ADAPTIVE RECOGNITION METHOD BASED ON POSTERIOR USE OF DISTRIBUTION PATTERN OF OUTPUT PROBABILITIES*

Jin-Song Zhang[+], *Beiqian Dai*[+], *Changfu Wang*[+],*Hingkeung Kwan*[++], *Keikichi Hirose*[++]

[+] Dept. of Elec. Engr., USTC, Hefei, Anhui 230026, PRC
email {jszhang, bqdai, cfwang@ee.cc.ustc.edu.cn}

[++] Dept. of Elec. Infm., Univ. of Tokyo, Hongo, Bunkyo-ku 7-3-1, Tokyo, Japan
email{kan,hirose@gavo.t.u-tokyo.ac.jp}

ABSTRACT

We propose a new adaptation scheme for speaker independent recognition. The basic idea lies in the change of the likelihood from the ordinary HMM scores to the combined observational scores. The new likelihood is computed based on a combination of HMM scores which we called a Distribution Pattern of Output Probabilities (POPD). The system needs to calculate only the POPD for each new speaker. Re-estimation of acoustic model parameters is unnecessary. Preliminary experiments on Chinese icolated -syllable recognition indicate the method's effectiveness.

1. INTRODUCTION

Recognition rate decrease for new speakers has traditionally been dealt with adaptation techniques which adjust the model parameters to new speakers or new environments, such as feature normalization and parameter modification.

We avoid such performance deterioration using an adaptation method which does not re-estimate acoustic model parameters. A conventional HMM based speech recognition system usually works under the maximum likelihood decision, and the HMM output scores are used as the likelihood measure. Most systems select only the candidate with the maximum probability. Whenever gaps exist between the acoustic characteristics of the model and the new speaker, recognition deteriorates. The most common way to mitigate the gap is the re-estimation of the HMM models so that the HMM output scores can be reliable likelihood.

We propose a likelihood measure for models deviating from the new speaker. Instead of considering the model with the largest output probability as the recognition result, we assume that the output probabilities for one's utterances computed according to the models of the system should have patterns of consistent distribution, so we can find a new way to develop likelihood from the combination of HMM scores rather than the one-model-score. In our system, an adaptive decision-maker is applied to a basic HMM recognizer, the decision-maker first computes likelihood based on the POPD and the HMM scores for the input utterance, and then decides according to the likelihood. Because of the mechanism, we call it posterior adaptation based on POPD.

As for the whole adaptation recognition performance, The system first trains the POPD during the adaptation process. During the recognition process, the likelihood is calculated from the stored POPD and the HMM outputs for the input speech. The new decision-maker selects candidates according to the new likelihood form rather than HMM output probabilities.

2. PRINCIPLES

2.1. Maximum Likelihood Classification

As known already, all statistic pattern recognition methods are based on Bayesian Equation below

$$\mathbf{P}(\mathbf{c}|\bar{\mathbf{x}}) = \frac{\mathbf{P}(\bar{\mathbf{x}}|\mathbf{c})\mathbf{P}(\mathbf{c})}{\mathbf{P}(\bar{\mathbf{x}})} \qquad (1)$$

in the case of speech recognition, the class random variable **c** may represent the syllable index in a vocabulary, say $\mathbf{c}^i$ (**i**=1, 2, ..., K), the random vector $\bar{\mathbf{x}}$ models a list of acoustic features extracted from the utterance of a syllable to be recognized. Ideally, given a feature vector outcome $\bar{\mathbf{x}}=\mathbf{x}$, we would select the class for which the conditional probability is the highest. That is $\mathbf{c}^i$ is the selected class (syllable) if

$$\mathbf{c}^i = \arg\max_{\mathbf{c}} \mathbf{P}(\mathbf{c}|\bar{\mathbf{x}} = \mathbf{x}) \qquad (2)$$

but we can not compute the probabilities $\mathbf{P}(\mathbf{c}|\bar{\mathbf{x}} = \mathbf{x})$, instead what we can get is the probability that a given class will generate

