# Speaker Independent Speech Recognition Method using Constrained Time Alignment near Phoneme Discriminative Frame

*T. Konuma[1], T. Suzuki[1], M. Yamada[1], Y. Ohno[1], M. Hoshimi [1) 3)] and K. Niyada [2)]*
tkonuma@mrit.mei.co.jp

[1)] Matsushita Research Institute Tokyo, Inc.
3-10-1 Higashimita Tama-ku, Kawasaki, 214-8501 JAPAN
[2)] Matsushita Electric Industrial Co., Ltd. [3)] Tohoku University

## ABSTRACT

We present constrained time alignment acoustic models based on phonetic knowledge and a speaker independent speech recognition method using our proposed models. Japanese syllable and isolated word recognition experiments show that the models have robustness to intra- and inter- speaker varieties such as acoustic diversity. Furthermore we experiment with word recognition tests under the condition such as noise environments and endpoints free matching, it reveals the feasibility of our proposed models.

## 1. INTRODUCTION

Recently, on speaker independent speech recognition, most approaches use hidden Markov model to absorb that intra- and inter-speaker variability. These approaches assume that appropriate acoustic models can be trained automatically with the time structure of given training speech data.

On the other hand, a phonetic subject about which temporal part of acoustic features can be cues for the identification of phonemes have been studied independently. Cooper *et al* indicated that dynamic spectral features had phoneme discriminative information [1]. Furthermore, Ide *et al* showed that explosive part was especially valid for identification of plosives and dynamic features near transitional part was for nasals [2]. Also we have proposed the effective speech recognition methods for plosive consonants and consonants of word-head and word-body [3-5]. All these researches showed that dynamic features around explosive parts or following vowels have phoneme discriminative information.

Therefore, we assumed that a speech recognition method using such phonetic knowledge should have strong robustness to intra-and inter-speaker variability [12].

## 2. PROPOSITION of NEW METHOD

We propose to employ constrained time alignment into acoustic models. The constrained time alignment is based on phonetic knowledge such as phoneme. A speech recognition method we discuses here is based on syllable acoustic models, DP matching and statistical distance measure.

### 2.1. Definition of Discriminative Frame

We describe phonetic knowledge on which the proposed acoustic models are based, and then discriminative frames for each phoneme are defined as follows (**Table 1**).

Unvoiced plosives (/p/, /t/, /k/) have dynamic spectral feature from explosive parts to the following vowel. Difference in explosive part spectrum and dynamic features near transitional part is the effective cue for /p/, /t/, /k/ identification.

Voiced plosives (/b/, /d/, /g/) have dynamic spectral feature from pre-voicing section to explosive part toward the following vowel.

Nasals (/m/, /n/, /ng/) have dynamic spectral feature from nasals murmur to the following vowel. The manner of articulation of nasal to the following vowel is the same as that of plosive case, and /m/ manner is the same as that of as /b/ and /p/ case. Similarly, /ng/ is the same as /g/ and /k/ case and also /n/ is the same as /d/and /t/.

Though it's difficult to specify an exact discriminative frame for unvoiced fricatives (/h/, /s/) due to their long stationary spectral features, we employed the same rule as the other consonants that discriminative frame is located at the end of the phonemes.

**Table 1:** Our Definition of Phoneme Discriminative Frame

| Phonemes | Phoneme discriminative frame |
|---|---|
| c, p, t, k | Explosive part |
| g, b, d, r, z | End of pre-voicing section / Explosive part |
| m, n, ng | Dynamic feature following vowels |
| h, s | End of phoneme |
| j, w | Begin of phoneme |

### 2.2. New Acoustic Models

We propose new acoustic models emphasizing dynamic spectral features near phoneme discriminative frame.

Clues for the identification of phonemes are found not only in the exact phoneme discriminative frame but also its neighbor frames including dynamic features. Therefore we propose new acoustic models for dynamic spectral features where the time structure is presented with constrained time alignment near phoneme discriminative frame.

