TR-IT-0105

Non-uniform unit based HMMs for continuous speech recognition

松村 壮史
Takeshi, Matsumura

1995.3

A novel acoustic modeling algorithm that generates non-uniform unit HMMs to effectively cope with spectral variations in fluent speech is proposed. The algorithm is devised for the automatic iterative generation of long-span units for non-uniform modeling. This generation algorithm is based on an entropy reduction criterion using text data and a maximum likelihood criterion using speech data. The effectiveness of the non-uniform unit models is confirmed by comparing likelihood values between long-span unit HMMs and conventional phoneme-unit HMMs. Results of classification tests showed that the non-uniform unit HMMs provide more precise representation than do conventional phoneme-unit HMMs, and preliminary phrase recognition tests suggest that non-uniform unit HMMs achieve higher performance than phoneme-unit HMMs.

Contents

1 INTRODUCTION 3

2 NON-UNIFORM UNIT MODELING 3
2.1 Long-unit Candidate Selection 4
2.2 Selection of Non-uniform Unit HMMs 4

3 Database 5

4 Context-independent Non-uniform Unit HMM 5

5 Context-dependent Non-uniform Unit HMM 6

6 Conclusion 8

参考文献 9

1 INTRODUCTION

Our objective is to recognize continuous speech for speech translation, in which there is a high degree of co-articulation and many speech variations such as similation, elision and filled pauses.

We believe that there are two fundamental problems in continuous speech recognition. The first is the limitations of phoneme-sized models. An important technique for achieving higher recognition is to generate a precise acoustic model. The context-dependent phoneme hidden Markov model (HMM) is widely used as an acoustic model because phoneme variations are highly dependent on the phonetic context (Schwartz, 1985; Moor, 1993; Takami and Sagayama, 1992). However, if the model's length is fixed as a phoneme, the context-dependent phoneme HMM cannot sufficiently represent long-distance contextual influences. Accordingly, it is more appropriate to model these typical variations as long-span units than to represent them as a concatenation of uniform context-dependent phoneme models, as done in the conventional approach.

The other problem is the difficulty of target-set-independent modeling for high recognition performance; for example, a task independent model, a speaking style independent model, and so on. It is important to cover the major allophonic variations expected to be contained in the target speech by using a small quantity of training or adaptation speech data.

This work mainly concerns the first problem, but a solution is proposed for each problem. The first solution involves using a unit-sized free model, which is longer than a phoneme-sized model, to represent highly co-articulated speech. The model's unit size depends on the context. That is, if the acoustic co-articulation of a context is high, the context is modeled as a long-unit-sized model to represent the co-articulation.

The other proposed solution is a modeling that takes into account the linguistic and acoustic characteristics of the target speech. Aiming toward a target-speech-dependent model, these characteristics are incorporated in the acoustic model when the unit-sized free model is to be generated.

To realize these solutions, we propose an acoustic modeling algorithm which generates non-uniform unit (unit-sized free) HMMs. The non-uniform unit HMMs include long-span units and phoneme units to cope with spectral variations having longer periods than the phonemes.

2 NON-UNIFORM UNIT MODELING

The algorithm is based on an entropy reduction criterion using text data to select a long-unit as a candidate for the non-uniform unit HMM and a maximum likelihood criterion using retraining speech data to train each long-unit HMM and check whether the HMM is appropriate as a non-uniform unit HMM.

Both text and speech data are target data of the same type, from the viewpoint of task and speaking style. Accordingly, models generated by this algorithm with these data can represent the characteristics of the target data.

In automatic generation, the long-units must be properly chosen because it is futile to generate long-units not entirely contained in the target speech. Furthermore, since the amount of training data is restricted, the longer a unit is, the less training data there is. This sometimes reduces recognition performance. Therefore, the long-unit model must be trained robustly; the following strategy can be used to generate the non-uniform unit HMMs.

If there is sufficient data to train the long-unit HMM and as well as the necessary contextual variations to recognize the target speech, the long-unit HMM is generated as a non-uniform unit HMM. Otherwise, then the long-unit HMM need not be a non-uniform unit HMM. A block diagram of the algorithm is shown in Figure 1.

