MODELING PHONE CORRELATION FOR SPEAKER ADAPTIVE SPEECH RECOGNITION

Baojie Li, Keikichi Hirose and Nobuaki Minematsu

School of Engineering, University of Tokyo
Bunkyo-ku, Tokyo, 113-8656, Japan
{lbj, hirose, mine}@gavo.t.u-tokyo.ac.jp

ABSTRACT

Information of phone relationships is regarded as acting an important role in speech recognition. It has been successfully exploited in many speaker adaptation approaches. In this paper, we propose a new approach, named Phone Pair Model (PPM) re-scoring, to utilize phone relationships for speaker-adaptive speech recognition. PPM re-scoring approach does not really adapt model parameters to a new speaker. It just uses some pre-registered phones' samples from the speaker being recognized, to re-calculate the likelihood of phones that has been calculated on conventional phone HMMs, resulting in a more correct recognition result. Additionally, it can deal with not only inter-speaker acoustic variations but also intra-speaker acoustic variations adequately. Results of two recognition experiments, one using phone HMMs only and the other incorporating phone HMMs with the PPMs, showed that even by using only a few vowel samples as the pre-registered phones, PPM re-scoring approach brought an increase in recognition rate .

1. INTRODUCTION

Because of the importance of speaker adaptation in speech recognition, its techniques have been broadly studied. However, most of them suffer from insufficient adaptation data. In view of human's ability of accurately recognizing speech in spite of large distributions of acoustic features of each individual phone, information of phone relationship should play an important role in speech recognition. Extended Maximum *a posteriori* estimation[1] and Regression-based Model Prediction[2] gave some suggestions in utilizing phone relationships. Since the former deals with all the phones simultaneously, the complicated implementation limits its practical applications. The latter approach requires that the parameters of well adapted phones should highly correlate with those of poorly adapted ones and linear relationships must exist between them. These conditions are often not easy to meet in actual cases. To make use of information of phone relationships easily in speech recognition, we propose a new approach, called Phone Pair Model (PPM) re-scoring. As described in Section 2, PPM is proposed to describe the relationship between two phones in a statistical fashion. In Section 3, by applying PPM to speech recognition, we investigate the properties of PPM and give some suggestions on its implementation and improvement. Section 4 concludes the paper.

2. PHONE PAIR MODEL

When we have some phones already known in the decoding stage, we can determine input unknown phones based on the probabilities calculated on the known-unknown phone pairs. For example, if the whole phone set is $\{P_1, \cdots, P_M\}$, $x = x_1, x_2, \cdots, x_{Tx}$ and $y = y_1, y_2, \cdots, y_{Ty}$ are two observation sequences generated from a known phone P_x's model λ_{P_x} and an unknown phone P_y's model λ_{P_y} respectively, we can calculate the conditional probability on each phone model pair $(\lambda_{P_x}, \lambda_{P_i})$, where $i \in \{1, \cdots, M\}$. If model pair $(\lambda_{P_x}, \lambda_{P_k})$ gives the highest probability, then λ_{P_k} is considered having generated y, i.e. $P_y = P_k$.

$$\begin{aligned}\hat{\lambda_{P_y}} &= \underset{\lambda_{P_i}}{argmax}\, p(y|x, \lambda_{P_x}, \lambda_{P_i}) \\ &= \underset{\lambda_{P_i}}{argmax}\, \frac{p(x, y|\lambda_{P_x}, \lambda_{P_i})}{p(x)} \\ i &= 1, \cdots, M \end{aligned} \tag{1}$$

Since $p(x)$ is invariant to i, we have

$$\hat{\lambda_{P_y}} = \underset{\lambda_{P_i}}{argmax}\, p(x, y|\lambda_{P_x}, \lambda_{P_i}). \tag{2}$$

It is too complicated to calculate the joint probability $p(x, y)$ when x, y are two observation sequences. In a conventional HMM-based recognizer, each phone is modeled as an HMM. By forced-aligning each observation sequence into states of its corresponding HMM, the joint probability $p(x, y)$ can be approximated by

$$p(x, y) \approx \prod_{i,j} p(\overline{x_i}, \overline{y_j}) \tag{3}$$

where $\overline{x_i}$ is the average of vectors aligned to state i of the HMM of P_x, and $\overline{y_j}$ is the average of vectors aligned to state j of the HMM of P_y.

To calculate the joint probability of $\overline{x_i}$ and $\overline{y_j}$, we first concatenate the two vectors to create a joint vector $(\overline{x_i}, \overline{y_j})$, then assume that it is a random vector normally distributed around its mean value

$$\mu = \begin{bmatrix} \mu_{xi} \\ \mu_{yj} \end{bmatrix}$$

with covariance matrix

$$\Sigma = \begin{bmatrix} \Sigma_{xi.xi} & \Sigma_{xi.yj} \\ \Sigma_{yj.xi} & \Sigma_{yj.yj} \end{bmatrix}. \tag{4}$$

The estimation of μ and Σ will be introduced in Section 3.4.

When the four sub-matrices of Σ are assumed diagonal, $\Sigma_{xj.yi}$ is equal to $\Sigma_{yj.xi}$. With this assumption, the computation load of calculating joint probability of two vectors will be largely reduced. See Section 3.2.2.

3. APPLYING PPM TO SPEECH RECOGNITION

3.1. Integrating PPM with Phone HMM

We use HVite of HTK(Ver.2.1.1)[3] as the baseline recognizer. In HTK, each word is represented as a sequence of phone HMMs (see the recognition network in Figure **1**. The square boxes represent word-end nodes, and the circles denote HMMs of the phones that constitute the words). Each HMM has 3 states with self-transitions, one initial state and one final state, totally 5 states.

For simplicity, the current approach only exploits the information of state 3.

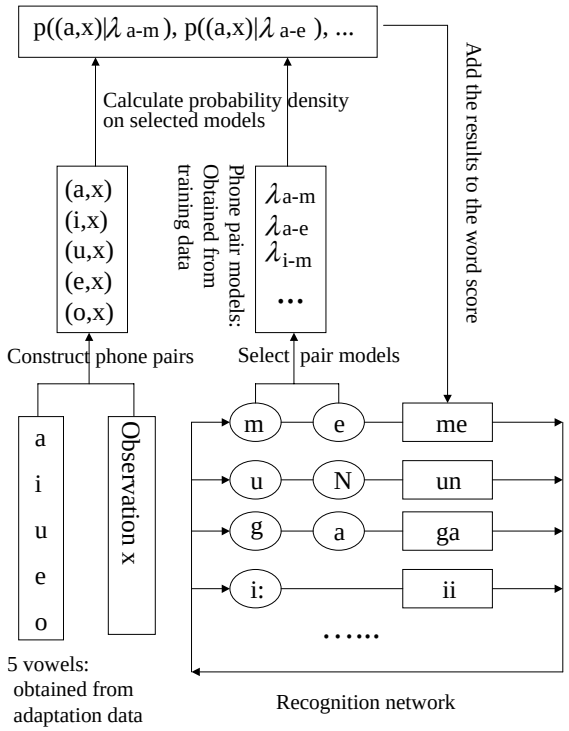

Fig. 1. Recognition process using PPMs

Before recognition, adaptation observations of an utterance from the speaker to be recognized are aligned into states of the corresponding HMMs. Then the phones which we are interested in (here are the five Japanese vowels $\{a, i, u, e, o\}$) can be extracted. Although the detected phone boundaries may include errors, they are accurate enough for our purpose. Then we average the vectors in state 3 of each HMM and get 5 vectors $\{v_a, v_i, v_u, v_e, v_o\}$ for the 5 vowels. The 5 vectors are registered for decoding stage.

During the recognition process, when a token (refer to [4]) reaches a word-end node, boundaries of the phones that constitute the current hypothesized word are known. For each phone P_y of the word, we can make up 5 vector pairs $(v_k, o_y)(k \in \{a, i, u, e, o\}$ and o_y is the average of vectors aligned to state 3 of the HMM of phone $P_y)$ based on the boundary information. By calculating the probability density of (v_k, o_y) generated by the corresponding PPM λ_{k-y} respectively, we get a *PPM score* $p_{P_y}^{pair}$, which is the average of the 5 scores, for phone P_y.