Speaking speed effects the variability of phoneme duration. In general, to handle the variability of speaking speed, DP matching or phoneme duration control technique is used. Time alignment technique such as DP matching is employed all over the speech segments whichever is a consonant or a vowel. In such a time-alignment mechanism, duration information of each phoneme tends to be missed. But phonetic researches so far clearly show that variation in duration exists dominantly in vowel acoustically stationary part. In contrast, little variation in duration is observed in transition part from a consonant to a following vowel. Furthermore, dynamic features in this

transitional part are reported as effective for identification of the proceeding consonant. Therefore DP matching is effective in vowel parts, but not in consonant. Phoneme duration control technique is proposed for this problem. But it is difficult to control it suitably for the variation of the dynamic spectral features.

To solve such a difficult problem, we applied the phonetic knowledge concerning discriminative frame in temporal structure of speech utterance into our speech recognition method. More precisely, we think a robust speech recognition algorithm is attainable through designing phonetic models with constrained time alignment conditions based on such phonetic studies. Dynamic spectral features near the discriminative frame are effective for identification and robust for the intra- and inter speaker variability. We propose acoustic models for consonant where time structure is prohibited self-loops and skips over successive near frames of discriminative frame. And also we propose acoustic model for vowel where time structure is allowed to extend and shrink. With these mechanisms in models training and recognition testing, proposed acoustic model are expected to statistically absorb the variation of duration information.

## 2.3. Example of Newly Proposed Models

We propose the word recognition method using new acoustic models with constrained time alignment near phoneme discriminative frame.

**2.3.1. Japanese Syllable Acoustic Models** We propose new acoustic models of Japanese syllables. Syllable models were trained from isolated word speech data. The training data was labeled for phoneme endpoints and phoneme discriminative frames by labeling experts.

A syllable model for word head is prepared as different one from that for word body. The selection of the discriminative frame of each sample is carried out based on hand labeling.

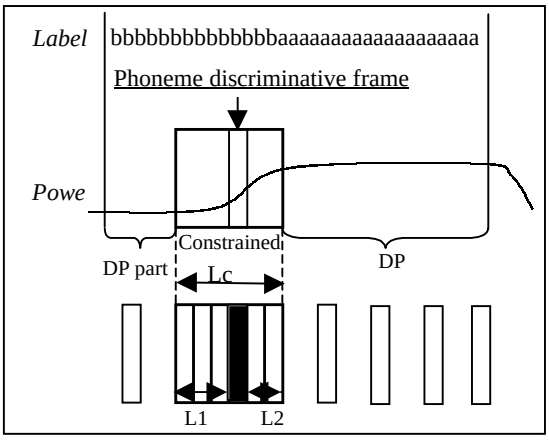


**Figure 1:** Proposed acoustic model (Ex. /ba/)

For instance, **Figure 1** shows how Japanese syllable /ba/ is trained in our newly proposed model. An explosive frame of the phoneme /b/ is selected as the discriminative frame. Then a pre-defined number of successive neighbor frames of the discriminative frame including transitional part toward the following vowel /a/ is segmented. The detail rules of the segmentation such as Lc (the length of the successive frames), L1 and L2 are defined by phonetic knowledge dependent on phonetic categories.

When training Japanese phonetic model of /ba/ from a lot of samples, successive frames with constant length are simply averaged without any time alignment and the rest part of a sample other than the successive frames are time aligned using DP technique.

**2.3.2. The Distance Measure** We use the distance measure based on Bayes' theorem for DP matching. Here each state of the model is assuming a Gaussian distribution with a mean value of feature parameters including static and dynamic acoustic parameters and its full-covariance matrix, provided that covariance value between static and dynamic parameters can be neglected.

$$d_{ij} = (x_i - \mu_j)^t W_{j-1} (x_i - \mu_j) + \log|W_j| \quad (1)$$

$$W_j = \begin{bmatrix} w_{11} & \cdots & w_{1n} & & & \\ \vdots & CEP & \vdots & & 0 & \\ w_{n1} & \cdots & w_{nn} & & & \\ & & & \Delta w_{11} & \cdots & \Delta w_{1n} \\ & 0 & & \vdots & \Delta CEP & \\ & & & \Delta w_{n1} & \cdots & \Delta w_{nn} \end{bmatrix} \quad (2)$$

$x_i$ : A feature parameter vector of input speech data *i*-frame

$\mu_j$ : Mean vector of a reference *j*-frame

$W_j$ : Covariance matrix of a reference *j*-frame (equation *(2)*)

The equation (1) indicate the distance between an input speech data frame and a reference frame.