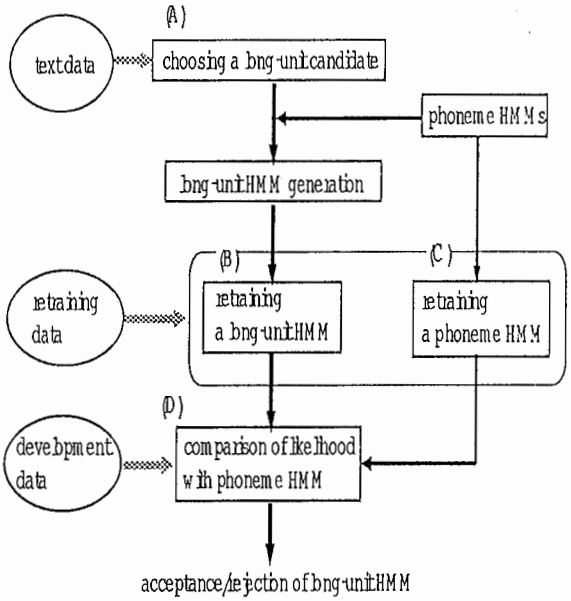

Figure 1: Block diagram of non-uniform unit generation

The non-uniform unit generation algorithm uses text data expressed in phoneme symbols to choose long-unit HMM candidates and three kinds of speech data: training speech data for initial phoneme modeling, retraining speech data for long-unit models and phoneme models and development speech data for selecting the long-unit HMM as a non-uniform unit HMM. Text data, training data, retraining data and development data all differ but concern the same task, that is, conference registration.

2.1 Long-unit Candidate Selection

A statistical approach to selecting the long-unit candidates is to choose the phoneme sequence that minimizes the entropy of the training text data (Tamoto, 1992). A heuristic approach is to choose the phoneme sequence that most frequently appears in the text data, thus reducing the entropy. In this research, we adopt the latter approach because it requires less computation.

Long-unit candidate selection procedure

The candidates for the long-unit HMM using text data were selected as follows ((A) in Figure 1). First, the frequency of all combinations of two neighboring phonemes in the text data were calculated. The phoneme (sequence) pair that had the highest frequency was selected as a long-unit candidate. If the candidate satisfied the acoustical conditions described in Section 2.2, this procedure was re-executed to get a new candidate on the condition that this newly selected phoneme sequence had to be regarded as a newly defined phoneme unit. If the conditions were not satisfied, the phoneme sequence was not selected as a candidate, and the phoneme (sequence) pair with the second highest frequency was selected instead.

2.2 Selection of Non-uniform Unit HMMs

Each long-unit HMM chosen from the text data was retrained and checked by using the following procedures to determine whether or not it was acceptable as a non-uniform unit HMM.

Retraining procedure

The initial long-unit model was obtained by concatenating the phoneme HMMs already given. The long-unit HMMs were retrained by using the Baum-Welch algorithm with retraining speech data appearing to have the same phonetic characteristics as the target speech data ((B) in Figure 1). The original phoneme HMMs were also retrained by using the same retraining speech ((C) in Figure 1).

Next, the long-unit HMM was checked by using the development speech data to verify whether it could be used as a non-uniform unit HMM as follows.

Verification procedure

First, the likelihood using the long-unit HMM was calculated for the development data. Another likelihood was then calculated using the concatenated phoneme HMMs.

These two likelihood values were then compared, and if the likelihood for the long-unit HMM was higher, the long-unit HMM was used as the non-uniform unit HMM. If

not, the long-unit HMM was rejected, and the phoneme models were used for recognition ((D) in Figure 1).

These three procedures were iteratively executed to get all of the non-uniform models.

3 Database

As mentioned before, we used one text data and three acoustic data to generate the non-uniform unit HMM. The text data consisted of 93,136 Japanese phrases represented in phoneme symbols. As for the acoustic data, the training data used to train the initial phoneme HMMs consisted of a labeled Japanese database of 2,620 common words. