論文 / 著書情報
Article / Book Information

Title	Elaborate Acoustic Modeling for Japanese Connected Digit Recognition
Author	Tatsuo Matsuoka, Naoko Uemoto, Tomoko Matsui, Sadaoki Furui
Journal/Book name	IEEE Automatic Speech Recognition Workshop, , , pp. 169-170
発行日 / Issue date	1995, 12
権利情報 / Copyright	

Elaborate acoustic modeling for Japanese connected digit recognition

Tatsuo Matsuoka†, Naoko Uemoto††, Tomoko Matsui†, and Sadaoki Furui†,††
†NTT Human Interface Laboratories, ††Tokyo Institute of Technology
3-9-11 Midori-cho, Musashino-shi, Tokyo 180, JAPAN
Tel: +81 422 59 3341, Fax: +81 422 60 7808, e-mail: matsuoka@splab.hil.ntt.jp

Introduction

Connected digit speech recognition is vital to a number of applications such as voice dialing, automatic data entry, credit card number entry, and transaction access code entry. Although grammar or language-model constraints play important roles in continuous speech recognition, they cannot be effectively used for connected digit recognition, since arbitrary sequences of digits are usually allowed in applications. Therefore, good acoustic modeling is crucial for connected digit speech recognition.

There are two approaches to improving acoustic models: one is to model co-articulation or allophones using context-dependent models and the other is to reinforce classification ability of the models by using discriminative training. Another important issue is robustness against speaker variation, channel characteristics and noise. This paper describes an elaborate acoustic modeling based on sub-word modeling to cope with co-articulation problem and Minimum Classification Error (MCE) training for connected digit speech recognition.

Sub-word modeling for connected digit recognition

Context-dependent sub-word modeling such as the *generalized triphone* is known to be effective for coping with co-articulation in continuous speech recognition[1]. If the vocabulary size is small, such as connected digits, a model unit need not be a phoneme. For such cases, longer units are better for modeling the co-articulation if a sufficient amount of training data is available.

Pieraccini et al. proposed sub-word modeling that divides a word model into three parts, *head*, *body*, and *tail*[2]. The *head* and *tail* represent the left and right contexts respectively. Since these sub-word models are usually longer than phoneme models, they can model co-articulation phenomena more accurately than phoneme models can. This technique is suitable for connected digit recognition.

Our task is to recognize Japanese connected digits of known length (four digits). The numbers from zero to nine are pronounced /zero/, /ichi/, /ni/, /saN/, /yoN/, /go/, /roku/, /nana/, /hachi/, and /kyu/. Figure 1 illustrates how sub-word models are used depending on the context. The preceding digit /roku/ has several *tails* and the succeeding digit /kyu/ has several *heads*. One of the *tails* of /roku/ and one of the *heads* of /kyu/ especially model the acoustical characteristics of the context /roku/-/kyu/. The straightforward way to design these sub-word models is to have 10 *heads* and 10 *tails* for each digit. This makes a total of 210 models, and 1000 different contexts are necessary to train them. In order to reduce the number of models, contexts are bundled if the co-articulation phenomena can be considered as being close to each other.

Contexts are bundled according to the manner of articulation, that is, fricatives, affricates, plosives, semivowels, and nasals. Since the combinations in Table 1 are thought to cause remarkable co-articulation, we first set *heads* and *tails* for those cases. Even after these rules have been applied, the number of models is still not small enough, so we applied some heuristic rules to reduce the number of models. The resulting set of models consists of 72 models, i.e., 29 *heads*, 10 *bodies*, and 33 *tails*.

