音声言語情報処理　29－12
（1999. 12. 20）

音素間相互情報を利用した音素ペアモデルによる話者適応

李 宝潔 † 　　広瀬啓吉 ‡
† 東京大学大学院・工学系研究科
‡ 東京大学大学院・新領域創成科学研究科
〒 113-8656 東京都文京区本郷 7 丁目 3 番 1 号
lbj@gavo.t.u-tokyo.ac.jp, hirose@gavo.t.u-tokyo.ac.jp

あらまし　異なる話者が発話した場合、同一音素は特徴空間内で広い範囲で分布する。同一話者が発話した場合には集中するが、なお相当の広がりが見られる。一方、同一話者が発話した各音素の特徴空間での相対位置についてみれば、よりよい安定性があると思われる。我々が以前に提案した音素ペアモデルは、二つ音素の特徴ベクトルの共起確率を利用してこのような音素間の相互関係を記述するものである。本論文は音素ペアモデルを音素 HMM に組み込んだ不限定話者認識タスクの評価実験結果を報告する。認識結果として、単語の正解率と精度の両者とも音素 HMM のみの場合よりも大幅に上昇した。

キーワード　音声認識, 頑健, 音素相関, 話者適応, ペアモデル, 認識システム

Application of Phone Pair Model to Robust Speech Recognition

Baojie Li[†] 　　Keikichi Hirose[‡]
[†]School of Engineering, University of Tokyo
[‡]School of Frontier Sciences,University of Tokyo
〒 113-8656 University of Tokyo, 7-3-1 hongo, Bunkyo-ku, Tokyo, 113-8656,Japan
lbj@gavo.t.u-tokyo.ac.jp, hirose@gavo.t.u-tokyo.ac.jp

Abstract　In the acoustic feature space, all phone positions suffer rather large shifts across speakers. Even within a speaker they are not so much stable due to utterance-to-utterance variations. However, their relative positions are considered to be rather stable for a speaker. In previous works, we proposed a Phone Pair Model which utilizes the joint probability of two phones' acoustic feature vectors to catch these inter-speaker tied movements. In this paper, we report the results of our comparative experiments between one using solely phone HMMs and the other uses phone HMMs combined with Phone Pair Model. Remarkable increases in both *Correct rate* and *Accuracy* were achieved.

key words　speech recognition, robust, correlation, speaker adaptation, pair model, recognition system

1 Introduction

Speaker adaptation techniques have been broadly studied [2][3][4]. But most of them suffer from insufficient adaptation data. Considering the fact that all the phones uttered by a speaker, are commonly influenced by the individuality of that speaker, correlations do exist among phones. If these phone correlations can be used effectively, we will be able to greatly enhance the adaptation process. Hazen [5] and Zhang *et al.* [6] did beneficial attempts to exploit correlations among phones, but several limitations prevented them from being applied to practice.

After the success in describing the correlations among phones with Phone Pair Model (PPM)[1], we continued to investigate the effect of the PPM on speech recognition systems. In this paper, we will report the comparative experiments for a speaker-independent recognition task, between one using solely phone HMMs and the other combining with PPM. After a brief review of the theory of PPM in section 2, the way of combining it with HTK is introduced in section 3 . The results of the recognition experiments are reported in section 4, which are followed by section 5, the discussions, and section 6 concludes this paper.

2 Phone Pair Model

When we have some phones known in the decoding stage, we can determine the unknown phones based on the probability calculated on the known-unknown phone pairs. For example, when $\boldsymbol{Y} = \boldsymbol{y}_1, \boldsymbol{y}_2, \cdots, \boldsymbol{y}_{Ty}$ is an observation sequence for $phone_Y$ and $\boldsymbol{X} = \boldsymbol{x}_1, \boldsymbol{x}_2, \cdots, \boldsymbol{x}_{Tx}$ for $phone_X$, the joint probability of $(phone_X, phone_Y)$ can be given approximately by

$$P(\boldsymbol{X}, \boldsymbol{Y}) \approx \prod_{i,j} (p(\overline{\boldsymbol{x}_{si}}, \overline{\boldsymbol{y}_{sj}})) \tag{1}$$

where $\overline{\boldsymbol{x}_{si}}$ is the average of vectors aligned into state s_i of the HMM of $phone_X$, and $\overline{\boldsymbol{y}_{sj}}$ the average of vectors aligned into state s_j of the HMM of $phone_Y$. Let variable vector $\overline{X_{si}}$ has the distribution $N(\boldsymbol{\mu}_{si}, \boldsymbol{\Sigma}_{si})$ and $\overline{Y_{si}}$ has $N(\boldsymbol{\mu}_{sj}, \boldsymbol{\Sigma}_{sj})$, Then the distribution of joint vector $(\overline{X_{si}}, \overline{Y_{sj}})$ is $N(\boldsymbol{\mu}_{ij}, \boldsymbol{\Sigma}_{ij})$ in which