* The work is sponsored by the China Natural Science Foundation and JSPS, Japan

certain feature vectors, introduce (1) to (2)

$$\mathbf{c}^{i} = \arg\max_{c} \frac{\mathbf{P}(\bar{\mathbf{x}} = \mathbf{x}|\mathbf{c})\mathbf{P}(\mathbf{c})}{\mathbf{P}(\bar{\mathbf{x}} = \mathbf{x})} \qquad (3)$$

under the condition of no linguistic knowledge, the class (syllable) probabilities are assumed to be equal,

$$\mathbf{P}(\mathbf{c} = \mathbf{c}^{1}) = \mathbf{P}(\mathbf{c} = \mathbf{c}^{2}) = \cdots = \mathbf{P}(\mathbf{c} = \mathbf{c}^{K}) \qquad (4)$$

then

$$\mathbf{c}^{i} = \arg\max_{c} \mathbf{P}(\mathbf{c}|\bar{\mathbf{x}} = \mathbf{x}) = \arg\max_{c} \mathbf{P}(\bar{\mathbf{x}} = \mathbf{x}|\mathbf{c}) \qquad (5)$$

Therefore, the class decision is based on the maximization of the likelihood $\mathbf{P}(\bar{\mathbf{x}} = \mathbf{x}|\mathbf{c})$.

In a speech recognition system where Hidden Markov Models are used as the acoustic models, each class **c** (syllable) is represented by a HMM model $\lambda = \lambda_j (\mathbf{j} = 1,2,\cdots,\mathbf{K})$ Using the Viterbi search algorithm, we can compute the Viterbi score that a given model $\lambda_j (\mathbf{j} = \mathbf{1},\cdots,\mathbf{K})$ will generate the input speech feature **x**, this is denoted by the form $\mathbf{Prob}(\bar{\mathbf{x}} = \mathbf{x}|\lambda = \lambda_j)$. When parameters for model λ can be pre-estimated suitably, $\mathbf{Prob}(\bar{\mathbf{x}} = \mathbf{x}|\lambda = \lambda_j)$ can be used as a good likelihood measure for class decision in equation (5),

$$\mathbf{c}^{i} = \arg\max_{j=1,\cdots,K} \mathbf{Prob}(\bar{\mathbf{x}} = \mathbf{x}|\lambda = \lambda_j) \qquad (6)$$

2.2 Probabilistic Overlap Measure

This section briefly describes the concept Probabilistic Overlap Measure to ease explaining POPD in the following section.

The simplest techniques for measuring probabilistic class overlap (or separation, interclass distance) are based on distance metrics in multidimensional space, especially the Euclidean distance and its variants. These measures generally do not utilize much of the probabilistic structure of the classes and therefore do not faithfully represent the degree of overlap of the classes in a statistical sense. The Probabilistic Overlap Measure represent an attempt to capture that information in the evaluation.

Given two probabilistic classes, say $\mathbf{c}^{1}, \mathbf{c}^{2}$, each class has a class-conditional probabilistic density function $\mathbf{f}_{\bar{x}|c}(\mathbf{x}|\mathbf{c} = \mathbf{c}^{1})$ or $\mathbf{f}_{\bar{x}|c}(\mathbf{x}|\mathbf{c} = \mathbf{c}^{2})$, the definition of the Probabilistic Overlap Measure for the two classes take the form

$$\mathbf{J} = \int_{-\infty}^{\infty} \mathbf{g}\{\mathbf{f}_{\bar{x}|c}(\mathbf{x}|\mathbf{c} = \mathbf{c}^{1}), \mathbf{f}_{\bar{x}c}(\mathbf{x}|\mathbf{c} = \mathbf{c}^{2})\}\mathbf{dx} \qquad (7)$$

where $\mathbf{g}(\cdot)$ is an appropriate function, and $\int_{-\infty}^{\infty}(\cdot)\mathbf{dx}$ indicates the integral over the entire N-dimensional hyperplane with N the dimension of the feature vector **x**. Probabilistic Overlap Measure have the following properties:

① **J** is nonnegative, $\mathbf{J} \geq 0$.

② **J** attains a maximum and **J**=1 when $\mathbf{f}_{\bar{x}|c}(\mathbf{x}|\mathbf{c} = \mathbf{c}^{1}) = \mathbf{f}_{\bar{x}|c}(\mathbf{x}|\mathbf{c} = \mathbf{c}^{2})$

③ **J**=0 when either $\mathbf{f}_{\bar{x}|c}(\mathbf{x}|\mathbf{c} = \mathbf{c}^{1}) = 0$ or $\mathbf{f}_{\bar{x}|c}(\mathbf{x}|\mathbf{c} = \mathbf{c}^{2}) = 0$

