Using Spatial Correlation Information in Speech Recognition

Yu Peng, Wang Zuoying

Department of Electronic Engineering
Tsinghua University, Beijing 100084, P.R.China
Email: yupeng@thsp.ee.tsinghua.edu.cn

Abstract

Acoustic model training is very important in speech recognition. But in traditional training algorithm, we take each state separately, and the relationship between different states is not considered. In this paper we bring forward a novel idea of using the correlation information between states, which is called "spatial correlation". We describe this correlation information as linear constraints. According to phonetic knowledge, we firstly divide states into small groups named "correlation sub-space". In every sub-space, we use eigen value decomposition to get linear constraints. The constraints are then used in a new training algorithm. Experiments of the new training algorithm show significant improvement over traditional training algorithm.

1. Introduction

In a Hidden Markov Model based speech recognition system, the HMM parameters of one speaker represent the acoustic property of this speaker. If we take each speaker as one sample point in a probability space, then the HMM parameters are themselves random variables. The probability distributions of these parameters describe the distribution of each state in feature space, so we name these distributions "spatial structure".

Information of spatial structure provides priori knowledge for HMM parameter estimation. This kind of information has already been used in speaker adaptation algorithm. For example, in MAP[1] method, speaker-independent model parameters are used as priori knowledge. Another case is the MLMI[2] method, where more information is used: the model parameters of a set of speakers are used as priori knowledge.

However, in these cases, the information about correlation between different states is not explicitly used. As we know, the parameters of different states are not independent. We call the correlation between different states "spatial correlation". Sometimes this kind of information is more useful, especially when some states are absent from the training data. In [3], Scott tempted to use this information in adaptation. He predicts one state from another state by the correlation coefficient. In this paper, we will discuss spatial correlation in a different way. We describe spatial correlation as constraints of HMM parameters. We develop methods for finding the constraints and using the constraints. Experiments show that applying the spatial correlation in training algorithm gain performance improvement over traditional training algorithm.

The paper is organized as follows. In Section 2, we give the method of discovering the spatial correlation. In Section 3, we propose a new training algorithm applying this correlation. Some experimental results are shown in Section 4. Finally, we summarize our findings in Section 5.

2. Discovering the spatial correlation

In a speech recognition system base on Continuous Density Hidden Markov Model, model parameters include the mean vector and the covariance matrix for all states. If we focus on the mean vector, then the model parameters can be regarded as a large vector accumulated by mean vectors of all states. Assume the state number is NS, the dimension of feature vector is NF, then the HMM parameters can be represented

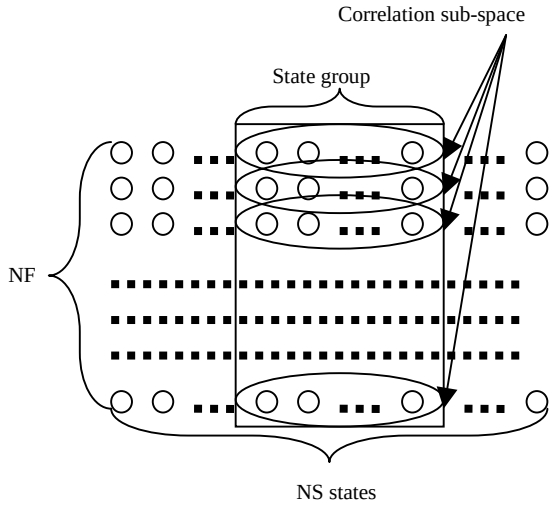

Figure 1 codebook space, state group and correlation subspace

by an NS×NF dimensional vector. Usually we call this vector a "codebook", denote it as $\vec{C}$. And we call the space of codebooks the "codebook space", this is a NS×NF dimensional space, denote it as $R_{NS\times NF}$.

We represent spatial correlation as some constraints of codebooks. In mathematics, it can be described by equations, $SC_n(\vec{C}) = 0, n = 0,1,...,NT$, where NT represents the number of constraints. Currently, we assume these constraints are linear,

$$SC_i = (\vec{C}, A_i) - b_i = \sum_{s=1}^{NS\cdot NF} c_s \cdot a_{is} - b_i = 0, i = 1,2,...,NT \quad (1)$$

Here $(\cdot,\cdot)$ represents the inner product in space $R_{NS\times NF}$.

We will discover the constraints from a given set of codebooks. Firstly we choose a set of speakers, and denote the set of their codebooks as Ψ. Our goal is to find a group of constraints so that all the codebooks in Ψ satisfy these constraints. As we have seen, a linear constraint is represented by a (NS×NF+1) dimensional vector $(\vec{A}_i, b_i)$. To estimate such a high dimensional vector, we will have to get a tremendous amount of training data. So we divide the states into groups to reduce the dimension of constraints.

Our grouping method are made up of two steps,

1. We suppose that spatial correlation only exists among similar phonetics. So firstly we divide the states into small groups according to phonetic knowledge. Assume the number of group is NG and each group has a state number of $NS^i, i = 1,2...,\ NG$.
2. We suppose the correlation between different states only exists in the same dimension of the mean vectors of these states. Then one state group is further divided into NF smaller groups.