Since each word consists of a different number of phones, we average the *PPM score*s over the phones that constitute the word, to assure that PPM contributes to every word equally. Since *PPM score* of a word is added to the score of a partial path one time when the word is added to the path, and *PPM score* is usually less than 0, PPM decreases the score of a longer hypothesized sentence more largely than that of a shorter one, thus involving errors in recognition results. A positive constant is used to alleviate this effect, named *PPM compensation*. Additionally, *PPM scale* is used to weight *PPM score*. If the logarithmic likelihood of the partial path up to the current word is ψ (which is calculated in conventional way), we add the logarithmic *PPM score* to ψ and get the modified score ψ^{mod} as

$$\psi^{mod} = \psi + k(\frac{1}{N}\sum_{n=1}^{N} p_{P_n}^{pair} + p_{comp}) \qquad (5)$$

where k is *PPM scale*, p_{comp} is *PPM compensation* and N is the number of phones that constitute the word. Detailed explanations are schematically shown in Figure **1**.

3.2. Issues in Implementation

A recognition task is designed to test PPM. An utterance from the new speaker to be recognized is necessary for adaptation. Two types of recognition experiments are conducted for comparison: one using HVite only and the other incorporating PPM into HVite.

3.2.1. PPM Score

Since we are aiming at distinguishing the correct word hypothesis from the others, only the partial score, which causes difference in likelihood scores between words, is exploited in calculating the *PPM score* of a particular observation o

$$p^{pair}(o) = -0.5 * (log|\Sigma| + (o-\mu)'\Sigma^{-1}(o-\mu)). \qquad (6)$$

where Σ and μ are the covariance matrix and mean vector of the corresponding PPM respectively.

3.2.2. Computational Load

In Equation (**6**), the determinant and the inverse matrix of Σ need to be calculated. When the input vector is D-dimensional, Σ is a $2D \times 2D$ square matrix, a heavy computational load is imposed on the recognizer. This computational problem can be solved by assuming all the 4 sub-matrices of Σ being diagonal (refer to Equation (**4**)).

When Σ is represented as

$$\begin{pmatrix} a_{11} & 0 & \cdots & 0 & b_{11} & 0 & \cdots & 0 \\ 0 & a_{22} & \cdots & 0 & 0 & b_{22} & \cdots & 0 \\ \vdots & & & & \vdots & & & \\ 0 & 0 & \cdots & a_{DD} & 0 & 0 & \cdots & b_{DD} \\ b_{11} & 0 & \cdots & 0 & c_{11} & 0 & \cdots & 0 \\ 0 & b_{22} & \cdots & 0 & 0 & c_{22} & \cdots & 0 \\ \vdots & & & & \vdots & & & \\ 0 & 0 & \cdots & b_{DD} & 0 & 0 & \cdots & c_{DD} \end{pmatrix}, \quad (7)$$

then Σ^{-1} is

$$\begin{pmatrix} A_{11} & 0 & \cdots & 0 & B_{11} & 0 & \cdots & 0 \\ 0 & A_{22} & \cdots & 0 & 0 & B_{22} & \cdots & 0 \\ \vdots & & & & \vdots & & & \\ 0 & 0 & \cdots & A_{DD} & 0 & 0 & \cdots & B_{DD} \\ B_{11} & 0 & \cdots & 0 & C_{11} & 0 & \cdots & 0 \\ 0 & B_{22} & \cdots & 0 & 0 & C_{22} & \cdots & 0 \\ \vdots & & & & \vdots & & & \\ 0 & 0 & \cdots & B_{DD} & 0 & 0 & \cdots & C_{DD} \end{pmatrix}, \quad (8)$$

where

$$A_{ii} = \frac{c_{ii}}{a_{ii}c_{ii} - b_{ii}^2}, \quad B_{ii} = \frac{-b_{ii}}{a_{ii}c_{ii} - b_{ii}^2},$$

$$C_{ii} = \frac{a_{ii}}{a_{ii}c_{ii} - b_{ii}^2}, \quad i \in \{1, \cdots, D\}.$$

The determinant is given by

$$|\Sigma| = \prod_{i=1}^{D} (a_{ii}c_{ii} - b_{ii}^2). \quad (9)$$

3.3. Training HMMs

ASJ (Acoustic Society of Japan) Continuous Speech Corpus for Research is used as the training and test data. 150 utterances from each of 6 male speakers, totally 900 utterances are used to train speaker-independent (SI) monophone HMMs (called SI-6 models for later references). 3 other speakers are used as test speakers. Their speaker-dependent (SD) HMMs are also trained for comparison.