2.3 Pattern of Output Probability Distribution

For the speech recognition task mentioned in session 2.1 for a speaker $\hat{\mathbf{T}}_0$, $\mathbf{x}'$ denotes his speech feature, let us hypothesize that a set of HMM models $\hat{\lambda} = \hat{\lambda}_j (\mathbf{j} = 1,2,\cdots\mathbf{K})$ best models the acoustic characteristics of $\hat{\mathbf{T}}_0$ for the vocabulary **c**. According to equation (6), the class (syllable) recognition decision is based on

$$\mathbf{c}^{i} = \arg\max_{j=1,\cdots,K} \mathbf{Prob}(\bar{\mathbf{x}} = \mathbf{x}'|\hat{\lambda} = \hat{\lambda}_j) \qquad (8)$$

in the case of speaker independent recognition task, $\hat{\mathbf{T}}_0$ is a new speaker, and the system HMM model set is $\lambda = \lambda_j (\mathbf{j} = 1,2,\cdots,\mathbf{K})$, the degree of superposition of the existing system models for the new speaker can be evaluated by the Probabilistic Overlap Measure

$$\begin{aligned} \mathbf{J}_i &= \mathbf{J}(\hat{\lambda}_i|\lambda_i) \qquad\qquad \mathbf{i} = \mathbf{1},\cdots,\mathbf{K} \\ &= \int_{-\infty}^{\infty} \mathbf{g}\{\mathbf{Prob}(\bar{\mathbf{x}} = \mathbf{x}'|\hat{\lambda} = \hat{\lambda}_i), \mathbf{Prob}(\bar{\mathbf{x}} = \mathbf{x}'|\lambda = \lambda_i)\}\mathbf{dx} \end{aligned}$$

if $\mathbf{J}_i$=1, i=1,...,K, then the system model set $\lambda = \lambda_1, \lambda_2, \cdots, \lambda_K$ completely fits the new speaker $\hat{\mathbf{T}}_0$, and $\mathbf{Prob}(\bar{\mathbf{x}} = \mathbf{x}'|\lambda = \lambda_i)$ is a suitable likelihood for decision. However, the actual situation is $0 < \mathbf{J}_i < 1, \mathbf{i} = 1,\cdots,\mathbf{K}$; hence $\mathbf{Prob}(\bar{\mathbf{x}} = \mathbf{x}'|\lambda = \lambda_i)$ is not a good likelihood measure for class decision.

In order to make the score $\mathbf{Prob}(\bar{\mathbf{x}} = \mathbf{x}'|\lambda = \lambda_i)$ a good measure for likelihood, various adaptation methods have been proposed to modify the model set $\lambda = \lambda_j (\mathbf{j} = 1,2,\cdots,\mathbf{K})$ to satisfy the condition $\mathbf{J}_i$=1, i=1,...,K, and some of them seem to be efficient.

However, if we can find something more suitable for the likelihood measure than the score $\mathbf{Prob}(\bar{\mathbf{x}} = \mathbf{x}'|\lambda = \lambda_i)$, we may solve the problem above differently. From equation (8) we may know the suitable likelihood for class $\mathbf{c}^{i}$ must be related to the model $\hat{\lambda}_i$. Although we can not acquire the desirable model $\hat{\lambda}_i$, we can indirectly get some information about it. Under the condition of $0 < \mathbf{J}_i < 1, \mathbf{i} = 1,\cdots,\mathbf{K}$, it is obvious that the probabilistic overlap measures $\mathbf{J}_{i,j}$ $(\mathbf{i} \neq \mathbf{j}, \mathbf{j} = \mathbf{1},\cdots\mathbf{K})$ can also be defined between $\hat{\lambda}_i$ and $\hat{\lambda}_j$

$$\mathbf{J}_{i,j} = \mathbf{J}(\hat{\lambda}_i|\lambda_j) \qquad \mathbf{i,j=1,\cdots,K}$$
$$= \int_{-\infty}^{\infty} \mathbf{g}\{\mathbf{Prob}(\bar{\mathbf{x}} = \mathbf{x}'|\hat{\lambda} = \hat{\lambda}_i),\mathbf{Prob}(\bar{\mathbf{x}} = \mathbf{x}'|\lambda = \lambda_j)\}\mathbf{dx}$$

In other words, a vector consists of probabilistic overlap measures

$$\bar{\mathbf{J}}_i = \{\mathbf{J}(\hat{\lambda}_i|\lambda_1), \mathbf{J}(\hat{\lambda}_i|\lambda_2), \cdots, \mathbf{J}(\hat{\lambda}_i|\lambda_i), \cdots, \mathbf{J}(\hat{\lambda}_i|\lambda_K)\}$$

can be a feature vector reflecting the acoustic characteristics of the model $\hat{\lambda}_i$.