By these two steps, the codebook space can be divided into NG×NF orthogonal product subspaces.

$$R_{NS\times NF} = \overbrace{\underbrace{R^1_{NS^1}\times R^2_{NS^1}\times...\times R^{NF}_{NS^1}}_{NF}\times...\times\underbrace{R^1_{NS^{NG}}\times...\times R^{NF}_{NS^{NG}}}_{NF}}^{NG} \quad (2)$$

We call every subspace a "correlation subspace". The grouping method is illustrated in figure 1.

We restrict the spatial constraints into one correlation sub-space, and then we can represent a constraint by a (NS^i+1) dimensional vector. We have sufficient data to estimate such a vector robustly.

Now what we need to do is to find a group of $(\vec{A}_i, b_i)$, such that for every $\vec{C}\in\Psi, (\vec{C},\vec{A}_i) = b_i$. This can be done by eigen value decomposition.

In every correlation sub-space, denote

$$\mathbf{R} = \sum_{k=1}^{NC} (\vec{C}'_k - \overline{\vec{C}'})\cdot(\vec{C}'_k - \overline{\vec{C}'})^T / NC \quad (3)$$

as the covariance matrix of vectors in Ψ, where NC is the number vectors in Ψ, $\vec{C}'_k$ represents the truncation of *k*th vector $\vec{C}_k$ on the current correlation sub space, $\overline{\vec{C}'} = \sum_{k=1}^{NC}\vec{C}'_k / NC$ represents the mean of truncation vectors. Here truncation means get several dimensions from a large vector to make up a small vector.

Do eigen value decomposition for **R**

$$\begin{aligned}\mathbf{R} &= \mathbf{E}\cdot\mathbf{\Sigma}\cdot\mathbf{E}^T \\ &= (\vec{e}_1,\vec{e}_2,\cdots,\vec{e}_{NS^i})\cdot diag(\sigma_1,\sigma_2,\cdots,\sigma_{NS^i})\cdot(\vec{e}_1,\vec{e}_2,\cdots,\vec{e}_{NS^i})^T\end{aligned} \quad (4)$$

Where σ_j is the eigen value of **R**, $\vec{e}_j$ is the corresponding eigen vector.

For $\vec{C}' \in \Psi$, define $\vec{C}^* = \vec{C}' - \overline{\vec{C}}'$. If one eigen value σ_j approaches to zero, then any vector $\vec{C}^*$ is almost orthogonal to the corresponding eigen vector $\vec{e}_j$. So we sort the eigen values and select NM^i smallest values, assume they are $\sigma_{NS^i-NM^i+1}, \sigma_{NS^i-NM^i+2}, \cdots, \sigma_{NS^i}$, which satisfies $\sum_{j=NS^i-NM^i+1}^{NS^i} \sigma_j < (1-\lambda) \cdot \sum_{j=1}^{NS^i} \sigma_j$. Here λ is a threshold parameter approaches to 1. Since the eigen values are small enough, we have

$$(\vec{C}' - \overline{\vec{C}'}) \perp span(\vec{e}_{NS^i-NM^i+1}, \vec{e}_{NS^i-NM^i+2}, \cdots, \vec{e}_{NS^i}) \quad (5)$$

Those are linear constraints of $\vec{C}'$.

$$\begin{aligned} \sum_{j=1}^{NS^i} c'_i \cdot e_{NS^i-k,j} &= (\vec{C}', \vec{e}_{NS^i-k}) = (\overline{\vec{C}'}, \vec{e}_{NS^i-k}) \\ &= \sum_{j=1}^{NS^i} \overline{c'_j} \cdot e_{NS^i-k,j}, k = 0,1,...,NM^i - 1 \end{aligned} \quad (6)$$

Since $\vec{C}'$ is a truncation of $\vec{C}$, so the above is also a group of constraints of $\vec{C}$. We can gather the constraints from all the correlation sub-spaces, then get the final constraints for $\vec{C}$.

3. Training algorithm with spatial correlation

The spatial constraints can easily be integrated into the Maximum Likelihood training algorithm, as shown in the following formula.

$$\begin{cases} \hat{\vec{C}} = \arg\max(P(S \mid \vec{C})) \\ s.t. \quad \mathrm{SC}_n(\vec{C}) = 0, n = 1,2,..., NT \end{cases} \quad (7)$$

Where S represents the feature vectors of training corpus, NT is the number of spatial correlation constrains.

To focus on the spatial correlation information, we use a segmental *k*-means training procedure. Then equation (7) can be written as,

$$\begin{cases} \hat{\vec{C}} = \arg\min(\sum_{i=1}^{NS} \sum_{j=1}^{NF} \sum_{k=1}^{N_i} (c_{i \times NF + j} - s_{k,j}^i)^2) \\ s.t. \ \mathrm{SC}_n(\vec{C}) = \sum_{s=1}^{NS \cdot NF} c_s \cdot a_{n,s} - b_n = 0, n = 1,2,..., NT \end{cases} \quad (8)$$

Where $c_{i \times NF + j}$ represents the *j*th dimension of the *i*th state in the codebook vector $\vec{C}$, $s_{k,j}^i$ represents the *j*th dimension of the *k*th training vector that corresponds to the *i*th state, N_i represents the number of training vectors correspond to the *i*th state.

