

# Statistical Estimation of Phoneme's Most Stable Point Based on Universal Constraint

Shigeki OKAWA

Department of Network Science  
Chiba Institute of Technology, Japan  
okawa@net.it-chiba.ac.jp

Katsuhiko SHIRAI

Department of Computer Science  
Waseda University, Japan  
shirai@waseda.jp

## Abstract

In this paper, we present a statistical approach for phoneme extraction based on *universal constraint*. Inspired by former phonological studies, we assume a fictitious point in each phoneme that exhibits the most stable information to explain the phoneme's existence. With the universal constraint of phoneme definitions, the point is statistically estimated by an iterative procedure to maximize the local likelihood using a large amount of speech data. We also mention a context dependent modeling of the proposed approach and its integration strategy to obtain more stability. The experimental results show favorable convergencies of both the fictitious points and their likelihoods, which give usefulness for the stable phoneme modeling.

## 1. Introduction

In order to find accurate phonetic features, phoneticians and phonologists have pursued universal characteristics of phonemes since ancient years. In classical phonology, represented by Saussure, Trubetzkoy and Prague School of Linguistics, a phoneme was determined by oppositions of distinctive features. In other words, they believed that different phonemes had some physical differences which human beings could distinguish. Afterward, Courtenay *et al.* defined a phoneme as a mental unit which belongs to the same phonetic space [1]. In generative phonology by Chomsky *et al.*, they gave a proposition that the phonemic system was motivated phonetically and discussed the natural generative process of phonemes [2].

Behind those former discussions, there was a simple idea that each phoneme has its characteristics universally and it had plain boundaries between the neighboring phonemes. In 1970s, the problem of phoneme extraction was expanded to the field of speech signal processing, which aimed to automatic speech synthesis and recognition. By that time, many hearing experiments were performed to clarify human perceptual structure. Researchers gradually found a large discrepancy between phoneme descriptions and acoustic characteristics brought by the sound spectrograph.

In recent years, people tend to believe that the information of a phoneme just spreads over time and human perceptual organs somehow unify them to recognize speech. Pursuing the universal characteristics of phonemes is considered an old topic. However, it is also interesting to apply statistical processing by present computer technologies to solve the problem. Because those old experiments had poorer conditions in terms of the quantities of data and computation.

---

This work is partly supported by the *Grant-in-Aid for Young Scientists from Japan Society for the Promotion of Science*, No. 13780302.

In this study, we attempt a phoneme extraction based on universal constraint by a statistical processing using a large amount of speech data. Note that the authors formerly proposed "Statistical Phoneme Center" (SPC) to improve acoustic models for speech recognition [3]. Whereas the SPC also had the similar strategy, this paper mainly aims to phoneme extraction in phonological point of view.

## 2. Universal Constraint of Phonemes

While it is very difficult to find universal characteristics of phonemes, it is rather easy to define universal constraints by a general phoneme categorization. Although we consider only Japanese phonemic system here, the same approach could apply to another language, at least it has the phoneme definitions.

The constraints given in this study are as following:

- (1) Phoneme categories basically depend on Japanese orthography.
- (2) Affricates /ch/ and /ts/ have two parts each corresponding to the burst and the fricative.
- (3) Palatalized semivowels are expressed by /j/.
- (4) For unvoicing of vowels, in cases of CVC and CV where V is {/i/, /u/} and C is {/p/, /t/, /k/, /s/, /h/}, the vowel could be omitted.
- (5) Long vowels (e.g. /aa/) have two parts at an interval over 20 ms. This is just because of the convergence properties of the estimation algorithm, which is introduced later, and not by any phonological grounds.

By these constraints, we use 29 phoneme categories; {a, i, u, e, o, p, t, k, b, d, g, s, sh, h, f, z, dj, ch, ts, m, n, N, w, y, j, r, sil}. Some exceptional events, which occur sometimes in Japanese language, are dealt with the context dependent models (see Section 4).