**2.3.3. Word Recognition** In word recognition, we make word models by concatenating syllable models. When DP matching is carried out along input speech axis basis, time alignment is prohibited over the successive frames near the discriminative frame.

# 3. EVALUATION EXPERIMENTS

We show the validity of the proposed acoustic models through two evaluation experiments. One is syllable and isolated word recognition tests under endpoints-fixed conditions. This experimental result shows the validity of our proposed models that constrain time alignment near phoneme discriminative frame. And the other is a large vocabulary word recognition test in real-time on such a PC where noisy speech data was evaluated under endpoints-free conditions.

## 3.1. Experiment with Endpoints Fixed

**3.1.1. Syllable Recognition Test** We compare the efficiency of three sets of acoustic models to prove the validity of the proposed acoustic models. We used syllable test data that are selected by hand labeling experts from word test data. We show the experimental result divided into three groups of syllable including voiced consonants (/m/,/n/,/ng/,/b/,/d/,/r/,/z/), unvoiced plosives (/c/,/p/,/t/,/k/) and unvoiced fricatives (/z/,/h/,/s/).

1) Experimental conditions
Acoustic models (1) are trained by samples whose frames are all subject to time alignment using DP technique (**Figure 2**) between labeled endpoints.
Acoustic models (2) are trained by samples whose exact discriminative frames are aligned and two DP matching sessions are carried out before/after the discriminative frames between labeled endpoints and the discriminative frames (**Figure 3**).
Acoustic models (3) (the proposed models) are trained by samples, which have labels of both endpoint and discriminative frame and successive part nearby the discriminative frame without time alignment. The rest of frames before or after the successive part are time aligned using DP technique (**Figure 4**).

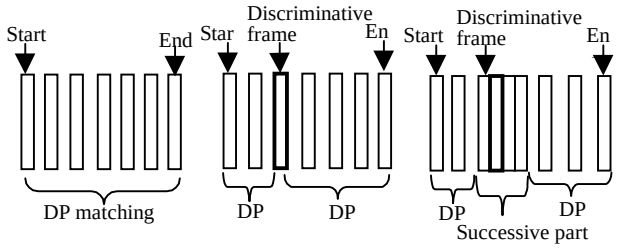


**Figure 2**: models(1) **Figure 3**:models(2) **Figure 4**:models(3)

When testing, syllable endpoints are given by label information, but location of phoneme discriminative frames is unknown. Continuous DP was carried out, where starting point was free between seven frames from three frames ahead to three frames behind of the labeled start frame and end point was free similarly. **Table 2** shows the condition of experiment. Words and speakers of the test data differ from the training data.

**Table 2:** The condition of experiment

| | |
|---|---|
| Acoustic Analysis | 12kHz sampling, 10ms frame |
| Training data | phoneme-balanced 32,580 words |
| Test data | phoneme-balanced 4,240 words |

2) Experimental result and discussion
**Figures 5** to **7** show experimental results divided into groups of syllable including voiced consonants, unvoiced plosives and unvoiced fricatives.

**Figure 6** shows that for all groups models (2) are more effective for consonant identification than models (1). This result indicates that (1) can't estimate dynamic spectral features sufficiently. Furthermore models (3) (the proposed models) are more effective for consonant identification than (2). This result indicates that (3) can estimate dynamic spectral features near phoneme discriminative frames sufficiently. On the other hand, for unvoiced fricatives group, (3) is as effective as (2). This result shows that models for unvoiced fricatives does not need to consider the successive part, because the spectrums of /s/ and /h/ are stationary and the duration of /s/ and /h/ is highly variable similarly as vowel.

**Figure 7** shows that, for all groups, (2) and (3) are less effective for vowel identification than (1). This result indicates that (2) and (3) use fewer frames for vowel than (1).