As the model $\hat{\lambda}_i$ is unknown, what we can observe are the feature vectors $\mathbf{x}_i'$ generated by the model $\hat{\lambda}_i$, the vector ξ_i of HMM output scores for $\mathbf{x}_i'$ computed by the $\lambda = \lambda_1, \lambda_2, \cdots, \lambda_K$

$$\xi_i = \{\mathbf{Prob}((\bar{\mathbf{x}} = \mathbf{x}_i'|\hat{\lambda}_i)|\lambda_1), \cdots,$$
$$\mathbf{Prob}((\bar{\mathbf{x}} = \mathbf{x}_i'|\hat{\lambda}_i)|\lambda_i), \cdots, \mathbf{Prob}((\bar{\mathbf{x}} = \mathbf{x}_i'|\hat{\lambda}_i)|\lambda_K)\}$$
$$= \{\mathbf{Prob}(\bar{\mathbf{x}} = \mathbf{x}_i'|\lambda_1), \cdots,$$
$$\mathbf{Prob}(\bar{\mathbf{x}} = \mathbf{x}_i'|\lambda_i), \cdots, \mathbf{Prob}(\bar{\mathbf{x}} = \mathbf{x}_i'|\lambda_K)\}$$

should have some relations with the vector $\bar{\mathbf{J}}_i$. Since $\mathbf{x}_i'$ are random vectors, the ξ_i is also a random vector with a distribution related to $\mathbf{x}_i'$. The ξ_i can represent the pattern of model $\hat{\lambda}_i$ and is expressed in the form of HMM output scores, we call it Pattern of Output-Probabilities-Distribution (POPD).

For the purpose of modeling the $\hat{\lambda}_i$, it is important how to acquire the POPD vector ξ_i, including how to estimate each element, how many elements in the ξ_i is enough for the characterization, how to choose the units. How to use POPD for recognition task is also important.

By establishing a POPD database after extracting POPD information for the new speaker, the original speech recognition system it was based on will work for the new speaker without the require for re-estimation of HMM model parameters.

3. POSTERIOR ADAPTATION BASED ON POPD

Various applications of POPD are possible for speech recognition. As POPD represents matching scores of HMMs, it is easily obtained online. The system continues training the POPD database during supervised recognition, effectively adapting itself to a new speaker or environment through the POPD learning process.

We set up an isolated speech recognition system based on POPD, which can be effectively used for speaker independent recognition task. Through a database called Confusion & Discrimination Knowledge the system use the POPD information to discriminate the confusing syllables.

The adaptation procedure (vocabulary-dependent adaptation) is described in figure 1. During the adaptation period, the one important process is to compute the POPD for the new speaker, and the other is confusion and discrimination analysis. Discriminating knowledge is extracted from the POPD information.

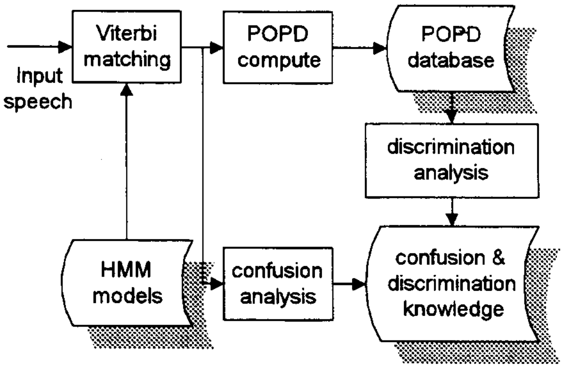

Figure 1: system block diagram for adaptation procedure

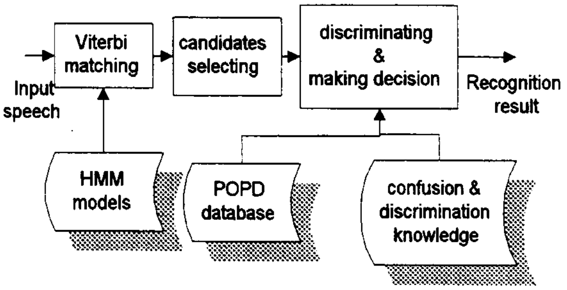

Figure 2: system block diagram for recognition procedure

The recognition procedure is illustrated in figure 2. Several candidates with the highest scores are selected after the ordinary Viterbi matching. During the process of discriminating & making decision, both the candidates and confusing classes are analyzed based on the knowledge base.

Our adaptation scheme differents itself from an ordinary adaptation system, as it is applied to the posterior stage of an basic recognizer based on the adaptation of the POPD. Therefore we call this system Adaptation Recognition based on Posterior User of POPD.

4. Experiments' Results and Comments

Two preliminary experiments have been conducted on the system mentioned above, the experiments' results are shown in .Fig. 3 and Fig. 4 .