Next, let us consider how to express the universal characteristics of phonemes. Needless to say, acoustic signal patterns corresponding to a phoneme fluctuate very much by the surrounding environments. Therefore, although a phoneme is regarded as the same category in terms of the perception, the observed speech signal is sometimes completely different. By thinking this explicit fact, the assumption that every phoneme has physically invariant cues that express the typical characteristics of the phoneme, seems senseless and thoughtless. However, it is possible to determine a point which has most stable information to explain the phoneme's existence by a statistical investigation of relations between acoustic signal patterns and phoneme categories using the universal constraints.

By the way, current methodologies of speech synthesis or recognition basically assume boundaries of phonemes to create

models, whether consciously or not. The boundaries are usually determined by hand- or automatic labeling. In other words, it is necessary to solve a problem of segmentation (to detect the boundaries) in the process of phoneme modeling. When other subword unit such as syllables or semi-syllables is used instead of phoneme, we have the same problem.

In this study, we solve the problem in a little different way in which the above mentioned Most Stable Point (called MSP hereafter) is estimated. The idea of MSP might seem to be analogous to the classical idea. Nevertheless, the novelty of this study is to assume a fictitious point for each phoneme. It does not mean the most remarkable point that exhibits some special properties of the phoneme in a classical sense. The MSP is fictitiously determined by a statistical procedure. Also, it is not related directly to some physical characteristics as seen in spectrum, but it reflects complex properties found in speech sound for a long period.

### 3. Most Stable Point of Phoneme

In this section, we first introduce a phoneme extraction method when the MSP's are known; then propose an estimation algorithm of the MSP.

When an acoustic feature vector sequence (e.g. FFT or LPC cepstrum)  $X_t$  ( $t = 0, 1, \dots$ ) is obtained, the conditional probability  $p(Y|X_t)$  could be defined by collecting the events that the MSP of phoneme  $Y$  exists at  $t$ .

Since we normally could not estimate the probability from the observed speech itself, the problem results in the maximization of the right side of Bayes's law:

$$p(Y|X_t) = p(X_t|Y)p(Y)/p(X_t), \quad (1)$$

where  $p(Y)$  is a *a priori* probability, brought by a linguistic condition, that the phoneme  $Y$  occurs.  $p(X_t|Y)$  is a conditional probability that the acoustic feature  $X_t$  is observed with the assumption of the MSP of phoneme  $Y$ . This probability could be estimated by collecting a large amount of data. Since  $p(X_t)$  in the denominator is independent from  $Y$ , it is not important in the maximization of the numerator.

Since the acoustic feature  $X_t$  is obviously effected by the coarticulation, not only  $p(Y|X_t)$  but also the preceding features should be considered in the phoneme extraction process. The simple Markov probability  $p(Y|X_{t-1}, X_t)$ , in which the previous one frame is considered, is widely used in the framework of well-known Hidden Markov Model, the most common algorithm in current ASR systems.

In speech recognition, one phoneme corresponds to several frames, so that the HMM probability calculation is more efficient. Here we just want to decide one point (over the time) for a phoneme. In actual speech patterns, since the events to determine a phoneme are distributed *around*  $t$ , it is desired to calculate:

$$p(Y|\dots, X_{t-1}, X_t, X_{t+1}, \dots). \quad (2)$$

It might seem strange to import the future information  $X_{t+k}$  to explain the event at  $t$ . This is one of the reasons that the proposed MSP is "fictitious." The calculation of (2) is, however, not realistic because it needs huge amount of training data and computation. So the calculation discontinues over  $\Delta t_p$  before and  $\Delta t_n$  after the aiming point, and the probability is defined by a segment over  $(\Delta t_p + \Delta t_n + 1)$  frames. Also an approximation, which the feature  $X_t$  at each  $t$  is independent together (which means the correlation between neighboring frames is 0), is imported. This approximation means the averaging of conditional probabilities given by the surrounding frames. Therefore the

number of considering frames  $\Delta t_p$  and  $\Delta t_n$  should be decided carefully. Thus, the probability is:

$$\begin{aligned} & p(Y|\dots, X_{t-1}, X_t, X_{t+1}, \dots) \\ & \approx p(Y|X_{t-\Delta t_p}, \dots, X_t, \dots, X_{t+\Delta t_n}) \\ & \approx p(Y|X_{t-\Delta t_p}) \cdots p(Y|X_t) \cdots p(Y|X_{t+\Delta t_n}). \end{aligned} \quad (3)$$

