g(x)/\partial x)\right| p_y(g(x))\ l(x)$ can be regarded as a new bounded linear functional on $L_2(x)$ and formula (4) becomes

$$F'(g)h = 2\overline{[g(\mathbf{x}) - \mathbf{x}]h} + + \\ g\left[\left\langle (\partial g(x)/\partial x)^{-1}, \partial h(x)/\partial x \right\rangle_{R^n} - \left\langle R_y^{-1} g(x), h(x) \right\rangle_{R^n}\right] = 0 \\ \forall h \in L_2(\ x\ ) \tag{9}$$

From (5) an equivalent restriction

$$A^t R_y^{-1} A = R_x^{-1} \tag{10}$$

can be obtained. A is the linear transformation between R_x^{-1} and R_y^{-1}. In fact a Gaussian distribution is characterized uniquely by its covariance. Formula (8) becomes

$$F'(g)h = \left\langle (A-I)R_x, h_A \right\rangle_{R^n} + g\left\langle A^{-1} - R_y^{-1}A(xx^t), h_A \right\rangle_{R^n} = 0 \tag{11}$$

Because g is a linear functional on $L_2(x)$, the result of its operation on a linear combination of several functions is equal to the linear combination of results by operation on each components. Therefore, g in (6) may operate on every components of matrix $A^{-1} - R_y^{-1}A(xx^t)$ first and then calculate the inner dot with h_A in R^n. Formula (6) becomes

$$\left\langle (A-I)R_x + A^{-1}g(I) - R_y^{-1}Ag(xx^t), h_A \right\rangle = 0$$

then

$$(A-I)R_x + A^{-1}gI - R_y^{-1}Ag(xx^t) = 0 \tag{12}$$

Finally we can obtain

$$A = R_y^{1/2}\left(R_y^{-1/2} R_x^{-1} R_y^{-1/2}\right)^{1/2} R_y^{1/2} \tag{13}$$

The similarity measure $d(x,h)$ is defined as the minimum of $J(A)$

$$d(x,h) = J_{\min}(A) = trace\left(R_y - 2\left[R_x^{1/2} R_y R_x^{1/2}\right]^{1/2} + R_x\right) \tag{14}$$

3.Adaptation Algorithm based on GSA

The adaptation procedure is made up of two steps. The first step is to construct a binary decision tree offline. When the adaptation data was input, try to update the acoustic model parameters according to these data online. This is the second step. Our recognition system is based on a modified HMM called the duration distribution based hidden Markov model (DDBHMM)[6], and we assume that each state comprises of a single Gaussian with the full covariance matrix.

3.1 Binary Decision Tree Definition

With the similarly measure defined in (14), we can get the distance between i^{th} and j^{th} state, which called as S_i and S_j:

$$d(S_i, S_j) = J_{\min}\left(A_{i,j}\right)$$
$$= trace\left(R_j - 2\left[R_i^{1/2} R_j R_i^{1/2}\right]^{1/2} + R_i\right) \quad (15)$$

Where R_i and R_j is corresponding covariance matrix, $A_{i,j}$ is the transformation matrix from R_i to R_j:

$$R_j = A_{i,j} R_i A_{i,j}^t$$

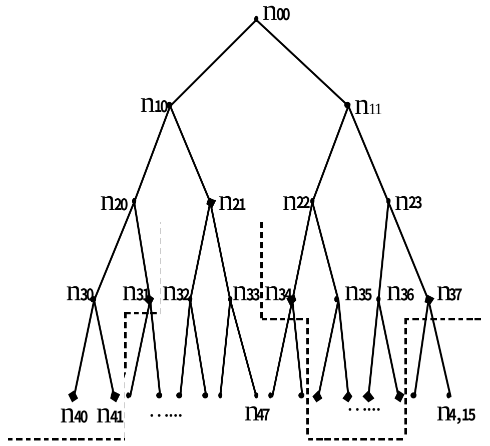
First, all states are put in the root node. Then in the second level, the node split into two nodes, and the K-means algorithm is used. The covariance of the center in the f^{th} cluster c_f is the weighted means of all covariance matrices in this cluster. So the covariance matrix of this cluster center is:

$$R_f = \frac{\sum_{m \in c_f} N_m R_m}{\sum_{m \in c_f} N_m}, \quad f = 1, \Lambda\ , \Phi \qquad (16)$$

$N_i, i = 1,2\Lambda\ \Lambda\ I$ is the number of vectors to estimate each Gaussian covariance in training SI model. I is the total number of HMM states.

In the next level, if the nodes of last level have enough Gaussian-distributed variables, it splits into two. If not, the node is noted as a leaf node. Keeping on this process, until all leaf nodes are found. Then for all nodes, we obtain the transformation matrix from cluster center to all Gaussian of HMM states with formula (8).

We have constructed a binary decision tree, which is shown in Figure 2.



Choose node depending on the corresponding vectors number

Figure 2. The Binary Decision Tree Based on GSA

3.2 Speaker Adaptation Based on GSA

When a small amount of speaker-specific speech is input, which serve as the adaptation data, the adaptation algorithm is as follows:

From SI models and adaptation data, calculate the mean vector of SD model with MLLR.

Find the segmentation of the input data with the mean vector of SD models, then denote each vectors which state it belongs to.

Decide the clusters for adaptation according to the number of vectors. It means, counting the number of vectors in each leaf node, if the number is larger than the threshold predefined, this node can be used as the adaptation cluster; if not, we should trace back to its father-node, until we can find a node has enough vectors. The dot line in Figure 1 is an demonstration. This step is to make it sure that we can have enough data to get a reliable covariance matrix of the cluster center.

