音素ペアモデルによる音素間情報の表現に関する検討

李 宝潔[†] 広瀬啓吉[‡] 峯松信明[†]
[†] 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

ある音素の特徴は発話毎に変動し、もし話者が異なればその分散は大きな値を示す。このような事実にもかかわらず、人間が良好に音声を認識できるのは、音素間の特徴が比較的安定していて、これを有効に使っているからと考えられる。そこで、このような音素間の関係を音素ペアモデルで表す。音素ペアモデルは２つの音素間の関係を音響ベクトルの結合確率として表現するものである。音素ペアモデルを音素 HMM による音声認識システムに統合して認識実験を行い、音素ペアモデルの尤度を付加する割合などに検討の余地があるものの、顕著な認識率の向上が得られた。さらに、頑健な音声認識にも有効であることが示された。

Representation of Phone Correlation by Phone Pair Model

Baojie LI[†] Keikichi HIROSE[‡] Nobuaki MINEMATSU[†]
[†]School of Engineering, University of Tokyo
[‡]School of Frontier Sciences, University of Tokyo

In view of human's ability of accurately recognizing speech in spite of large distributions of phone features, information on the relationship between phones should play an important role in speech recognition. The phone relationship can be modeled by our proposed Phone Pair Model(PPM). PPM statistically models the relationship between two phones using joint probability of acoustic features. After integrating PPM into HMM-based recognition systems, recognition experiments are conducted. The results showed remarkable increases in recognition rates only by a short phrase for adaptation. The Phone Pair Modeling was also shown valid for robust speech recognition.

1 Introduction

Speaker adaptation 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 phone features, information on the relationship between phones 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, but their effects were rather limited. We proposed a Phone Pair Model (PPM) re-scoring adaptation approach previously[3]. In this paper, we report our recent works on this approach. We investigate the property of PPM, give some suggestions on its implementation and improvements. As described in Section 2, PPM is proposed to describe the relationship between two phones in a statistical fashion. In Section 3, PPM is then integrated with phone HMM to re-calculate the likelihood of words in the recognition network within a HTK [4] framework. Its application to speaker adaptation is introduced in Section 4, where two series of recognition experiments are conducted. In Section 5, the robustness of PPM in speech recognition is shown by some experiment results. Section 6 concludes the paper.

2 Phone Pair Model

When we have some phones known in the decoding stage, we can determine the unknown phones based on the probabilities calculated on the known-unknown phone pairs. For example, if $X = x_1, x_2, \cdots, x_{Tx}$ and $Y = y_1, y_2, \cdots, y_{Ty}$ are two observation sequences generated from phone $phone_X$'s model λ_X and an unknown phone model respectively, we can calculate the conditional probability on each phone model pair $(\lambda_X), \lambda_i, \lambda_i \in \{\lambda_1, \cdots, \lambda_M\}$, ($M$: the number of phones), and then select the model that generated Y as follows.

$$\begin{aligned} \hat{\lambda_Y} &= \underset{\lambda_i}{argmax}\, p(Y|(X, \lambda_X, \lambda_i)) \\ &= \underset{\lambda_i}{argmax} \frac{p((Y, X)|(\lambda_i, \lambda_X))}{p(X)} \\ i &= 1, \cdots, N \end{aligned} \tag{1}$$

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

$$\hat{\lambda_Y} = \underset{\lambda_i}{argmax}\, p((Y, X)|(\lambda_i, \lambda_X)). \tag{2}$$

In conventional HMM-based recognizer, each phone is modeled with an HMM. The joint probability of (X, Y) can be approximated by

$$p(X, Y) \approx \prod_{i,j} (p(\overline{x_{si}}, \overline{y_{sj}})) \tag{3}$$

where $\overline{x_{si}}$ is the average of vectors belonging to state s_i of the HMM of $phone_X$, and $\overline{y_{sj}}$ is the average of vectors belonging to state s_j of the HMM of $phone_Y$. Let variable vector $\overline{X_{si}}$ has normal distribution $N(\mu_{si}, \Sigma_{si})$ and $\overline{Y_{si}}$ has $N(\mu_{sj}, \Sigma_{sj})$, Then the distribution of joint vector $(\overline{X_{si}}, \overline{Y_{sj}})$ is $N(\mu_{ij}, \Sigma_{ij})$ in which

$$\mu_{ij} = \begin{bmatrix} \mu_{si} \\ \mu_{sj} \end{bmatrix}$$

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

$$\Sigma_{ij} = \begin{bmatrix} \Sigma_{si.si} & \Sigma_{si.sj} \\ \Sigma_{sj.si} & \Sigma_{sj.sj} \end{bmatrix} \tag{4}$$