The  $Y$  that maximizes this probability provides the result of phoneme extraction at  $t$ . The MSP of phoneme  $Y$  could be decided by using the local maximum of the probability.

Here we define the *MSP likelihood* that the MSP of phoneme  $Y$  exists at  $t$ . For the sake of calculation convenience, the probability is expressed by logarithmic likelihood. The MSP likelihood which indicates how likely a phoneme  $Y$  exists at  $t$  is denoted by:

$$L(t, Y) = \frac{1}{\Delta t_p + \Delta t_n + 1} \sum_{i=-\Delta t_p}^{\Delta t_n} \log p(Y|X_{t+i}) \quad (4)$$

To estimate the probability distribution of  $p(Y|X_t)$  for the training data, the *true* points of MSP's are needed as a teacher signal. However, unlike phoneme labels, they are unknown for the initial dataset because they do not relate directly to the observation. We therefore apply the following iterative procedure for a large amount of speech data.

- (1) Set an initial MSP point of  $y^*$  for each phoneme  $Y$  in the training data based on the uttered text. If phoneme labels are given in advance, the initial MSP positions could be decided using the equally divided points for example. If there are no labels, the initial MSP's are set by positioning with same intervals for entire phoneme sequence. As mentioned later, the convergency of MSP position is warranted and this effects only to the speed of convergence.
- (2) Estimate distributions of feature  $X_{t \pm i}$  around  $y^*$ .
- (3) Calculate the MSP likelihood  $L(t, Y)$  for each phoneme.
- (4) Move  $y^*$  toward the local maximum within  $\pm 10$  ms (2 frames in our condition) from the original  $y^*$  of the likelihood.
- (5) Iterate steps (2)-(4) until all  $y^*$ 's converge.

During this process, long vowels, expressed by two same phonemes, could come together at the same point by the iteration. To avoid this, we give a heuristic condition that two MSP's have 20 ms or longer interval.

### 4. Consideration of Phonemic Environments

As mentioned in Section 2, the phonetic properties are frequently diverged by the contexts or the environments for specific phonemes. To realize higher accurate phoneme extraction, therefore, a context dependent model is often employed. When we consider more detailed environments, however, the number of fundamental models increases and larger amount of training data are necessary for the parameter estimation.

Here we apply the triphone structure as a context dependent MSP model, which contains one phoneme and its front and behind phonemes. We first calculate the probability distribution of the MSP for each combination of the triphone unit. Then several units which exhibit more similar properties are integrated successively using the MSP likelihood as an evaluating measure. The integration is executed according to the quantity of the training data. By iterating and integrating procedure which

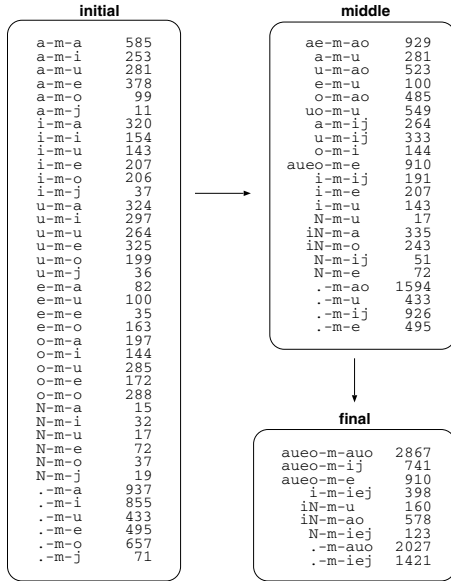


Figure 1: An example of integrating triphones (phoneme /m/).

maximizes the local likelihood, we could obtain an optimal set of MSP's. Thus the grouping of the MSP's is realized considering their phonemic environments.