$$\boldsymbol{\mu}_{ij} = \begin{bmatrix} \boldsymbol{\mu}_{si} \\ \boldsymbol{\mu}_{sj} \end{bmatrix}$$

is the mean of joint vectors $(\overline{X_{si}}, \overline{Y_{sj}})$, and

$$\boldsymbol{\Sigma}_{ij} = \begin{bmatrix} \boldsymbol{\Sigma}_{si.si} & \boldsymbol{\Sigma}_{si.sj} \\ \boldsymbol{\Sigma}_{sj.si} & \boldsymbol{\Sigma}_{sj.sj} \end{bmatrix} \tag{2}$$

the covariance. When the four sub-matrix of $\boldsymbol{\Sigma}_{ij}$ are assumed diagonal, $\boldsymbol{\Sigma}_{sj.si}$ is equal to $\boldsymbol{\Sigma}_{si.sj}$

3 Combination of PPM with HTK

In HTK[7], each word is represented as a sequence of phone HMMs. In Fig.1, the square boxes represent work-end node, and the circles denote HMMs of phones that compose word.

Token Passing Model is used as an alternative formulation of Viterbi algorithm in HTK. Each state of a HMM holds a movable token which contains the partial log likelihood score of the rout it has passed through. When an observation from the input sentence is processed, every token is copied to all connecting states with the log likelihood score increased and all the tokens are discarded except the highest one. When a token is propagated out from the exit state of a word, the boundary for this word is recorded.

In our experiments, each HMM has 3 states (except the entry and the exit states). For simplicity, our current PPM only exploits state 2. Before recognition, we align the given adaptation sentence from the new speaker with Speaker-Independent (SI) HMMs. Despite the boundary of each phone may includes errors to some extent, we can extract phones we are interested in (in the current method, the five Japanese vowels $\{a, i, u, e, o\}$) with enough accuracy. Then we average the vectors of state 2 for each vowel and get vectors $\{v_a, v_i, v_u, v_e, v_o\}$ for the 5 vowels $\{a, i, u, e, o\}$ respectively.

When a token reached a word-end node, the boundary for each phone composing this word can be easily known. For example, assume a Japanese word "*ni* (meaning *two* in English)" has a boundary of frame (n_1, n_N) for phone n and frame (i_1, i_I) for phone i, corresponding to observations (o_{n1}, o_{nN}) and (o_{i1}, o_{iI}). We construct vector pairs (o_{nk}, v_a), (o_{nk}, v_i), (o_{nk}, v_u), (o_{nk}, v_e), (o_{nk}, v_o) where $k = 1...n_N$, then calculate the logarithmic probability density of the vector pairs generated by PPMs $M_{n-a}, M_{n-i}, M_{n-u}, M_{n-e}, M_{n-o}$, respectively. Thus a score P_n^{pair}, which is the average on k and the 5 vowels, is obtained for phone n. Also we can get P_i^{pair} for phone i in the same way.

Since each word is composed with different number of phones, we average the scores of all the phones composing a word to assure that the PPM contributes to every word equally, no matter what the length of the word is, short or long.

If the partial path till current word has a likelihood score ψ (which is calculated in the conventional way), we add the PPM score of this word to ψ and get the modified ψ^{mod} score as

$$\psi^{mod} = \psi + \frac{1}{M}\sum_{m=1}^{M}(\frac{1}{m_N}\sum_{j=1}^{m_N} P_{o_j}^{pair}) \qquad (3)$$

, where $P_{o_j}^{pair}$ is the average score of frame j on vowel set $\{a, i, u, e, o\}$, m_N the number of frames aligned into *state* 2 of $phone_m$ and M the number of phones composing this word.

For an observation o_x of *state*2 of $phone_x$, the score of vector pair (o_x, v_z) generated by pair model M_{x-z} would be ideally higher than that generated by M_{y-z} where $phone_y$ is any phone except $phone_x$. See Fig.1 for further explanations.

4 Experiment implementation

We designed a SI recognition task to test our PPM. As many other adaptation algorithms, we need a fragment of an utterance from the new speaker. We use HTK(Ver.2.1.1) to train SI models. The recognition experiments are done comparatively using the original HTK and the modified HTK, which is integrated with PPM.

