A Statistical Phonemic Segment Model for Speech Recognition Based on Automatic Phonemic Segmentation

Katsura Aizawa and Chieko Furuichi

Toin University of Yokohama
1614 Kurogane-cho, Aoba-ku, Yokohama, Kanagawa, 225-8502 JAPAN

ABSTRACT

This paper presents a method of constructing a statistical phonemic segment model (SPSM) for a speech recognition system based on speaker-independent context-independent automatic phonemic segmentation. In our recent research, we proposed the phoneme recognition system using the template matching method with the same segmentation, and confirmed that 5-frame-fixed time sequence of feature vectors used as a template represents features of phoneme effectively. This time, to improve a mass of these templates to a smarter model, we introduced a statistical method into modeling. The structure of SPSM connects 5 distributions of Gaussian N-mixture density in series. By the experiment of closed Japanese spoken word recognition, using VCV balanced 4920 words spoken by 10 male adults including 34430 phonemes in total, the rate of phoneme recognition using SPSM was up to 90.23 % compared with the rate using phoneme templates, 80.39 %.

1. INTRODUCTION

In recent years, there is few research of speech recognition based on automatic phonemic segmentation because of its difficulties to achieve an effective and reliable method. It is considered to be largely due to the diversity in the acoustic properties of speech sounds arising from different inter-phonemic contexts, speaking rate, and variety of speakers.

To solve these difficult problems, we used a hierarchical segmentation method with 6 selected segmentation parameters and phonetic category labeling algorithm based on acoustic phonetic knowledge concerning continuous Japanese speech [1]. It was confirmed by experiment that this segmentation method is applicable to English continuous speech using the knowledge concerning English language [2]. And it is considered that this method may be also applicable to the other languages.

In segmentation, the phonetic category labels are decided for each segment. These labels are reliable and helpful clues for efficient recognition by reason that the derivation of these labels is based on acoustic analysis of speech data and that utilizing of these labels limits the valuable models for recognition per segment before computation of likelihood.

For the feature extraction, several model types are defined based on the phonetic labeling, and the most important 5 frames of each segment are decided for stationary part, transitional part or broad part. It is reasonable to extract features from different part because the appearance of feature is different by kinds of phoneme. To represent the feature of phoneme using such a limited number of vectors is effective for low cost computation. Furthermore, using the most suitable type of models per each segment based on phonetic category labels, we can expect high performance recognition.

As we know, there is static and dynamic acoustic feature in speech signal so that it is regarded as important to model not only static but also dynamic acoustic feature of speech. To satisfy this condition, for example, the modeling method using frame cluster as input of a model is proposed in other recent research [3].

The sequence of 5-frame-fixed feature vectors described above also satisfies this condition. It holds the order of sequence so that it can represent both static and dynamic acoustic feature of phoneme. In previous research, we proposed the speech recognition system based on template matching method [4]. Templates consist of the sequence of vectors. We confirmed by experiment that such a phoneme template could represent the feature of phoneme effectively to recognize unknown phonemes.

But in case of recognition of multiple speakers' speech, the amount of templates becomes enormous and consequently the cost of computation becomes higher. Therefore we introduced a statistical method for modeling phonemes this time. The proposed statistical phonemic segment model (SPSM) consists of sequential 5 distributions of N-mixture Gaussian connected in series trained by the sequences of feature vectors as same as that used as a phoneme template with a one-to-one correspondence of its sequence number.

To examine the performance of SPSM, the comparative experiment of 3 different methods, use of SPSM, template matching and use of basic continuous HMM, were carried out. In this paper, the methods of automatic phonemic segmentation, acoustic feature extraction, training of SPSM and recognizing are discussed first. Then the conditions of this comparative experiment and results are shown.

2. METHODS

2.1 Outline

A SPSM represents acoustic features of phonemic segments statistically using 5 sequential Gaussian mixture density distributions. The structure of a SPSM corresponds to 5-state Gaussian mixture density Hidden Markov Model with no duration and no skip of any states (i.e. only transitions concatenating neighboring states are exist).