3.4. Training Phone Pair Models

The same data as we used in training SI HMMs is used to train PPMs, through the following steps:

1. Align the training data to states of SI phone HMMs.
2. Average the vectors aligned to state 3 of the SI phone HMMs.
3. Select samples of phone pair to estimate the mean vector μ and covariance matrix Σ of the corresponding phone pair model. Figure **2** illustrates this process for a given utterance "aka da" (meaning "It is red").

   As depicted in Figure **2**, 10 phone pair samples can be obtained. Here, to catch the intra-speaker acoustic variations, two phone samples are chosen to make up a phone pair sample only when they appear in the same utterance.

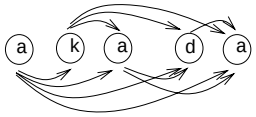

Fig. 2. Selecting phone pair samples for training PPM

3.5. Preparing Adaptation Data

Given an adaptation utterance and its transcription, by performing forced-alignment with SI HMMs, the average vectors of state 3 of the HMMs of the 5 vowels $\{a, i, u, e, o\}$ can be extracted. They are registered for the following decoding stage.

3.6. Test Conditions and Results

The parameter vector is 38-dimensional for both HMMs and PPMs, containing 12th order $MFCC$s, $\Delta MFCC$s, $\Delta\Delta MFCC$s, Δ power, $\Delta\Delta$power. The dictionary consists of 886 words. 50 utterances from each of the 3 test speakers are tested.

The results are shown in Figure **3** and Figure **4**, where SI, SD mean recognizing with SI models and SD models, respectively. The other curves show the results using SI HMMs incorporated with PPMs, with different *PPM scale*s and *PPM compensation*s.

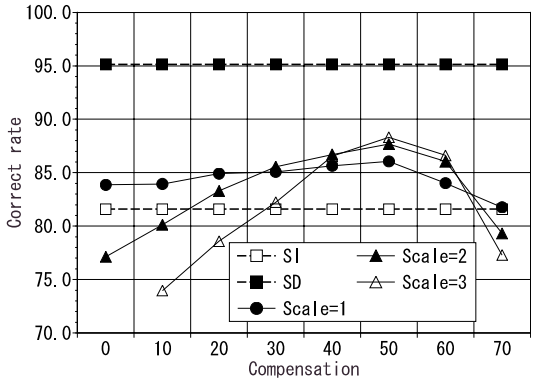

Fig. 3. Word correct rate using SI-6 HMMs incorporated with PPMs

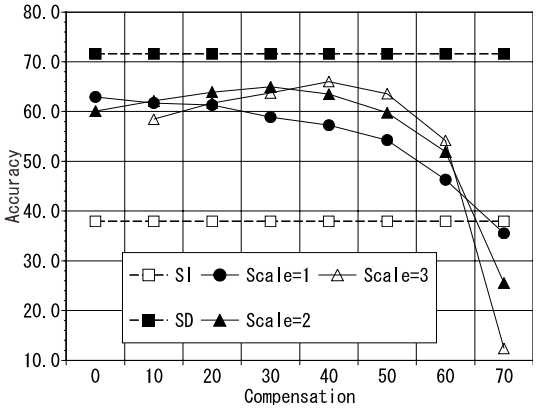

Fig. 4. Word accuracy using SI-6 HMMs incorporated with PPMs

The figures show that PPM re-scoring generally results in an obvious increase in both word correct rate and word accuracy across a wide range of *PPM compensation*.

3.7. Further Investigations on PPM

We still have to address 2 problems

1. How does the effect of PPM vary when the performance of SI models is improved?
2. How do we set *PPM scale* and *PPM compensation* to appropriate values?

In order to solve these problems, further experiments are conducted using the SI mono-phone HMMs provided by Information-technology Promotion Agency, Japan (called IPA-SI models in contrast to SI-6 models, with 16 mixture components in each state of HMM, trained with the *ASJ Continuous Speech Corpus for Research* and *Japanese newspaper article sentences*, totally 20k sentences uttered by 132 speakers. The parameter vector is 25-dimensional containing 12th order $MFCC$s, $\Delta MFCC$s and Δ power) under the same conditions except that the dictionary is extended from 886 words to 2947 words.