The MSP triphone integration algorithm is as follows:

- (1) Using the MSP estimation algorithm (see Section 3), estimate the MSP's for all combinations of  $Y_j^{(n)} = \{Y_j^-, Y_j, Y_j^+\}$  in the training dataset, where  $Y_j$  is the target phoneme,  $Y_j^-$  is the front phoneme,  $Y_j^+$  is the behind phoneme and  $n$  is the number of combinations.
- (2) For all  $Y_j^{(n)}$ , integrate two sets  $\{Y_j^{(a)}, Y_j^{(b)}\}$  that have the most similar properties. The similarity is evaluated by looking for one  $Y_j^{(b)}$  which the difference of MSP likelihoods is minimum at the MSP. In this case, triphones which have different  $Y_j$  are not integrated.
- (3) Reestimate MSP's for the integrated set  $Y_j^{(a+b)}$ .
- (4)  $n = n - 1$ ; iterate from step (2) until the target model size.

Since the positions of two MSP's are usually different, the updated MSP is reestimated in step (3) severally.

Figure 1 shows an example of the integration process of triphone models for phoneme /m/. In this case, 2,131 triphones are integrated into 134 models. The figure shows the contents of both initial and final states along with the middle state (at 488 triphones).

## 5. Experiments and Discussion

In order to verify the proposed method, first, we investigate the change of MSP likelihoods by applying the above iterative procedure for a large amount of speech data. Next, to examine the effect of integrating triphone models, we experiment a continuous phoneme recognition based on One Pass DP in which the MSP likelihood is used as a similarity score.

### 5.1. Experimental Conditions

For the experiments, we employ multi-speaker (10 males: MAU, MHT, MMS, MMY, MNM, MSH, MTK, MTM, MTT and MXM) speech data (5,240 common words and 216 phonetic balanced words) in ATR Japanese Speech Database[4].

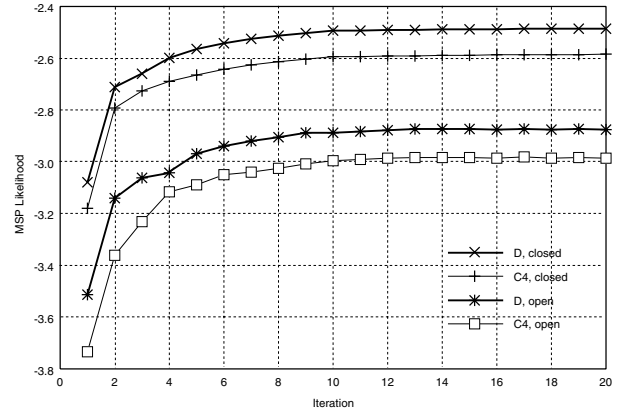


Figure 2: Convergence of MSP Likelihood (all phonemes)

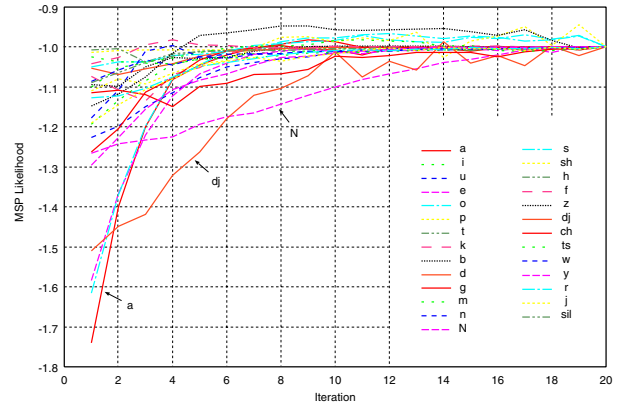


Figure 3: Convergence of MSP Likelihood (each phoneme; discrete distribution)

The data are digitized at 16 kHz, and LPC mel-cepstrum and its  $\Delta$ ,  $\Delta\Delta$  parameters are calculated every 5 ms, using a 24 ms Hamming window. The total dimension of the feature vector is 48.

