ESTIMATION OF STATISTICAL PHONEME CENTER CONSIDERING PHONEMIC ENVIRONMENTS

Shigeki Okawa and Katsuhiko Shirai

Department of Information and Computer Science, Waseda University
3-4-1, Okubo, Shinjuku, Tokyo, 169 Japan
E-mail: {ookawa, shirai}@shirai.info.waseda.ac.jp

ABSTRACT

This paper presents a new scheme of acoustic modeling for speech recognition based on an idea of Statistical Phoneme Center. The Statistical Phoneme Center has several properties that are feasible to realize a higher-reliable phoneme extraction. First, we assume that there is a fictitious center point in every phoneme. The center is determined statistically by an iterative procedure to maximize the local likelihood using a large amount of speech data. Next, in order to evaluate the performance of phoneme extraction, phoneme recognition is realized by optimizing the likelihood based on Dynamic Time Warping technique. As the experimental result, 71.6% recognition accuracy is obtained for speaker independent phoneme recognition. This result demonstrate that the proposed SPC is a new effective concept to obtain more stabilized acoustic model for speaker independent speech recognition.

1. INTRODUCTION

Improving acoustic model is an essential problem in the research of speech recognition. Recently, several statistical methods including Hidden Markov Modeling are generally employed as acoustic modeling. There is, however, a difficult problem in the speech pattern, which is fluctuated by various factors and the distribution is overlapped between the categories. Although current methods are very powerful, it is desired further to improve the basic technique.

In order to realize a robust speech recognition system for large vocabulary, speaker independent, and continuous utterance conditions, a sub-word often used as a unit of the acoustic model. In the case, the problems of co-articulation and allophones are unavoidable. Therefore, a phoneme extraction method that is not dominated by such problems as well as possible is desired. Especially for speaker independent condition, the recognition performance becomes lower unless using the model that expresses variable environments.

In our previous studies, we have investigated an effective estimation method of Statistical Phoneme Center (SPC) from acoustic features and realize phoneme and word level recognition by uniting the conventional HMM technique, where we assumed context independent SPCs [1].

To solve a problem of fluctuation depending the phonemic contexts, in this paper, we attempt to extend and accurate the SPC, in which the phonemic environment is considered. Furthermore, we evaluate the effectiveness of the extended SPC by DTW-based phoneme recognition experiments.

2. STATISTICAL PHONEME CENTER

When a large amount of speech data with the uttered texts is given, it is possible to determine a point statistically that includes the most stabilized information of the phoneme existence. Here we call the point *Statistical Phoneme Center* (SPC). We define the SPC as follows: (1) Every phoneme has its own SPC. (2) The SPC is determined simply by the surrounding acoustic information.

The problem to associate speech signal with phoneme category has been considered over the past years. Especially during the 1970s many researchers investigated the process of speech perception by hearing experiments [2]. The idea proposed in several decades ago was that every phoneme has the invariant cue that expresses the typical characteristics. Until now, however, the exact solution has not been obtained.

The idea of SPC discussed in this paper seems to be analogous to the classical idea. Nevertheless, the novelty of this study is to assume a fictitious center point for each phoneme. It does not mean necessarily the most remarkable point that exhibits the special property of each phoneme in classical sense. First, it is fictitiously defined and determined as well by a statistical procedure. Second, it is not related directly to some physical characteristics as seen in spectrum, but it reflects complex mixed properties found in speech sound during considerably a long interval.

The advantages of considering the SPC are: (1) Since the SPC exists around the most possible region, more stabilized phoneme extraction can be realized. (2) Unnatural settlement of phoneme boundary is unnecessary at the training of acoustic models. (3) The likelihood that expresses the phoneme existence can be calculated simply for every frame.

2.1. Calculation of SPC Likelihood

We define the SPC likelihood that shows a probability of SPC existence. It is calculated in statistical way using *a posteriori* probability considering the surrounding acoustic information.

When an acoustic vector sequence X_t (t: time) is obtained, conditional probability $p(y_j^*|X_t)$ could be defined by collecting the events that y_j^* (the SPC of phoneme Y_j) exists at a time t.

Since we normally cannot estimate the probability distribution from speech input itself, the problem results in the maximization of the right side of Bayes's law.

$$p(y_j^*|X_t) = \frac{p(X_t|y_j^*)p(y_j^*)}{p(X_t)} \quad (1)$$

where $p(y_j^*)$ is *a priori* probability that the phoneme Y_j occurs, and this is based on a linguistic condition. $p(X_t|y_j^*)$ is a conditional probability that the acoustic feature X_t is observed in the assumption of SPC y_j^*. This can be estimated by observing a large number of speech data. Since $p(X_t)$ in the denominator is independent from Y_j, it is unrelated to the maximization of the numerator.