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

3 Incorporating PPM into Phone HMM

We use HTK(Ver.2.1.1)[4] as the baseline recognizer. In HTK, each word is represented as a sequence of phone HMMs(see the recognition network in Figur 1). The square boxes represent word-end node, and the circles denote HMMs of phones composing the word.

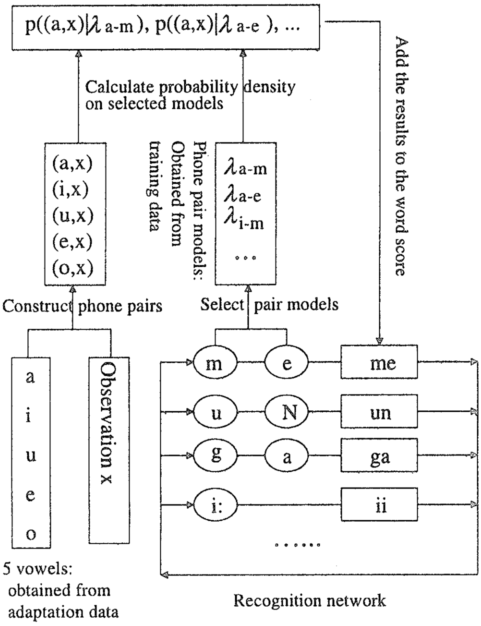
p((a,x)|λ a-m), p((a,x)|λ a-e), ...
Calculate probability density on selected models
Add the results to the word score
(a,x)
(i,x)
(u,x)
(e,x)
(o,x)
Phone pair models: Obtained from training data
λa-m
λa-e
λi-m
...
Construct phone pairs
Select pair models
a
i
u
e
o
Observation x
5 vowels: obtained from adaptation data
m
e
me
u
N
un
g
a
ga
i:
ii
......
Recognition network

図 1: Recognition process using PPM

Each HMM has 3 states with self-transitions, one initial state and one final state, totally 5 states. For simplicity, our current PPM only exploits the information of state 3. Before recognition, the adaptation observations from the new speaker are aligned into the states of the HMMs. Although the detected boundaries of each phone may include errors, we can extract the phones we are interested in (here are the five Japanese vowels $\{a, i, u, e, o\}$) with enough accuracy. Then we average the vectors in state 3 of each HMM and get 5 vectors $\{\boldsymbol{v}_a, \boldsymbol{v}_i, \boldsymbol{v}_u, \boldsymbol{v}_e, \boldsymbol{v}_o\}$ for the 5 vowels $\{a, i, u, e, o\}$ respectively.

During the recognition process, when a token (refer to [4]) reached a word-end node, boundaries of the phones composing the current word are known. With the boundary information, we can make up 5 vector pairs $\boldsymbol{v}_k - \boldsymbol{o}_x^i(k \in \{a, i, u, e, o\}$) for each observation $\boldsymbol{o}_x^i$ (the i'th observation in state 3 of phone ph_x's model). Then the probability density of each $\boldsymbol{v}_k - \boldsymbol{o}_x^i$ generated by the corresponding PPM λ_{k-x} is calculated. Thus a *PPM score* p_x^{pair}, which is the average on the 5 PPMs and all the observations in state 3, is obtained for phone ph_x.

Since each word consists of different number of phones, we average the *PPM scores* of all the phones that compose the word, to assure that PPM contributes to every word equally. Moreover, since PPM decreases the score of a longer sentence more largely than that of a shorter one, and therefore involves errors in recognition results, *PPM compensation* is used to alleviate this effect. Additionally, PPM scale is used to weight *PPM score.* If the logarithmic likelihood of the partial path till current word is ψ (which is calculated in conventional way), we add the logarithmic *PPM score* of this word to ψ and get the modified score psi^{mod} as

$$\psi^{mod} = \psi + k(\frac{1}{M}\sum_{m=1}^{M} p_m^{pair} - p_{comp}) \qquad (5)$$

where k is *PPM scale*, $p_{compensation}$ is *PPM compensation* and M is the number of the phones composing the word. Detailed explanations are schematically shown in Figure 1.