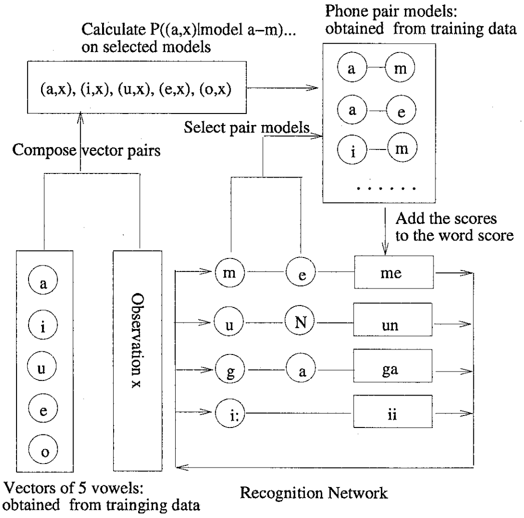

Fig 1. Diagram of combining PPM with HTK

A. Training SI HMMs We use ATR *Continuous Speech Corpus for Research* as our training and test data. 150 sentences of each of 6 male speakers (can0001, can0002, ecl0001, ecl0002, mit0001, mit0002) are used to train SI mono-phone HMMs.

B. Training Speaker-dependent (SD) HMMs for new speakers Also SD HMMs for new speakers (tsu0001, tsu0003, nec0001, fuj0001) are trained for comparison between recognition results of SI,SD and PPM.

C. Training PPMs We use the same data as we used to train SI HMMs to train PPMs, though the following steps:
(1) Aligning the training data into proper states with SD mono-phone HMMs. (2) Averaging the vectors in state 2. (3) Using each combination of two phones(which we are interested in) as one phone pair event to estimate the mean vector $\boldsymbol{\mu}_{ij}$ and covariance matrix $\boldsymbol{\Sigma}_{ij}$ of this phone pair.

D. Preparing adaptation data From sentence *a*01 of each test speaker, we cut a fragment as adaptation data (uttered as *ara yuru genjitsuwo*). Then vowels $\{a, i, u, e, o\}$ are extracted using SI models. In case of there are several samples of one vowel, we just average them.

E. Results Because we aim at distinguishing the correct word from the others, only the partial score causing difference in likelihood score of words is exploited as PPM logarithmic score,

$$P^{pair}(o) = -0.5 * (log|\Sigma| + (o-\mu)'\Sigma^{-1}(o-\mu)). \quad (4)$$

In addition, we use *PPM scale* to adjust the degree of the effect of PPM. Another problem is that PPM decreases score of longer sentence more largely than that of a shorter one. Consequently, *compensation* is neccesary in our experiments.

Several trials were conducted on 50 sentences of set "*a*" of speaker *tsu*0001(each speaker has "*a*" and two other sets) to find the proper *PPM scale*. The results are shown in Fig.2 and Fig.3.

In Fig.2 and Fig.3, *SI* means recognizing with HTK, using speaker-independent models. *SD* means using the speaker-dependent models. The others use speaker-independent models and PPMs, adjusted by different *PPM scale*s. The results of integrating PPM into HTK generally outperforms that of using speaker-independent model solely. Both the word *Correct rate*, which is defined as

$$\frac{Number\ of\ correct\ words}{Total\ number\ of\ words} * 100,$$

and *Accuracy*, which is defined as

$$\frac{Number\ of\ correct\ words - insertions}{Total\ number\ of\ words} * 100,$$

get relatively higher scores simultaneously with *PPM scale* = 3 and *Compensation* around 45. Further investigations were done to determine the *Compensation* on 50 sentences of set "*a*" of new speakers (*tsu*0001, *tsu*0003, *nec*0001 and *fuj*0001), and better results were obtained around *Compensation* = 45 (See Fig.4) for the 4 speakers.

Other trials were done on speakers *tos*0001 and *tos*0002, and the results supported our selection more firmly.

Although the test data are limited, and the selection of *PPM scale* and *compensation* are