The results are shown in Figure **5** and Figure **6**.

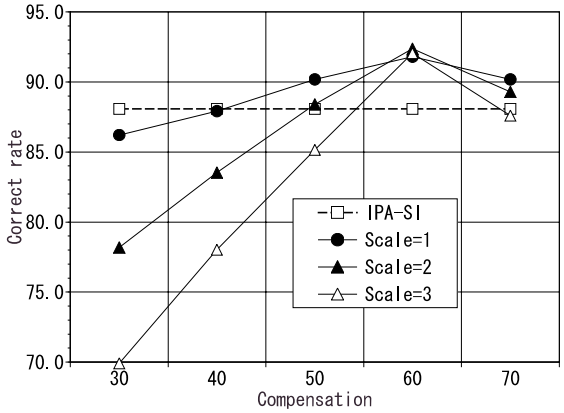

Fig. 5. Word correct rate using IPA-SI HMMs incorporated with PPMs

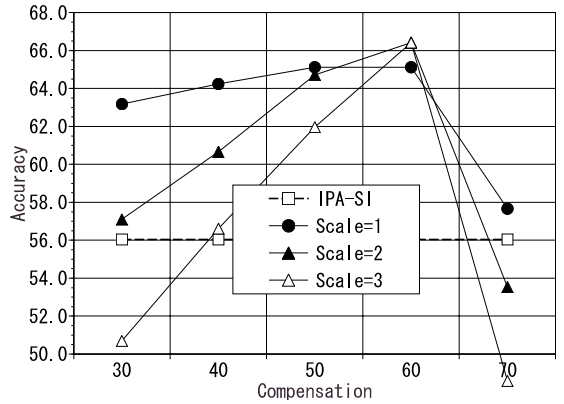

Fig. 6. Word accuracy using IPA-SI HMMs incorporated with PPMs

In the above experiments, even though the baseline SI models are rather well trained, PPM re-scoring still increases word correct rate and word accuracy. However, the *PPM active ranges* (where PPM re-scoring outperforms the baseline recognizer) become narrower. The relative reliability of PPMs and HMMs (high reliability when trained well) is considered as one main factor that influences *PPM active range*, *PPM scale* as well as the obtained improvement. When the PPMs are well-trained, a larger *PPM scale* may be preferable.

The results also suggest us that the optimum *PPM compensation* varies with SI HMMs. The vector size of model parameters is considered as another factor. It may be set properly by a few tests preceding the recognition.

4. CONCLUSION

A new approach, PPM re-scoring, was proposed to utilize phone relationships for speaker-adaptive speech recognition. It has the following main properties

1. Since it does not really adapt model parameters to a new speaker, its "adaptation" stage is very simple.
2. By modeling phone relationship of each phone pair separately, some particular phone pairs that involve errors in the recognition results (e.g. consonant-consonant pairs) can be removed. The re-scoring computational load may also be reduced largely.
3. Since only the phones appearing in the same utterance are selected to make up a phone pair sample in training PPMs, PPMs can also catch intra-speaker variations.

We incorporated PPM with phone HMM and tested it on a speaker-independent recognition task. A remarkable increase of recognition rate was achieved, even given only a few vowel samples of the new speaker.

REFERENCES

[1] 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

[2] S.M. Ahadi and P.C. Woodland, "Combined Bayesian and Predictive Techniques for Rapid Speaker Adaptation of Continuous Density Hidden Markov Models", Computer Speech and Language, Vol. 11, pp. 187-206, 1997

[3] S.Young, J.Odell, D.Ollason, V.Valtchev and P.Woodland, "HTK-Hidden Markov Model Toolkit", *Cambridge Research Laboratory*, 1997

[4] S.J. Young, N.H. Russell and J.H.S. Thornton, "Token Passing: a Simple Conceptual Model for Connected Speech Recognition Systems", *CUED Technical Report F_INFENG/TR38*, Cambridge University, 1989. Available by anonymous ftp from svr-ftp.eng.cam.ac.uk.