Consequently even with less efficiency in vowel identification, **Figure 5** shows the validity of our newly proposed acoustic models over the syllable recognition.

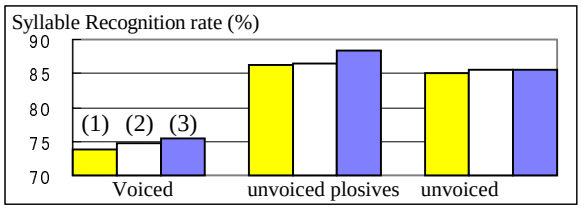


**Figure 5:** Syllable recognition rate

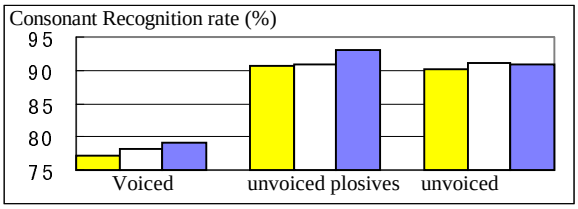


**Figure 6:** Consonant recognition rate

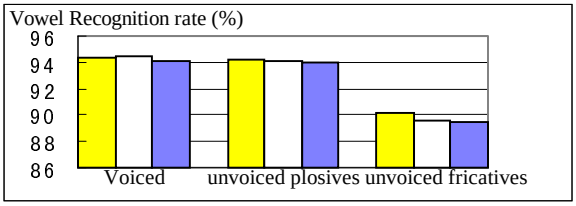


**Figure 7:** Vowel recognition rate

**3.1.2 Word Recognition Test** Using three syllable models made in 3.1.1., we experiment a large vocabulary word recognition test.
1) Experimental conditions
**Table 3** shows the condition of experiment with three acoustic models. The models were trained from Japanese phoneme-balanced 36,820 words uttered by speakers of both genders. When testing, endpoints are fixed. 100 words set each uttered by 25 male and female speakers were evaluated. The vocabulary size for recognition test was 5,168 words.

**Table 3:** The condition of experiment

| | |
|---|---|
| Acoustic Analysis | 12kHz sampling, 10ms frame |
| Training data | Phoneme-balanced 36,820 words |
| Test data | 100 city names uttered by 50 subjects |
| Vocabulary size | 5,168 words |

2) Experimental result and discussion
**Table 4** shows the experimental result of word recognition

**Table 4:** 5,168 word recognition rate (%)

| | Mean | Worst |
|---|---|---|
| models (1) (**FIG. 2**) | 88.80 | 64.0 |
| models (2) (**FIG. 3**) | 90.18 | 70.0 |
| Models (3) (**FIG. 4**) | 91.02 | 73.0 |

With the traditional syllable model [6], which was trained with free time alignment over whole frames in sample speech data, mean recognition rate 88.8% and worst speaker's accuracy rate 64.0 % were obtained. On the other hand, our proposal models which were trained under constrained time-alignment condition, mean recognition rate 91.0 % and worst speaker's one 73.0 % were obtained. Therefore, the validity of the proposed models was proved with the improvement of accuracy of 2.2 % in mean recognition rate and 9.0 % in the worst speaker recognition rate.

## 3.2. Experiment under Noisy Environments

We prove the validity of the acoustic models with constrained time alignment near discriminative frame under noisy environments. First, with proposed acoustic models, we develop a large vocabulary word recognition method, which works in real-time on such a PC. Next, we evaluate this method through the experiment with several additional noise at 20dB S/N ratio under endpoints free condition.

**3.2.1. Introduction of Word Spotting and Beam Search** We introduce word spotting and beam search into the word recognition method using the proposed acoustic models. Considering using the recognition system based on the proposed acoustic models under practical environments, noisy input speech and spontaneous speech have to be handled. Word spotting technique is effective such input speech because precise speech segmentation is not necessary.

But introducing word-spotting technique, in order to compare the likelihood of words calculated from different speech segments, we must alter the likelihood of each frame considering a posterior probability according to Bayes' theorem. And in order to decrease calculation so that it works in real-time on such a PC, we introduce beam search technique that was indicated the effectiveness on DP-matching [7-9].