4 Applying PPM to Speaker Adaptation

A recognition task is designed to test PPM. As many other adaptation algorithms, a fragment of an utterance from the new speaker is necessary for adaptation.

4.1 Training Phone Pair Models

We can 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 SI phone HMMs.

2. Averaging the vectors in the middle state of the SI phone HMMs.

3. Selecting the combinations of two phones(which we are interested in) as one phone pair event, using the average vectors of the middle state of their HMMs to estimate the mean vector $\boldsymbol{\mu}_{ij}$ and covariance matrix $\boldsymbol{\Sigma}_{ij}$ of this phone pair.

4.2 Preparing Adaptation Data

By performing forced-alignment, the mean vector of the middle states of the 5 vowels $\{a, i, u, e, o\}$ are extracted from a given adaptation utterance and registered for the following decoding stage.

4.3 Some Problems

4.3.1 *PPM Score*

Since we aim at distinguishing the correct word from the others, only the partial score, which causes difference in likelihood scores between words, is exploited as *PPM score*

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

where $\boldsymbol{o}$ is an observation of input speech, $\boldsymbol{\Sigma}$ and $\boldsymbol{\mu}$ the covariance matrix and mean vector of corresponding PPM respectively.

4.3.2 Computational Load

In Equation(6) the determinant and the inverse matrix of $\boldsymbol{\Sigma}$ need to be calculated, When the input vector is D-dimensional, Σ is a $2D \times 2D$ square matrix, and a heavy computational load is imposed on the recognizer. This computational problem can be solved by utilizing the property of Σ that its 4 submatrices are diagonal (refer to Equation (4)). When Σ is represented as

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

then Σ^{-1} is

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

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

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

4.3.3 Setting *PPM scale* and *PPM compensation*

PPM scale is used to adjust the relative effect (to phone HMMs) of PPMs. Additionally, since PPM decreases score of longer sentence more largely than that of a shorter one, *PPM compensation* is also necessary.

To find the proper *PPM scale* and *PPM compensation*, we conduct two series of recognition experiments, one uses poorly trained SI mono-phone models (SI-6 models), which are trained using 150 sentences of each of 6 male speakers from *ATR Continuous Speech Corpus for Research.* The other uses the SI mono-phone HMMs provided by IPA (called IPA-SI models in contrast to SI-6 models, with 16 mixture components in each state of HMM, trained with *ASJ Continuous Speech Corpus for Research* and *Japanese newspaper article sentences*, totally 20k sentences uttered by 132 speakers).

The test conditions are

Vector size The vector size for IPA-SI is 25-dimensional, containing 12th order $MFCC$s,$\Delta MFCC$s, $\Delta\Delta MFCC$s, Δ power. It is 38-dimensional for SI-6 models and PPMs, with $\Delta\Delta MFCC$s, Δ power, $\Delta\Delta power$ in addition.

Dictionary size 886 words for the experiments using SI-6 models and 2947 words for those using IPA models

Test data : 50 sentences from each of 3 new speakers from ATR *Continuous Speech Corpus for Research*

The results of the former series of experiments are shown in Figure 2 and Figure 3, where SI, SD mean recognizing with speaker-independent models and speaker-dependent models, respectively. The other curves show results using speaker-independent models integrated with PPM, with different *PPM scales* and *PPM compensations*. Figure 4 and Figure 5 are the results for the latter experiment series.