This is an optimization problem of a convex functional on a convex set. According to optimization theory, it has unique solution,

$$\begin{aligned} &\text{Denote} \bar{s}_{i,j} = \sum_{k=1}^{N_i} s_{k,j}^i / N_i, i = 1,2,...,NS, j = 1,2,...,NF \\ &v_{n,m} = \sum_{s=1}^{NS \cdot NF} a_{m,s} \cdot a_{n,s} / N_{[s/NF]}, n, m = 1,2,...,NT \\ &w_n = b_n - \sum_{i=1}^{NS} \sum_{j=1}^{NF} a_{n,i \times NF + j} \cdot \bar{s}_{i,j}, n = 1,2,...,NT \\ &U = \mathbf{V}^{-1} \cdot W, \\ &c_{i \times NF + j} = \bar{s}_{i,j} + \sum_{n=1}^{NT} u_n \cdot a_{n,i \times NF + j} / N_i \end{aligned} \quad (9)$$

In the final result, $\bar{s}_{i,j}$ is the codebook estimation without spatial correlation, and $\sum_{n=1}^{NT} u_n \cdot a_{n,i \times NF + j} / N_i$ is the revision of spatial correlation.

In the final solution, we must calculate the inverse of matrix **V**, which is NT×NT dimensional. We do not need to calculate it directly, because V is a block diagonal matrix made up of small matrices. Every small matrix is $NM^i \times NM^i$ dimensional.

4. Experimental results

The following experiments are implemented in Mandarin. The database is provided by Chinese National 863 High-Tech project. It comprises 77 male speakers. Each speaker has 600 sentences. The speech recognition system is based on a modified HMM model named DDBHMM [4]. Every state is described by a single-Gaussian distribution with full covariance matrix. Totally there are 857 states. The feature is combined with 14 dimensions MFCC, 1 dimension energy and their first and second order differences, totally 45

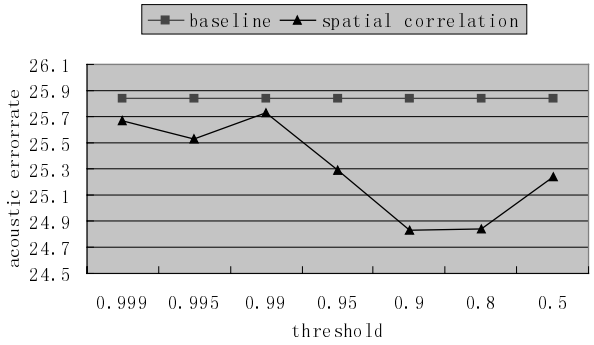

Figure 2 result of spatial correlation experiment

dimensions. In the experiments, 70 speakers are chosen for finding correlation constrains (used as set Ψ); the other 7 speakers are used for testing. For the testing speaker, the first 100 sentences are used to train speaker dependent model parameters, and the last 200 sentences are used for recognition. All the results shown below are acoustic syllable error rate, including insertion error, deletion error and substitution error.

All states are divided into 41 groups, forming 1845 correlation sub-spaces. Choose λ as different values, the results are shown in figure2.

We can see, when λ decreases from 1.0, the constraint number increases, and the spatial correlation shows more effect to the training performance, so the error rate goes down. However, when λ is too small, the accuracy of the constraints decreases, the spatial constraints will bring negative effect, shown in the figure, the error rate increases when λ is less then 0.8. The optimal value of λ is near 0.9, where 3.91% error rate reduction is obtained.

5. Summarization

In this paper, we discuss the application of spatial correlation in speech recognition training task. We describe the spatial correlation as linear constraints of the codebook. In order to get robust constraints, we use phonetic knowledge to divide the states into groups, and use eigen value decomposition to discover the constraints. With the spatial correlation, a new training algorithm is proposed, which outperform the original method by 3.91% error rate reduction. It can be imagined that the spatial constraints will also be useful in speaker adaptation algorithm.

In the paper, spatial constraints are assumed to be linear. It is only an approximation, there may be other more efficient choices, and it will be our next research interest.

6. Reference

[1] Chin-Hui Lee, Chin-Heng Lin and Bing-Hwang Huang, “A Study on Speaker Adaptation of the Parameters of Continuous Density Hidden Markov Models”, IEEE Trans. On Signal Processing, 1991

[2] Wang Zuoying, Liu Feng, “Speaker adaptation using maximum likelihood model interpolation”, Proceeding of ICASSP, 1999

[3] Scott Shaobing Chen, Peter DeSouza, “Speaker Adaptatiton By Correlation”, Proceeding of Hub4 Workshop, 1997

[4] Wang Zuoying, “ Duration Distribution Based HMM speech recognition model”, The 2nd national Chinese character and speech recognition conference, 1989.9