All results shown in this paper are in speaker independent condition; training by 9 speakers and test by another speaker. We use five combinations, in which the test speakers are MAU, MHT, MMS, MMY and MNM, and the results shown later are by the averaged value. For the training (calculating the probability distribution of MSP's), 23,580 words (half of  $9 \times 5,240$  words) are used, and for the test, (i) the same 23,580 words (closed test), or (ii) other 23,580 words (open test) are used.

As the probability  $p(Y|X_t)$ , we compare (i) discrete distribution using VQ which codebook size is 2,528, and (ii) continuous distribution using diagonal Gaussian, 4 mixtures. In case of (i), we apply the hierarchical clustering algorithm [5]. In this paper, the detail explanation is omitted because it's not the essential point. The number of considering frames in equation (4) is decided as  $\Delta t_p = \Delta t_n = 3$  also given by a preliminary experiment which evaluates the mutual information between acoustic features and phoneme categories [5].

### 5.2. MSP Convergence

When the uttered text (sequential phoneme symbols) is given for the training data, convergence of the MSP's is warranted. The summation of the MSP likelihood for all phonemes in a training dataset takes ordinarily minus value, upper bounded,

and monotonically increasing. In the strict sense, this assumption could be proved when the analyzed period is rather short. Since we use 5 ms period in the experiments, the monotonic convergency is not always warranted. Therefore, we evaluate the convergency by the following experiments.

At each point of 20 times iterations, we examine the average of MSP likelihoods. Since the probability distribution could become quite dependent to the training dataset, the average of MSP likelihoods for the test dataset is also calculated. Note that the triphones and their integrating models are not used in this experiment. The results are shown in Figure 2 and 3.

Figure 2 shows the convergency of the MSP likelihood for all phonemes (the averaged values) using discrete distribution (D) and continuous distribution (C4) models. In the figure, “closed” means the result by the training dataset and “open” means for the test dataset. The first iteration is provided by the likelihood using the initial MSP set.

Figure 3 shows the convergency for each phoneme in case of discrete distribution (D) with closed dataset. Since the absolute value of the likelihood is different for each phoneme, it is normalized to -1.0 at 20 iterations.

The results show that the MSP likelihoods are increased and converged by the iteration for any probability distributions. Also for the dataset which is not used in the training, the likelihoods are increased and converged. From Figure 3, although the converging speed is different, it is confirmed that the likelihood is converged after about 15 iterations for most phonemes.

### 5.3. Phoneme Recognition Accuracy

Next, we apply the MSP likelihood to One Pass DP based continuous phoneme recognition, which is the simplest way to use the MSP likelihood as a resemblance measure in the DTW. Only the connectivity of two phonemes, which is a loose linguistic constraint, is considered as a linguistic knowledge. The recognition algorithm is as follows:

- (1) Symbols:
  - $L(t, Y)$  : MSP likelihood of phoneme  $Y$  at  $t$ .
  - $G(t, Y)$  : optimal cumulative likelihood until  $t$ .
  - $C(\tilde{Y}, Y)$  : boolean connectivity of  $\tilde{Y}$  and  $Y$ .
  - $B(t, Y)$  : array for backtrack.
  - $P(t, Y)$  : array for phoneme connectivity.
- (2) Initialize for all  $Y$ :
 
$$B(0, Y) = P(0, Y) = 0; \quad G(0, Y) = L(0, Y)$$
- (3) for  $t = 1, 2, \dots, T$  (input frames):
- (4) for all  $Y$ :
- (5) 
$$\hat{Y} = \underset{y}{\operatorname{argmax}} \{G(t-1, y) \cdot C(y, Y)\}$$