As shown in the above figures, in both of the two series experiments, PPM generally results in an obvious increase in both word correct rate, defined as

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

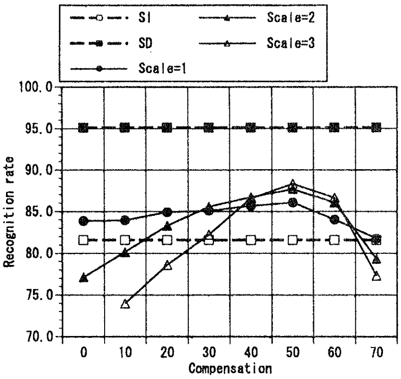

図 2: Recognition word correct rate of SI-6 models

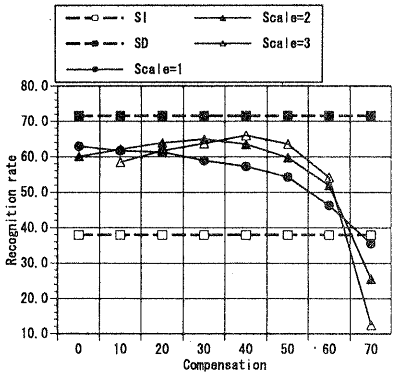

図 3: Recognition word accuracy of SI-6 models

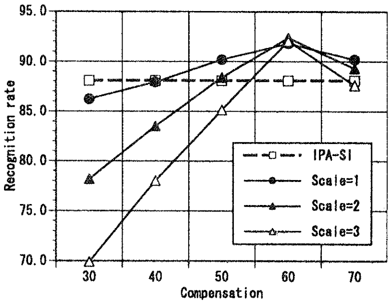

図 4: Recognition word correct rate of IPA-SI models

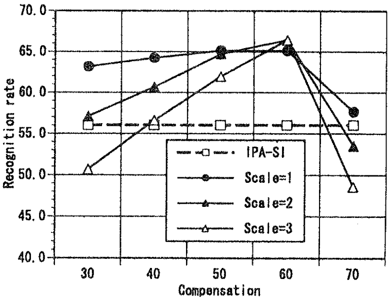

図 5: Recognition word accuracy of IPA-SI models

and word *accuracy*, defined as

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

In each experiment series, when the *PPM scale* increased, we get only a little increase in recognition rate, however the range of *PPM compensation*, within which PPM outperforms conventional HMM (called *PPM active range*), becomes narrower dramatically. And this also occurs when the performance of the SI models improved. This may be attributable to the relative performance of PPMs compared to the SI models. When the PPMs are well-trained, a larger *PPM scale* is preferable. *PPM scale* and *PPM compensation* may be set properly by a few tests preceding the adaptation.

4.3.4 Weighting Phone Pair Models

In the above experiments, each PPM contributes to the likelihood scores equally. But each PPM has a recognition error rate different from others. Hence we should find the optimal weight set to improve the recognition rate further. This problem is left for future investigation.

5 Robustness of PPM

After the investigation on the relationships between phones within a speaker, we continue to investigate how these relationships exist across speakers. Instead of extracting

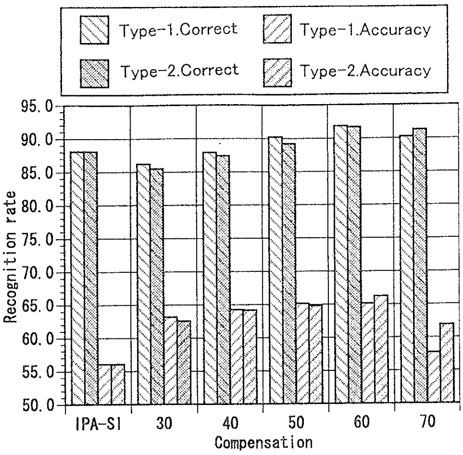

図 6: Comparison of the recognition results between the two type experiments

the 5 vowels from an utterance of the speaker to be recognized in the previous experiments (called Type-1 experiments), the vowels are extracted from an utterance of an arbitrary training speaker in the following experiments (called Type-2 experiments). The recognition results of the two types of experiments are shown in Figure 6. As shown in Figure 6, there is only a little difference between the results of the two type experiments. While this shows the robustness of PPM in speech recognition, larger improvements of Type-1 experiments are expected when compared with Type-2 experiments. This may be achieved by setting different PPM compensation for each PPM and modeling phone relationships more precisely.

6 Conclusion

We incorporated PPM into phone HMM and tested it on a speaker-independent recognition task. A remarkable increase of recognition rate was achieved, even given only one sample of each of the 5 vowels from the new speaker. The robustness of PPM was also shown by experiments. Some suggestions were given on properly setting *PPM scale* and *PPM compensation*. Further improvements of PPM may be made by defining the PPM more precisely.

参考文献

[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] 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

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