In actual speech patterns, the events to determine the phonemics are distributed around the time t. It is desired, therefore, to calculate a probability $p(y_j^*|..., X_{t-1}, X_t, X_{t+1}, ...)$. The calculation is, however, not realistic because it needs a huge amount of training data and quantity of calculation. So the calculation discontinues over L frames from the aiming point, and an approximation, which the acoustic feature at every time is independent together, is imported.

This approximation means the average of conditional probabilities given by the surrounding frames. Therefore the number of considering frames L should be decided carefully.

$$\begin{aligned} & p(y_j^*|..., X_{t-1}, X_t, X_{t+1}, ...) \\ \approx\ & p(y_j^*|X_{t-L}, ..., X_t, ..., X_{t+L}) \\ \approx\ & p(y_j^*|X_{t-L}) \cdots p(y_j^*|X_t) \cdots p(y_j^*|X_{t+L}) \end{aligned} \quad (2)$$

The y_j^* that maximizes this probability provides a result of phoneme extraction at the time t. The SPC of phoneme Y_j can be decided by using the local maximum of the probability.

Here we define the *SPC Likelihood* that the SPC y_j^* of phoneme Y_j exists at the time t as follows. X_t is a feature vector at the time t and L means the considering frames.

$$S^*(t, Y) = \frac{1}{2L+1} \sum_{k=-L}^{L} \log p(y_j^*|X_{t+k}) \quad (3)$$

2.2. SPC Re-estimation Algorithm

To estimate the probability distribution of $p(y_j^*|X_t)$ for training data, it is necessary to provide the true point of SPC as a teacher signal. However it is unknown for the initial data. Therefore, we apply the following iterative procedure for a large amount of speech data to determine the SPC.

1. Set an initial SPC point of y_j^* for each phoneme in the training data based on the uttered text.
2. Estimate distributions of feature $X_{t\pm k}$ around y_j^*.
3. Calculate the SPC likelihood $S^*(t, Y)$ for each phoneme.
4. Move y_j^* one frame toward the local maximum (within $\pm$ 25ms from original y_j^*) of the likelihood.
5. Iterate steps 2 – 4 until y_j^*'s convergence.

Although the phoneme category for the SPC basically depends on Japanese orthography, we provide several rules for the following acoustic events.

(1) Long vowels have two SPCs at an interval over 30ms. (2) Affricates (/ch/ and /ts/) have two SPCs corresponding with the burst and the fricative parts. (3) For unvoiced vowels, in each case of CVC or CV where V is {/i/, /u/} and C is {/p/, /t/, /k/, /s/, /h/}, the SPC of the middle vowel is not assigned.

Figure [IMAGE A371G01.GIF] (in CD-ROM) shows an example of the estimated SPCs and their likelihood.

3. PHONEMIC ENVIRONMENTS

It is well-known that the variation of phoneme characteristics is a significant problem in speech recognition. For specific phonemes, the properties are frequently diverged by the contexts or the environments. To realize higher accurate phoneme extraction, therefore, phoneme context dependent model is often employed.

When we consider more detailed environments, however, the number of basic models increases and larger amount of training data are necessary at the parameter estimation.

Here we apply the *tri-phone* structure as a context dependent SPC model, which contains a phoneme and its front and behind phonemes. We first calculate the probability distribution of the SPC for each combination of the tri-phone unit. Then several units including more similar properties are integrated successively using the above mentioned SPC likelihood as an evaluating measure. The integration is executed according to the quantity of training data. By iterating the maximization of the likelihood and the integration, we can obtain an optimal set of SPCs. Thus the distinction of the SPCs is realized considering their phonemic environments.

SPC Integration Algorithm

1. Estimate the SPC for each combination of $Y_j^{(n)} = \{Y_j^-, Y_j, Y_j^+\}$ where Y_j is the target phoneme, Y_j^- is the preceding phoneme, Y_j^+ is the following phoneme and n is the number of combinations.
2. Integrate two sets $\{Y_j^{(a)}, Y_j^{(b)}\}$ that have the most similar characteristics in all Y_j. The similarity is evaluated by using the following conditions for the same data-set.
   - The difference of SPC likelihood $< \varepsilon_s$
   - The distance of SPCs $< \varepsilon_d$
3. Re-estimate SPCs for the integrated set $Y_j^{(a+b)}$.
4. $n = n - 1$; iterate from step 2 until the target model size.

