The Description of the Intra-State Feature Space in Speech Recognition

Fang Zheng, Mingxing Xu, Wenhu Wu
Speech Lab, Dept. of Comp. Sci. & Tech., Tsinghua Univ., Beijing 100084, China
fzheng@cenpok.net, [fzheng, xumx]@sp.cs.tsinghua.edu.cn

Abstract

In speech recognition, the description of the intra-state feature space is an important issue in systems based on HMM-derived acoustic models. The existing techniques include the famous methods based on VQ technique and mixture Gaussian densities. In this paper, a method based on sub-space division is proposed. Experiments are done to find how many densities should be used to better describe the intra-state feature space, and the experimental results show that the number of densities should depend on the particular distribution of that space and can be judged by a kind of criterion.

1. Introduction

In speech recognition, how to describe or represent the intra-state feature space for a HMM-based system is an important problem.

In general, there is an assumptions for traditional HMM, i.e., the current observation depends only on the current system state, which indicates that the observation output of the intra-state is independent and identical distributed (***i.i.d.***). Though it is simple, the results are not bad, which introduces discrete HMM (DHMM), continuous density HMM (CDHMM), semi-continuous HMM (SCHMM) [Huang 1989] and some other similar models.

According to the above assumption, the intra-state feature space can be described according to the Theory of Probability. The common used approaches include Mixture Gaussian Densities (**MGD**) [Wilpon 1989, Huang 1989] and tied Mixture Gaussian Densities (**TMGD**) [Bellegarda 1990]. Zheng *et al* [Zheng 1996, 1997] describes the feature space by sub-space division (**SSD**) method.

No matter what kind of method is adopted, there is an important problem to solve, that is, how many probability density functions (**PDFs**) should be used to better describe the space when the maximum number of densities is limited ? According to the Theory of

Probability, the more Gaussian densities, the better the space is approached. Due to the limitation of computer processing, a suitable number of densities must be determined.

In that case, should the same density number be chosen for every speech recognition unit (**SRU**) ? The answer is no. In this paper, the experimental results will be given based on the CDCPM [Zheng 1996].

2. Sub-Space Division Method

The sub-space division (**SSD**) method is to divide the whole space into several independent sub-space by a certain criterion. If considering the probability distribution of a specified space, the scoring of a feature vector o_t at time t is given by

$$b(o_t) = \sum_{m=1}^{M} g_m f_m(o_t) \tag{1}$$

where $b(\cdot)$ is the PDF of the whole space, $f_m(\cdot)$'s are the M Gaussian PDFs, and g_m's are the corresponding weightings of the PDFs. If using the SSD method, the scoring equation is

$$b(o_t) = \max_{1 \le m \le M} f_m(o_t) \tag{2}$$

(Notice that $f_m(\cdot)$'s in Eq. (2) are often different from those in Eq. (1).) That is to say, the score of a feature vector is defined as the matching score with the closest sub-space. We found this kind of scoring scheme is identical to the human's cognition.

2.1 Considerations in the SSD method

In general, when dividing the space into sub-spaces in speaker-independent speech recognition tasks, the following factors will be considered: (1) gender-dependent (GD) information; (2)accent-dependent (AD) information; (3) speaker-dependent (SD) Information; (4) context-dependent (CD) information; (5) background noise (BN) information; and so on.

All these factors make the situation more complicated, it will cost much more model storage and database labeling even if only one or two of these factors are considered. It sounds not practical.

2.2 The motivation of the automatic SSD (ASSD) method

Actually, the feature space of the same SRU uttered by different speakers in different accents and different contexts has many common areas, as illustrated in Fig. 1.

In Fig. 1, symbols 1, 2, and 3 stand for three different speech sources, any two of them have a common region in the whole space. (It maybe more complicated actually.) Obviously, if we divide the feature spaces individually, then 4 densities are needed for both sources 1 and 3 while 3 densities for source 2, and totally 11 densities are needed. But actually, there are only 6 different densities. So if there are many speech sources, a great deal of redundant density storage are cost for individual space division and description.

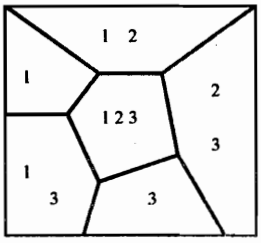

Fig. 1 The feature space of three different speech sources

The automatic SSD (ASSD) method tries to find as fewer common regions as possible for the whole feature space of several different speech sources.

2.3 Clustering: one ASSD method

As a matter of fact, many existing clustering methods can be used for sub-space division, such as LBG [Linde 1980], K-means [Furui 1989], and simulated anneal [Xu 1989] algorithms. But it is not too easy to decide how many densities should be used. A criterion will be given in this paper based on within- and between-class scatter degrees.

First of all, we will define the within-class scatter degree (**WCSD**) and between-class scatter degree (**BCSD**).