$$G(t, Y) = G(t-1, \hat{Y}) + L(t, Y)$$
- (6) 
$$B(t, Y) = (\hat{Y} == Y) ? B(t-1, Y) : t-1$$

$$P(t, Y) = (\hat{Y} == Y) ? P(t-1, Y) : \hat{Y}$$

? and : are like the conditional operators in C language.
- (7) Backtrack:
 
$$\hat{Y} = \underset{y}{\operatorname{argmax}} G(T, y), \quad t = T$$

while  $t > 0$   
 output  $\hat{Y}$  as a candidate phoneme  

$$Y = P(t, \hat{Y}), \quad t = B(t, \hat{Y}), \quad \hat{Y} = Y$$

To evaluate the recognition performance, three kinds of errors are considered; Substitution, Deletion, and Insertion versus Correct phonemes. Then we define RA (*recognition accuracy*) =  $(C - I)/(C + S + D)$  [6].

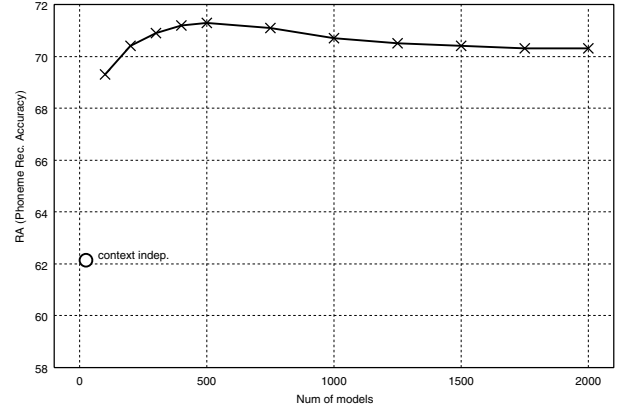


Figure 4: Phoneme recognition accuracy for the context independent and dependent models with several target model sizes.

Figure 4 summarizes the results. 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 MSP integration algorithm. For the experimental conditions in this paper, the best performance of phoneme extraction is obtained with 500 models.

## 6. Conclusion

In this paper, we have proposed a fictitious point which has most stable information to explain phoneme’s existence based on universal constraint of Japanese phonemes. The points could be estimated statistically by using an iterative procedure. We used the points to extract phonemic characteristics for the analyzed frames and also investigated the integration of triphone models to consider the context dependency of phonemes. As a method to express phoneme properties by a statistical basis, the proposed strategy is very stable and good to describe the phoneme’s characteristics even if there are rather strict constraints. Also the results are well contrasted with the former phonological studies.

The MSP and its likelihood could be useful not only for speech recognition and synthesis, but also to clarify the phoneme’s fundamental properties. As a future work, we would like to investigate the MSP with points of view of speech perception and prosody.

## 7. References

- [1] Clark, J. and Yallop, C., “An Introduction to Phonetics and Phonology,” *Blackwell Publishers*, Oxford, 1990.
- [2] Chomsky, N. and Halle, M., “The Sound Pattern of English,” *Haper and Row*, New York, 1968.
- [3] Okawa, S. and Shirai, K.: “Estimation of statistical phoneme center and its application to accurate phoneme modeling,” *Proc. EUROSPEECH*, 791–794, 1995.
- [4] Kuwabara, H. *et al.*, “Construction of a large-scale Japanese speech database and its management system,” *Proc. ICASSP*, 1989.
- [5] Okawa, S., Kobayashi, T. and Shirai, K., “Phoneme recognition in various styles of utterance based on mutual information criterion,” *Proc. ICSLP*, 1911–1914, 1994.
- [6] Lee, K. F., “Automatic Speech Recognition – Development of the SPHINX System,” *Kluwer Academic Publishers*, Boston, 1989.