Since we have to consider the difference of SPC positions at the integration of two SPCs, the SPC is re-estimated in step 3 severally.

4. PHONEME RECOGNITION ALGORITHM

Next, we apply the SPC likelihood to One Pass DP based continuous phoneme recognition. It is the simplest way to use the SPC likelihood as a resemblance measure of the DTW. Only the connectivity of two phonemes is considered as a linguistic knowledge [3]. The recognition algorithm is as follows:

1. Symbols

   $S^*(t,j)$: SPC likelihood of phoneme Y_j at a time t.
   $G(t,j)$: optimal cumulative likelihood till t.
   $C(j,k)$: boolean connectivity from phoneme Y_j to Y_k.
   $B(t,j)$: array for back-track.
   $P(t,j)$: array for phoneme connectivity.

2. Initialization

   for all Y_j:

   $$B(0,j) = P(0,j) = 0; \quad G(0,j) = S^*(0,j)$$

3. **for** $t = 1, 2, ..., T$ (input frame)
4. **for all** y_j (phoneme category)
5. $$\hat{j} = \underset{k}{\text{argmax}} \{G(t-1,k) \cdot C(k,j)\}$$
   $$G(t,j) = G(t-1,\hat{j}) + S^*(t,j)$$
6. $$B(i,j) = (\hat{j} = j) \ ? \ B(i-1,j) \ : \ i-1$$
   $$P(i,j) = (\hat{j} = j) \ ? \ P(i-1,j) \ : \ \hat{j}$$
7. Back-track

   $\hat{j} = \underset{j}{\text{argmax}}\, G(I,j), \quad i = I$

   while $i > 0$

   output $\hat{j}$

   $j = P(i,\hat{j}), \quad i = B(i,\hat{j}), \quad \hat{j} = j$

To evaluate the recognition performance, we consider three kinds of errors; *S*ubstitution, *D*eletion, and *I*nsertion versus *C*orrect phonemes. Then we define PC (*percent correct*) = $C/(C+S+D)$, and RA (*recognition accuracy*) = $(C-I)/(C+S+D)$.

5. EXPERIMENTS

As for the experiments, we employ multi-speaker (10 males) word data (5,240 words) in ATR Japanese speech database [4]. The data were digitized at 12kHz, and 16 dimensional mel-cepstrum and the linear regressive coefficients were calculated every 5ms, using a 21.3ms Hamming window.

The experiments in this paper are in the speaker independent condition, in which the training is performed by 9 speakers and the test is by another speaker. We experiments three combinations and each result shown later is the averaged value.

We use 30 categories of Japanese phoneme; {a, i, u, e, o, p, t, k, b, d, g, s, sh, h, f, z, dj, ch́, ch, tś, ts, m, n, N, w, y, j, r, sil}.

To calculate the distribution of $p(y_j^*|X_t)$, we employ diagonal mixture Gaussian distribution (4 mixtures). The number of considering frames is decided as $L = 3$ by a preliminary experiment to evaluate the mutual information between acoustic features and phoneme categories [3].

5.1. SPC Convergence

In order to verify the proposed SPC, first, we investigate the change of SPCs and their likelihood by applying the above mentioned iterative procedure for a large amount of speech data.

Since the uttered text is given for the training data, convergence of the SPC is warranted. Besides the summation of the likelihood for all phonemes in training data takes ordinarily minus value, upper bounded, and monotonically increasing. To evaluate the convergence by the actual speech data, we examine the SPC likelihoods at each point of 20 times iterations.

Figure 1 shows the experimental results using 26,200 words for 10 speakers, which SPC model is trained by other 26,200 words for all speakers. Here the first iteration is provided by the likelihood trained using the initial SPC-set. The results show that the SPC position (averaged length between the final SPC position) and their likelihood are increased and converged by the iterative procedure.