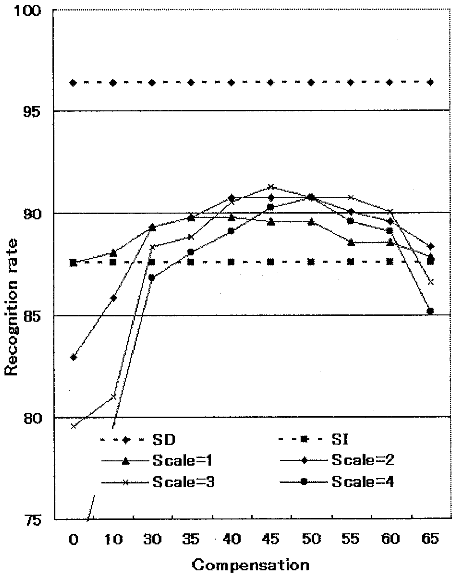

Fig 2. Word correct rate

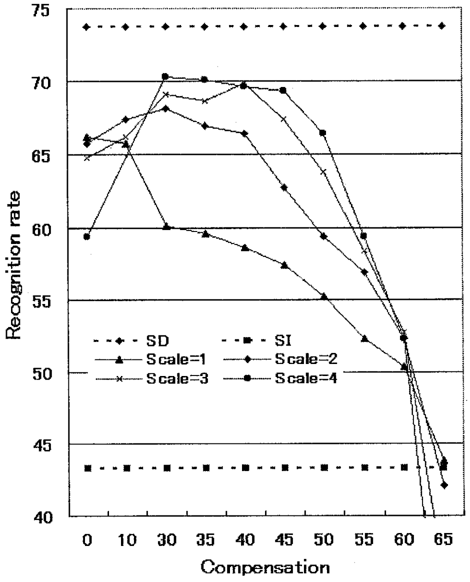

Fig 3. Word accuracy

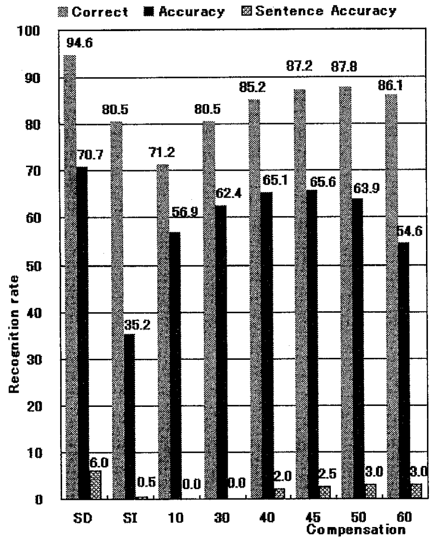

Fig 4. Recognition results for 4 new speakers

done through a cut and try process, these experiments confirm a fact that PPM do have a remarkable effect on improving both the word *Correct rate* and *Accuracy*, with the highest effect from (80.5%, 35.2%) to (87.2%, 65.6%) or to (87.8%, 63.9%).

5 Discussion

Computational load As equation(4) calculates the determinant $|\Sigma|$ and the inverse matrix of Σ, and Σ is a $2N$(N is the number of dimensions of the feature vector used in recognizing) dimensions square matrix, a heavy computational load is imposed on the recognizer. Since the 4 submatrix of Σ is diagonal (refer to equation (2)), we can simplify the computation and achieved a high processing speed close to that of original HTK.

Weighting PPMs In the above experiments, each PPM contribute to the likelihood scores equally. But [1] indicates that each PPM have a recognition error rate different from others. Hence we should find the optimal weight set to improve the recognition rate further.

6 Conclusion

We incorporate Phone Pair Model into HTK and test it on a speaker-independent recognition task. A remarkable increase of recognition rate is achieved, even given only one instance of each of the 5 vowels from the new speaker. Although how to find the optimal parameters of *PPM scale* and *compensation* have not been precisely discussed, the effectiveness of PPM is observed across a wide range of these two parameters. Further researches on PPM are expected.

References

[1] Li Bao Jie and Keikichi Hirose, " Use of phone feature correlations to robust speech recognition", *The 1999 autumn meeting of the ASJ*,pp. 129-130

[2] C.H. Lee,J.L. Gauvain, "Bayesian Adaptive Learning and MAP Estimation of HMM", *Automatic Speech and Speaker Recognition* ,pp. 83-107,1996

[3] C.J Leggetter, P. C. Woodland, "Maximum Likelihood Linear Regression for Speaker Adaptation of continuous Density Hidden Markov Models", *Computer Speech and Language*, pp. 171-185,September 1995.

[4] M.J. Lasry and R.M. Stern, "A posteriori Estimation of Correlated Jointly Gaussian Mean Vectors", *IEEE Trans. on Pattern Analysis and Machine Intelligence*, Vol. PAMI-6,No. 4,JULY,1984

[5] Timothy J. Hazen, "The Use of Speaker Correlation Information for Automatic Speech Recognition", *PhD Thesis, Massachusetts Institute of Technology*,1998

[6] J.S. Zhang, B. Dai, C.F.Wang, K. Hirose, "Adaptive Recognition Method Based on Posterior Use of Distribution Pattern of Output Probabilities", *Proc. ICSLP*,Vol. 2 pp. 1129-1132,1996

[7] S.Young, J.Odell, D.Ollason, V.Valtchev and P.Woodland, *The HTK Book*,1997