We developed a large vocabulary word recognizer and reported the validity of new acoustic models on a PC [11].

**3.2.2. Word recognition test** We experiment with word recognition test to prove the validity of our proposal acoustic models under noisy conditions.

1) Experimental conditions

**Table 5** shows the experimental conditions. To test noise robustness, we compared two models trained from two different additional noises. One was the models trained from phoneme-balanced 36,820 words data with additional only exhibition hall noise at 20dB S/N ratio. The other was the models trained from 5 sets of the same word data, each set was with additional noise selected from 5 different noise at 20dB S/N ratio.

For recognition test, we made 4 test sets of the same word data, each set was with additional noise selected from 4 different noises at 20dB S/N ratio. Introducing word spotting and beam search, we also experimented with 5,168 words recognition test under endpoint free condition.

**Table 5:** The condition of experiment

| | |
|---|---|
| Acoustic Analysis | 12kHz sampling, 10ms frame |
| Training data | Phoneme-balanced 36,820 words with additional 1) or 2) noise at 20dB S/N ratio<br>X) exhibition hall noise<br>Y) 5 noise (exhibition hall, small factory, computer room, luxury car A, popularly-priced car) |
| Test data | 100 city names uttered by 50 persons with additional bellow noise at 20dB S/N ratio<br>a) exhibition hall<br>b) luxury car B (another type of a car from training data)<br>c) office<br>d) around train ticket vender |
| Vocabulary size | 5,168 words |

2) Experimental result and discussion

**Table 6** shows the mean recognition rate of 5,168 words recognition test. In experiments for 4 test data sets, model Y obtained more accuracy than model X. For **b c d** sets, which were noise open test data, recognition method using model Y increased in the accuracy of 0.8% - 5%, also for **a** sets which was noise closed test data, using model Y, the increase in the accuracy of 1% was obtained. Under noisy environment, the validity of our proposed models trained from data with additional several noise, was proved with these results

**Table 6:** 5,168 words recognition rate

| | Test data sets (recognition rate(%)) | | | |
|---|---|---|---|---|
| | **a** | **b** | **c** | **d** |
| model X (1 noise) | 89.26 | 88.08 | 78.66 | 82.40 |
| model Y (5 noise) | 90.22 | 88.80 | 83.68 | 85.00 |

# 4. CONCLUSIONS

In this paper we proposed constrained time alignment acoustic models based on phonetic knowledge and a speech recognition method using our proposed models. The validity of the proposed acoustic models was proved with the Japanese syllable and isolated word recognition experiments show that the models have robustness to intra- and inter- speaker varieties such as acoustic diversity. Next we experiments with a large vocabulary word recognition method using our proposed models, which works in real-time with beam-search technique, under the condition models such as noise environments and endpoints free matching. It revealed the practicality of the proposed acoustic models.

In future, we will study to apply the proposed acoustic models to HMM.

# 5. REFERENCE


[1] F. S. Cooper, *et al*, J.Acoust. Soc. Am. 24(6), 597-606 (1952)

[2] K. Ide, *et al.*, Trans. of the Committee on Speech Research, ASJ82, S83-61

[3] K. Niyada and M.Hoshimi, Spring Meeting, ASJ84, 2-3-17

[4] M.Hoshimi and K.Niyada, Spring Meeting, ASJ84, 3-3-1

[5] K. Niyada, *et al.*, Fall Meeting, ASJ84, 2-9-6

[6] T. Kimura, *et al*., ICSLP-92, pp.169-172(1992)

[7] H. Ney, *et al*., ICASSP87, 20.10, pp. 833-836 (April 1987)

[8] H. Sakoe, *et al*.,Trans. IEICE, Vol. J71-D, No.9, pp.1650-1659

[9] M. Kohda, Trans. IEICE, Vol. J72-D-II

[10] M. Yamada, *et al*., Technical report of IEICE, SP95-99, pp. 39-44

[11]T.Suzuki, *et al*., Autumn Meeting, ASJ97, 3-1-4

[12]T.Konuma, *et al*. ,1997 IEEE Workshop, ASRU *proceedings* pp. 458-465