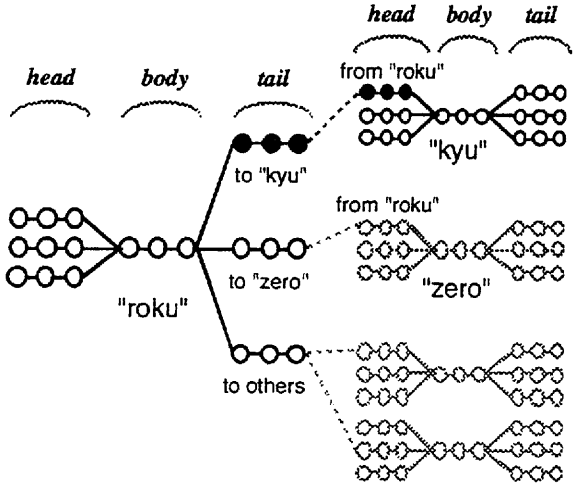

Figure 1 Sub-word model connection

Table 1 Digit combination with co-articulation

co-articulation pattern	preceding digit	succeeding digit
plosives + plosives	roku	go, kyu
affricates + plosives	ichi, hachi	go, kyu
plosives + fricatives	roku	zero, saN, hachi
affricates + fricatives	ichi, hachi	zero, saN, hachi
nasal + nasal	ni, saN, yoN, nana	ni, nana
semivowel + semivowel	zero	roku
/a/ + /a/	nana	saN, nana, hachi
/i/ + /i/	ichi, ni, hachi	ichi, ni
/o/ + /o/	go	yoN, go, roku

Minimum error training

Maximum Likelihood (ML) estimates maximize the likelihood of the models for the training data, but there is no guarantee that the parameter set is optimal for discriminating the classes. Several discriminative training methods have been proposed[3-6]. These methods directly use the classification ability of the models as the objective function. MCE training utilizes a measure of misclassification (loss) as the objective function, and it minimizes the recognition error by minimizing the loss. However, for the loss to be evaluated, competing models and their likelihood scores must be known

during the training process. If the vocabulary size is large it is rather difficult to find the competing models (N-best hypotheses) in the training. Therefore, MCE training is more feasible for a small vocabulary task such as connected digit recognition.

The discriminative function for class k is defined as the log-likelihood for the observation vector X_t given the model parameter set Λ.

$$g_k(X_t,\Lambda) = \log\left(L_k(X_t)\right) \quad (1)$$

The misclassification function is defined as

$$d(X_t,\Lambda) = -g_k(X_t,\Lambda) + G_k(X_t,\Lambda) , \quad (2)$$

where

$$G_k(X_t,\Lambda) = \log\left\{\frac{1}{K-1}\sum_{j,j\neq k}\exp\left(\eta\cdot g_j(X_t,\Lambda)\right)\right\}^{\frac{1}{\eta}} . \quad (3)$$

The first term of the right hand side of Eq. (2) is the log-likelihood of the correct class and the second term is the Lp norm of the log-likelihood of competing classes. Therefore, the larger $d(X_t,\Lambda)$ is, the more likely misclassification is to occur. The loss function is defined as

$$l\left(d(X_t,\Lambda)\right) = \frac{1}{1+\exp\left(-\alpha\left(d(X_t,\Lambda)\right)+\beta\right)} . \quad (4)$$

The total loss for all the data is

$$L(\Lambda) = \frac{1}{T}\sum_{t=1}^{T} l\left(d(X_t,\Lambda)\right) . \quad (5)$$

The optimal model parameter set is obtained by minimizing $L(\Lambda)$. The model parameter set is updated according to Eq. (6).

$$\Lambda_{n+1} = \Lambda_n - \varepsilon_n U_n \delta L(\Lambda_n) \quad (6)$$

Experiments

Utterances from 70 speakers were used for the training, and 5 speakers were used for the evaluation. Each speaker spoke 35 digit strings consisting of 4 digits. Speech data was sampled at 12 kHz, and the 16th-order LPC derived cepstrum coefficients and their first-order derivatives (delta-cepstrum) were calculated for each 8-ms frame. The acoustic models were left-to-right tied-mixture HMM with 256 prototypes (means and variances).