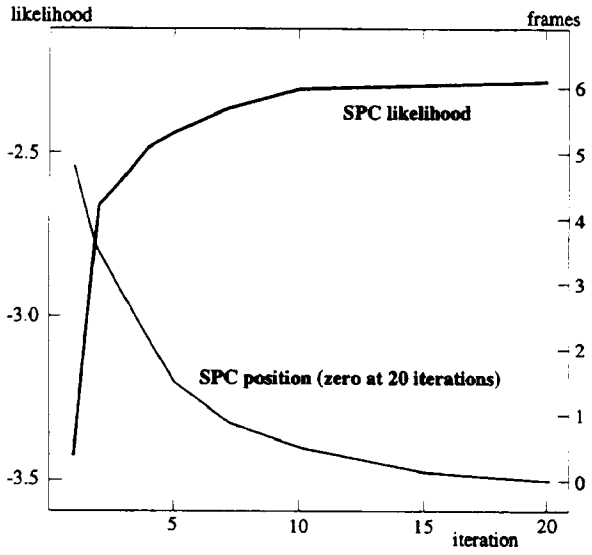

Figure 1: Convergence of the SPCs and their likelihood.

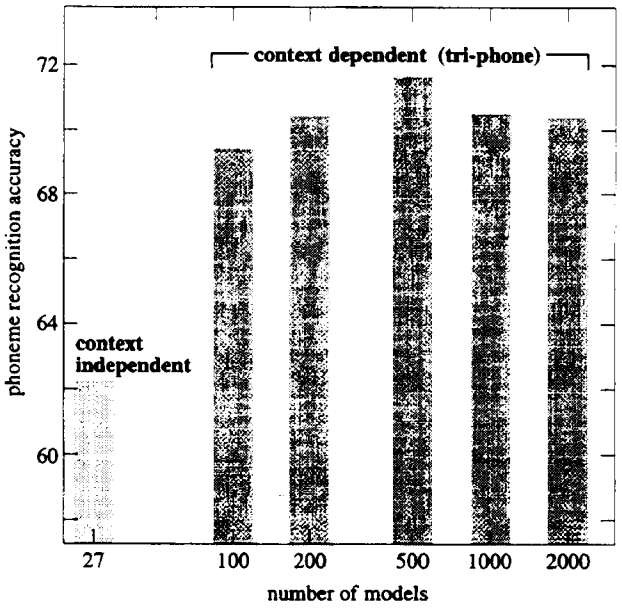

Figure 2: Results of phoneme recognition.

5.2. Phoneme Recognition

After 20 times iterative training, the SPC models are created by the obtained SPCs, then phoneme recognition is experimented for the test data-set.

Figure 2 shows the results by the context dependent SPC as compared with the former context independent models.

In case of the context dependent, the SPC integration is iterated until the target model size, where the initial model number is all kinds of tri-phone combinations (here 2,131) in the training data. Since the target model size is due to the linguistic condition and quantity of the training data, we provide several sizes in the experiments.

In the figure, though it is natural that the recognition performance of the context dependent condition is higher than the context independent one, it is also ascertained that the reduction of the performance is hardly observed by applying the SPC integration algorithm.

6. CONCLUSION

In this paper, first, we have presented a new concept called Statistical Phoneme Center to realize higher-reliable phoneme extraction, and showed an algorithm to estimate the SPC by the iterative training. Then, we have proposed an extension of the SPC to consider phonemic environments using the SPC integration algorithm. As the experimental results of DTW-based phoneme recognition, it is demonstrated that the proposed method is effective for stabilized phoneme modeling and phoneme recognition.

The SPC is also very suggestive to clarify the basic characteristics of phonemes. As some future works, we would like to investigate on some analysis of SPC in a point of view of speech perception and prosody.

ACKNOWLEDGMENTS

The authors are grateful to the members of Laboratory for Spoken Language Processing of Waseda University for their help and discussions. This work is partly supported by *the Grant-in-Aid for Scientific Research from the Ministry Education, Science and Culture of Japan*, No. 05241103.

REFERENCES

1. Okawa S. and Shirai K., "Estimation of Statistical Phoneme Center and its Application to Accurate Phoneme Modeling," *Proc. EUROSPEECH*, TUpm2E.7, 791-794, 1995.
2. Stevens K. N., and Blumstein S. E., "Invariant Cues for Place of Articulation in Stop Consonants," *J. ASA*, 64, 5, 1358-1368, 1978.
3. Okawa S., Kobayashi T. and Shirai K., "Phoneme Recognition in Various Styles of Utterance Based on Mutual Information Criterion," *Proc. ICSLP*, 31.20, 1911-1914, 1994.
4. Kuwabara H., *et al.*, "Construction of a large-scale Japanese speech database and its management system," *Proc. ICASSP*, S10b.12, 560-563, 1989.