The distributions of SPSM are trained by 5-frame-fixed time sequences of feature vectors extracted from representative 5 frames of stationary part, transitional part or broad part of the phonemic segment, which frames are decided automatically in segmentation process based on acoustic analysis.

2.2 Automatic Phonemic Segmentation

In our system, all speech data for both training and testing are segmented into phonemic units using speaker/context independent automatic phonemic segmentation precursor to training or recognition. In this method, we need not prepare any phonemic model. Only 6 segmentation parameters, which represent static/dynamic features of speech sound, and their thresholds, which are set up and fixed to the most suitable value in consequence of preparatory experiments using other speech data, are used. In this paper, we used this method optimized for Japanese speech [4].

The process is performed hierarchically. First voiced regions of input speech data are detected. Then voiced regions are segmented into phonemic units. After that silence segments are detected from non-voiced regions, and finally the remaining regions are segmented into phonemic segments.

At the same time, one of phonetic category labels as shown below is decided to every phonemic segment based on the behavior of segmentation parameters.

- *C*: consonant (for a voiced region)
- *W*: vowel (for a voiced region)
- *V*: voiced (for a voiced region)
- *F*: fricative (for an unvoiced region)
- *U*: unfricative (for an unvoiced region)
- *S*: silence

These labels indicate the approximate acoustic feature of a phonemic segment without recognition process.

2.3 Definition of Model Types

It is generally known that the features of phoneme appear at stationary part, transitional part or broad part of the phonemic segment by kinds of phoneme. Therefore it is ideal to use the most suitable type of models for recognition of each phonemic segment. In our system, several model types are defined based on the sequence of phonetic category labels. Table 1 shows the definition of model types for Japanese speech.

For example, if the target segment is labeled as *C* and the following segment is labeled as *W*, the models of both type C and type CW are adopted for both training and recognition.

2.4 Extraction of Acoustic Feature Vectors

Figure 1 shows the positions for the extraction of acoustic feature vectors. The vectors are extracted from delegated 5 frames which positions are decided related to the applied model type of each segment. For the type of stationary part (C/W/U), the frame where the change of acoustic feature is the most stable in a target segment is used as the center of 5 frames (i.e. the third frame), and the other frames are chosen at intervals of 1 frame. For the type of transitional part (CW/UW/SV), the phonemic boundary frame between a target segment and the next segment is applied for the center frame, and the other frames are chosen at the same intervals. For the type of broad part (X/F), first we divide the whole target segment into 6 equal portions, and then pick up sandwiched 5 frames with interpolation.

Table 1: Definition of model types based on the sequence of phonetic category labels (for Japanese speech). Model types of X and F are used only for the phonemic segments which time length is longer than 60 msec.

		Phonetic category label	
Modeling part	**Model type**	**Target segment**	**Following segment**
Stationary	C	*C* or *V*	Any
	W	*W* or *V*	Any
	U	*U* or *F*	Any
Transitional	CW	*C*	*V* or *W*
	UW	*U* or *F*	*V* or *W*
	SV	*S*	*V* or *W*
Broad	X	*U* or *F*	*S*
	F	*U* or *F*	Not *S*

2.5 Training Method

Figure 2 shows the image of training method. The time length of every training time sequence of feature vectors is fixed to 5 frames. The n-th vector in a sequence is used as a sample data for the estimation of the n-th distribution in SPSM.

The maximum mixture number is set previously, but it would be reduced if any diagonal element of the covariance matrix falls under 0.1 while training.