Assume there are N feature vectors totally in a space. In M'th clustering iteration, M classes (sub-spaces) are generated, and in Class m $(1 \le m \le M)$, the mean vector of the N_m vectors $\{x_m^{(i)}: 1 \le i \le N_m\}$ is μ_m. Denote the distance measure between two vectors by $y(\cdot,\cdot)$. Define average within-class scatter degree of Class m as

$$\tilde{d}_{wm} = \frac{1}{N_m}\sum_{i=1}^{N_m} y(x_m^{(i)}, \mu_m) \tag{3}$$

and the total average within-class scatter degree as

$$\tilde{d}_w = \frac{1}{N}\sum_{m=1}^{M} N_m \cdot \tilde{d}_{wm} = \frac{1}{N}\sum_{m=1}^{M}\sum_{i=1}^{N_m} y(x_m^{(i)}, \mu_m). \tag{4}$$

Define the average between-class scatter degree as

$$\tilde{d}_b = \frac{1}{M\times(M-1)}\sum_{m=1}^{M}\sum_{\substack{m1=1 \\ m1\neq m}}^{M} y(\mu_m, \mu_{m1}) = \frac{2}{M\times(M-1)}\sum_{m=1}^{M}\sum_{m1=m+1}^{M} y(\mu_m, \mu_{m1}) \tag{5}$$

After the definitions of the within-class and between-class scatter degrees, the criterion

function is define as

$$J_d(M) = \frac{\tilde{d}_w \cdot f(M)}{\tilde{d}_b} \tag{6}$$

where $f(\cdot)$ is a strictly increasing function.

In general, $\tilde{d}_w / \tilde{d}_b$ is a decreasing function of M. Consider the special case when M is equal to the number of all feature vectors, where there is only one vector in every class, so $\tilde{d}_w / \tilde{d}_b$ will decrease to 0. The increasing function $f(\cdot)$ is used as a penalty function to avoid unreasonable larger class number.

In the speech recognition system, a maximum value of class number or sub-space number is often given at the very beginning, say $M_{\max}$. So in this case,

$$M_s = \arg\min_{2 \le M \le M_{\max}} J_d(M) \tag{7}$$

is chosen as the suitable class (sub-space) number.

3. Database Description

A great deal of experiments have been done across a real-world spontaneous database. The speech data are taken from telephone network and sampled at 8KHz. The samples are 13-bit linear PCMs expanded from A-law codes. The database consists of speech data uttered by 200 people, and the amount is about 4GB. 10^{th} order LPC-based cepstral (LPCC) analysis is performed on 32 ms speech window every 16 ms. Auto-regressive analysis is also performed on 5 adjacent frames of LPCC vectors. The LPCCs and their corresponding auto-regressive coefficients are the features used for the CDCPM [Zheng 1996, 1997] in this paper.

The SRUs are 419 Chinese syllables.

4. Experimental Results

The experimental results are given in Tab.1, where the number of states (NOS) ranges from 3 to 6. In the number of densities (NOD) column, "Fix" stands for using the maximum number of densities, i.e., 16, while "Var" stands for using different number of densities for different syllables, but the maximum density number is 16.

From Tab. 1, we can draw the conclusion for any number of states for the CDCPM: choosing different density number for different syllables is better than using fixed density number, the former scheme can improve the system by 1.6% when NOS is 6.

Tab.1 Experiment on choosing the intra-state number of densities

NOS \ Top *n* candidates / NOD		1	2	3	4	5	6	7	8	9	10
3	Fix	69.44	77.94	82.05	84.70	86.51	87.95	89.14	90.04	90.72	91.39
	Var	69.95	78.47	82.59	85.29	87.11	88.54	89.75	90.64	91.34	92.02
4	Fix	74.03	81.42	85.03	87.34	88.95	90.02	90.96	91.79	92.49	92.99
	Var	74.90	82.32	85.96	88.24	89.87	90.94	91.84	92.68	93.41	93.92
5	Fix	77.41	83.79	86.68	88.44	89.75	90.85	91.56	92.20	92.79	93.39
	Var	78.66	85.07	87.95	89.73	91.04	92.13	92.85	93.48	94.08	94.67
6	Fix	78.86	85.12	87.84	89.40	90.71	91.41	92.12	92.75	93.27	93.67
	Var	80.46	86.71	89.45	91.03	92.32	93.03	93.77	94.40	94.91	95.31

5. Conclusion

In this paper, we study the description methods for intra-state feature space based on HMM-derived acoustic models. The experiments are done for choosing the suitable number of sub-spaces. The experimental results show that using the same number of densities to describe every state of every SRU performs worse, different state of different syllable should have different density number according to a reasonable criterion. Studies and experiments are focused on the SSD scheme, we think the conclusion is also right for the MGD method.

6. References

[1] **Bellegarda, J.R., Nahamoo, D.,** "Tied Mixture Continuous Parameter Modeling for Speech Recognition," *IEEE Trans. on ASSP*, vol.ASSP-38, No.12, pp.2033-2045, Nov. 1990

[2] **Linde, Y., Buzo, A., Gray, R.M.,** "An algorithm for vector quantization," *IEEE Trans. On COM*, 28(1), Jan. 1980

[3] **Furui, S.,** Digital Speech Processing, Synthesis and Recognition, *Marcel Dekker, Inc.*, 1989

[4] **Huang, X.-D., Jack, M. A.,** "Semi-continuous hidden Markov models for speech signals," *Computer Speech and Language,* 3:239-251, 1989.

[5] **Wilpon, J.G., Lee, C.-H., Rabiner, L.R.,** "Application of hidden Markov models for recognition of a limited set of words in unconstrained speech," *ICASSP-89*, 3: 254-257

[6] **Xu, L.,** "A kind of new clustering method: Simulated Anneal," *Pattern Recognition and Artificial Intelligence*, 2 (1), March 1989 (in Chinese)

[7] **Zheng, F., Wu, W.-H., Fang, D.-T.,** "CDCPM with its applications to speech recognition," *Chinese J. of Software*, 7: 69-75, Oct. 1996 (in Chinese)

[8] **Zheng, F., Chai, H.-X., Shi, Z.-J., Wu, W.-H., Fang, D.-T.,** "A real-world speech recognition system based on CDCPMs," *Int'l Conf. on Computer Processing of Oriental Languages (ICCPOL'97)*, Apr. 2-4, 1997, Hong Kong