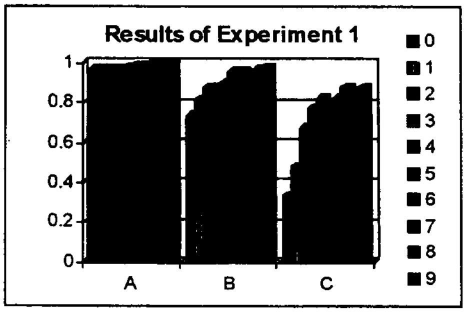

Figure 3: ten-digit experiment results

Experiment 1 is a ten-digit vocabulary recognition task in Chinese. Nine utterances per digit (which are monosyllables in Chinese) per speaker were chosen as the adaptation data set. To determine the degree of deviation of HMMs from the speaker POPD can overcome, the models were trained on utterances of several male speakers. Speaker A is one of the male speakers whose utterances were used to train the baseline model, speaker B and C are both new speakers to the system, B is a male and C is a female. In fig. 3, 0~9 are the numbers of utterances per syllable that are used for adapting the POPD.

In experiment 2, the vocabulary consists of 100 Chinese syllables including acoustic confusing syllables.

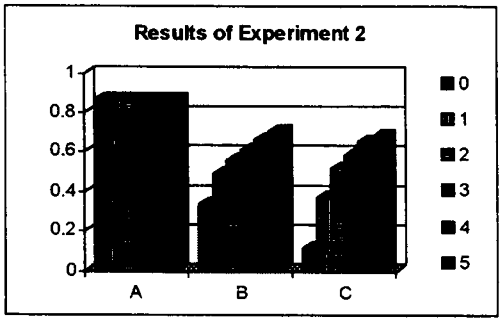

Figure 4: one-hundred-syllable experiment

The experimental results show that:

①POPD information classifies speech effectively, and suggests speaker adaptation can be done by means other than model re-estimation.

②As shown in figures 3 and 4, the recognition rate for speaker B is much higher than C's when the number of adaptation utterances equals 0, This indicates the degree of deviation of the system's HMMs is greater for speaker C than speaker B. Results show that as the number of adaptation utterances increases, the recognition rates for speakers B and C become closer. This means that the POPD information may be not much sensitive to the prior deviation, so that the pre-estimation of models may be simplified.

③The recognition accuracy for speaker A improved with POPD (although improvement is less than for speakers B and C). POPD is helpful for both speaker independent and speaker dependent recognition.

④The increase of recognition accuracy in experiment 2 is less than that in experiment 1. This is due to the larger vocabulary used in experiment 2. Future research is needed to determine how POPD can solve this problem.

⑤Adaptive decisions based on POPD do not conflict with the ordinary adaptation methods for the acoustic models. This is because the model adaptation can be considered as the adaptation in the prior stage of a recognizer, while our adaptive decisions are adaptations in the posterior stage.

5.CONCLUSION

A speaker-adaptation method was proposed where the distribution pattern of output probabilities is utilized to decide the recognition result. Results are promising, but further studies are necessary, especially on how to realize vocabulary independent adaptation based on POPD.

We also want to express our gratitude to Dr. Qiang Huo of ATR, Dr. Goh Kawai and hirose-lab of Univ. of Tokyo, Japan for their kind contributions to the work

REFERENCES

[1]J.-L. Gauvain and C.-H. Lee, "Maximum a posteriori Estimation for Multivariate Gaussian Mixture Observations of Markov Chains," *IEEE Trans. on Speech and Audio Processing, Vol. 2, pp.291-298,1994*

[2]T. Matsuoka and C.-H. Lee, "Study of On-line Bayesian Adaptation for HMM-based Speech Recognition" in *Proc. Eurospeech-93, pp.815-818, 1993*

[3]Y. Zhao, "An Acoustic-Phonetic Based Speaker Adaptation Technique for Improving speaker-Independent Continuous Speech Recognition" *IEEE Trans. on Speech and Audio Processing, Vol. 2, No.3, pp.380-384, July, 1994*

[4]Q.-Huo, Ch-Chan and C.-H. Lee, "Bayesian Adaptive Learning of the Parameters of Hidden Markov Model for Speech Recognition" *IEEE Trans. on speech and Audio Processing, VOL. 3, No.5, pp.334-345, Sept., 1995.*

[5]T. Matsuoka and S. Furui, "A Study of Speaker Adaptation Based on Minimum Classification Error Training", *Proc. Eurospeech-95, pp81-84, 1995*