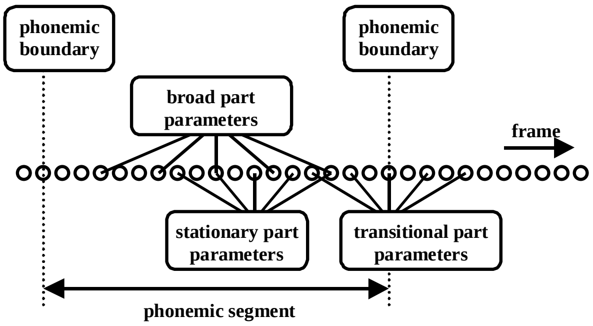

Figure 1: Positions of the extraction of acoustic feature vectors. The train of small circles represents a sequence of frames.

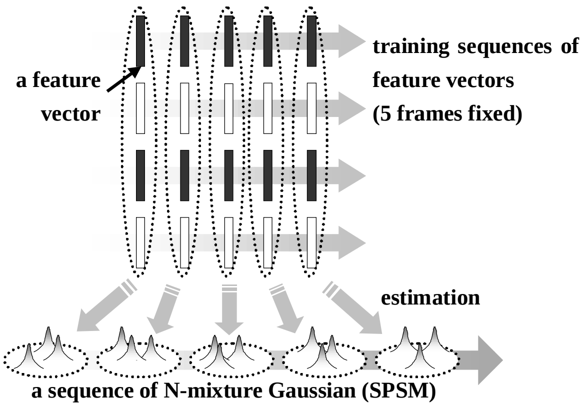

Figure 2: The image of training method of SPSM.

2.6 Recognition Method

In recognition, total output probability of each model belonging to applied model type is computed for each segment as the likelihood.
Let the sequence of feature vectors be defined as below.

$$\mathbf{V} = \mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_5$$

If a reference model **M** consists of *n*-mixture Gaussian, the formula of the total output probability *P*(**V**|**M**) is

$$P(\mathbf{V} \,|\, \mathbf{M}) = \prod_{i=1}^{5} \sum_{j=1}^{n} b_{i,j} N(\mathbf{v}_i ; \mathbf{u}_{i,j}, \mathbf{s}_{i,j})$$

where $b_{i,j}$ is the weight of j-th mixture in i-th distribution. For the ignored mixture, $b_{i,j}$ is set to zero. And $N(\mathbf{v};\mathbf{u},\mathbf{s})$ is a multivariate Gaussian with a mean vector **u** and a covariance matrix **s**.

3. EVALUATION EXPERIMENT

3.1 Conditions of Experiment

The experiment of phoneme recognition was carried out using VCV balanced 4920 Japanese words spoken by 10 males including 34430 phonemes in total for both training and testing. The speech data are processed by the sampling frequency of 10 kHz, the quantization by 12 bits, the frame length of 25.6 msec, Blackman window, and unbiased estimation of log spectrum [5]. For comparison of the SPSM method with the template matching (TMPLT) method and HMM method using basic continuous HMM with a diagonal covariance matrix, these three methods are used under the same conditions of speech data processing.

Note that in case of using the template matching method, every test pattern must be identical with one of the reference templates. So the identical reference template with test pattern is ignored to compute the likelihood in every time of recognition of a segment.

And note that in SPSM's and TMPLT's the automatic phonemic segmentation is carried out before training and recognition, while in HMM's, phonemic segments for training were decided manually.

3.2 Experimental Results

Table 2 shows the rate of phonemic recognition to the maximum number of mixtures of SPSM. By the result, we found that the SPSM with 6 mixtures maximum produced the best result in total, 90.23%. Speaker M19 produced the highest change in which the highest rate was 91.42 % to 6 mixtures' SPSM. On the other hand, speaker M15 produced the lowest change in which the highest rate was 88.60 % to 7 mixtures' SPSM.

Table 3 shows the best rate of phonemic recognition to three different methods of SPSM's, TMPLT's and HMM's. In HMM method, the model using different numbers of states and mixtures were tested. The range of state number was from 3 to 5, and the range of maximum mixture number was from 3 to 9. Then 3-state 7-mixture-maximum model produced the best rate of recognition.