After we have found all the clusters for adaptation, the covariance matrix R_f $(f = 1,2\Lambda\ \Lambda\ \Phi)$ can be estimated by all the vectors belong to this cluster.

With the covariance matrix R_f of cluster center and transformation matrix $A_{j,i}$, we can obtain all the adapted covariance matrices R_i: $R_i = A_{f,i} R_f A_{f,i}^t$. They are the covariance matrices of SD Model.

Following these steps, we can get a SD model with mean vector and covariance matrix both adapted. This speaker adaptation algorithm can dynamically changed the clusters according to the adaptation data, and it is convenient to be used in supervised and unsupervised adaptation.

4. Experiment Results and Discussion

The proposed speaker adaptation scheme has been implemented in THEESP Mandarin speech recognition system. In this system there are totally 856 states, each being modeled as a single Gaussian with full covariance. The signal observation vectors are MFCC, the first and second delta features are both included. The experiments are carried out on a large-vocabulary task based on *people' s daily,* the training and testing corpus is both provided by National 863 High Technology Project.

The test data comprises of 250 sentences each from 12 speakers. Another 100 sentences of corresponding speaker are used as the adaptation data. Supervised batch adaptation is used in the following tests.

Speaker	SI Model	Mean Adaptation	Mean and Covariance Adaptation
m80	28.86	23.47	21.46
m81	30.35	21.55	20.45
m82	31.26	23.04	17.15
m83	28.45	22.45	18.17
m84	40.24	23.61	19.66
m93	23.40	21.5	18.5
m94	27.34	23.64	20.84

m95	24.16	17.1	14.48
m96	40.84	34.28	31.08
m97	37.45	24.45	27.20
m98	23.46	18.74	13.83
m99	26.19	23.51	19.32
Average	30.17	23.11	20.18

Table 1 the error rate with different adaptation schemes.

In the first test, we assess the performance of different speaker adaptation scheme. Table 1 gives the error rate of SI model, mean adaptation with MLLR and covariance adaptation based on GSA. As can be seen in Table 1, the mean vector adaptation with MLLR decrease the error rate 23.39% relatively, and the covariance adaptation based on GSA get a further reduction 12.69% relatively. With only 100 sentences, about 5 minutes speaker-specific speech, we successfully achieve the Gaussian-covariance adaptation, which greatly improves the performance of recognition system.

In the second test, we try to find a proper threshold to decide the clusters for adaptation, which can assure that each cluster has enough observation vectors to estimate covariance matrix. In adaptation process, the total number of vectors is limited. If the threshold is too small, we don’t have enough data to estimate covariance matrix of cluster center. When the threshold increasing, because each cluster should have enough vectors over this threshold, the cluster number will decrease. It will affect the performance of the adaptation algorithm. As can be seen from Figure 2, when the threshold is very small at first part, the word error rate keeps falling, because the estimated covariance of cluster center should be more reliable with the increasing of threshold. In the next part, because we already have almost enough vectors to estimate covariance matrix, the reduction of cluster number let the error rate rise. And then the curve descends and rises for another time but in much smaller amplitude. This curve explains the relationship between the error rate and threshold vividly. We must find a threshold to make best trade-off between the reliability of covariance estimation and more cluster number as possible. So we can get best performance with the limit amounts of adaptation data.

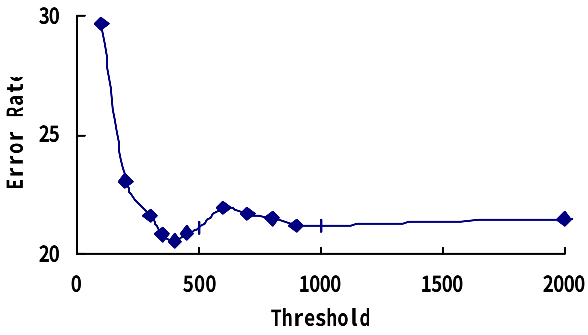


Figure 2 the preformance with different threshold.

4. Conclusion

The above GSA based speaker adaptation algorithm focuses on the Gaussian-covariance adaptation, which is different from the traditional adaptation schemes. It achieves covariance adaptation with quite limit adaptation data. Also, the online computation cost is small because the most time-consuming part of the algorithm, constructing a binary decision tree can be done off-line. Further research emphasis hope to be put on how to build the decision tree making use of phonetic knowledge to improve the performance and robustness of this adaptation scheme.

Reference:

[1] M.Padmanabhan, L.R.Baul, and M.A.Picheny, “Speaker Clustering and Transformation for Speaker Adaptation in Large-Vocabulary Speech Recognition Systems”, Proc. ICASSP-96.

[2] J.L.Gauvain and C.H.Lee, “Maximum a Posteriori Estimation for Multivariate Gaussian Observations of Markov Chains”, IEEE Trans. SAP, vol.2, no.2, pp291-298, Apr.1994

[3] C.J.Legetter and P.C.Woodland, “Maximum Likelihood Linear Regression for Speaker Adaptation of Continuous Density HMM’s”, Computer Speech and Language, vol.9, no.2, pp171-186.

[4] Zuoying Wang and Feng Liu, “Speaker Adaptation Using Maximum Likelihood Model Interpolation”, Proc, ICASSP-99

[5] Feng Liu, “Speaker Adaptation in Large-Vocabulary Continuous Speech Recognition for Chinese”, Thesis of Post-Doctor, Tsinghua Univ, Jan.2000(in Chinese)

[6] Zuoying Wang, “Inhomogeneous HMM for Speech Recognition and THED Recognition and Understanding System”, Telecommunication Science, Vol.9, No.4, July 1993, pp31-36(in Chinese)