The number of states for HMMs was 3 states per phoneme in a digit. Four-digit strings were uniformly segmented into four segments to train initial models for the whole-word models. Then embedded training was carried out by concatenating these initial models using 4-digit strings. The initial models of the sub-word models were obtained by uniformly segmenting these whole-word models into three parts. They were trained by embedded training with context constraints.

First the batch mode training, which updates the parameters after all the data has been used for the loss calculation, was tried for MCE training, but the process did not converge. Then the training data was divided into several blocks, and the parameters were updated for each block. The process began to converge with this block-sized training, but the convergence speed was still very slow. Finally, the parameters were updated only when the recognition was incorrect. When doing this, the training data was limited and the estimation may have had errors due to the lack of training data, so the updating vectors were smoothed as follows[7]. Let $\delta\mu_i$ be an updating vector for the ith distribution of the model. The smoothed updating vector $\delta\mu'_i$ is obtained by Eqs. (7) and (8).

$$\delta\mu'_i = \frac{\sum_{j=1}^{M} w_{i,j}\cdot\delta\mu_i}{\sum_{j=1}^{M} w_{i,j}} \quad (7)$$

$$w_{i,j} = \exp\left(\frac{-(\mu_i-\mu_j)^2}{s}\right) \quad (8)$$

In addition to the above schemes, we used multiple ε_n in parallel in order to accelerate the convergence by using the most effective ε_n. The parameters to be updated were means and weights. The cepstrum parameters were updated first; then the delta-cepstrum parameters were updated. This is for stabilizing the convergence process.

Since the numbers of occurrences of the models in the training data were not equal, the total loss in Eq. (5) was normalized by the number of occurrences.

Table 2 lists the experimental results. Sub-word modeling of *head*, *body*, and *tail* greatly reduced errors. The error reduction from the whole-word model to the sub-word model was 70%. Although the MCE training did not improve the performance very much, since the sub-word models were already very good, it still improved the performance slightly, and a string error rate of 0.9% was achieved.

Table 2 Error rates

Methods	Word error	String error	Error reduction
Whole-word	0.62 %	3.44 %	-
Sub-word	0.17 %	1.00 %	70.9 %
MCE	0.14 %	0.86 %	14.0 %

Summary

This paper proposed sub-word modeling for connected digit recognition. Using *head*, *body*, and *tail* models reduced the errors by more than 70%. Furthermore, MCE training reduced the error by 14% and a string accuracy of 99.1% was achieved. We will evaluate these methods for real PSTN telephone speech incorporating robust acoustic modeling technique.

Another important issue, in addition to the acoustic modeling itself, is extraneous words. How to reject extraneous words while achieving high accuracy for valid inputs is our next study topic.

References

(1) K. F. Lee, "Context-dependent phonetic hidden Markov models for speaker-independent continuous speech recognition," IEEE Trans. ASSP, Vol. 38, No. 4, pp. 599-609, 1990

(2) R. Pieraccini, et al., "Implementation aspects of large vocabulary recognition based on intraword and interword phonetic units," Proc. DARPA Speech and Natural Language Workshop, pp. 311-318, 1990

(3) R. Cardin, et al., "High performance connected digit recognition using maximum mutual information estimation," Proc. ICASSP, pp. 533-536, 1991

(4) R. Cardin, et al., "High performance connected digit recognition using codebook exponents," Proc. ICASSP, pp. I-505-508, 1992

(5) R. Cardin, et al., "Inter-word coarticulation modeling and MMIE training for improved connected digit recognition," Proc. ICASSP, pp. II-243-246, 1993

(6) W. Chou, et al., "Segmental GPD training of HMM based speech recognizer," Proc. ICASSP-92, pp. I-473-476, 1992

(7) K. Ohkura, et al., "Speaker adaptation based on transfer vector field smoothing with continuous mixture density HMMs," Proc. ICSLP, pp. 369-372, 1992