By the result, we found near 10 % improvement at the recognition rate of SPSM's upon the one of TMPLT's. Moreover, the performance of SPSM is almost the same as HMM in respect of recognition rate. As regards the cost of computation, SPSM is superior to the other two.

Table 2: The rate of phoneme recognition (%Correct) to the number of mixtures using SPSM. VCV balanced 492 Japanese words spoken by each speaker are used for both training and testing.

	Mixtures			
Speaker	**4**	**5**	**6**	**7**
M01	88.65	89.46	**90.29**	89.90
M03	88.16	88.80	89.51	**89.70**
M10	89.25	89.73	90.38	**90.86**
M15	87.47	88.33	88.52	**88.60**
M16	89.13	89.54	90.18	**90.82**
M18	88.58	89.43	**89.84**	89.79
M19	91.60	91.19	**91.42**	91.19
M20	89.58	90.61	**91.37**	91.37
M41	89.46	90.42	**90.83**	90.11
M42	88.59	89.43	**89.76**	89.63
TOTAL	89.05	89.71	**90.23**	90.22

Table 3: Comparison of the best rates of phoneme recognition (%Correct) using 6-mixture statistical phonemic segment model (SPSM), template matching method (TMPLT) and 3-state 7-mixture Hidden Markov Model (HMM). Note: For TMPLT, as test data is identical with training data, the template identical with the input pattern is ignored at every time of recognition of the segment.

Speaker	SPSM	TMPLT	HMM
M01	90.29	81.55	**90.42**
M03	**89.51**	78.97	89.02
M10	90.38	81.21	**91.17**
M15	**88.52**	75.97	88.24
M16	**90.18**	76.27	87.39
M18	**89.84**	81.44	89.31
M19	**91.42**	83.99	90.62
M20	**91.37**	85.29	91.00
M41	**90.83**	79.88	89.75
M42	**89.76**	77.97	88.00
TOTAL	**90.23**	80.39	89.49

4. CONCLUSIONS

In this paper, we proposed the statistical phonemic segment model (SPSM) as the alternative smarter model of the phoneme templates what we used in recent researches based on automatic phonemic segmentation. The training sequence of feature vectors for SPSM is the same with the one for phoneme templates, so that almost the same or higher performance of phoneme recognition was expected for SPSM method compared with the template matching method. By the comparative experiment of closed phoneme recognition, we found that SPSM method can produce near 10 % higher rate of phoneme recognition than the use of templates, and that is almost the same performance with the use of basic continuous HMM. Regarding the simple structure of SPSM, this proposed model has large advantage of low cost of computation.

REFERENCES

1. C. Furuichi and S. Imai, “Phonemic Units Segmentation in Various Phonetic Environments,” *(in Japanese) Trans. IEICE*, vol. J72-D-II, no. 8, pp.1221-1227, Aug. 1989.
2. C. Furuichi, K. Aizawa and S. Imai, “Automatic Phonemic Segmentation System of English Continuous Speech by Using Static/Dynamic Parameters,” *(in Japanese) Trans. IEICE*, vol. J78-A, no. 3, pp.295-304, Mar. 1995.
3. S. Nakagawa and K. Yamamoto, “Speech Recognition by Hidden Markov Model Using Segmental Statistics,” *(in Japanese) Trans. IEICE*, vol. J79-D-II, no. 12, pp.2032-2038, Dec. 1996.
4. C. Furuichi and S. Imai, “Speaker-Dependent Phoneme Recognition of Unspecified Vocabulary Japanese Speech,” *(in Japanese) Trans. IEICE*, vol. J73-D-II, no. 4, pp.501-511, Apr. 1990.
5. S. Imai and C. Furuichi, “Unbiased Estimation of Log Spectrum,” *(in Japanese) Trans. IEICE*, vol. J70-A, no. 3, pp.471-480, Mar. 1987.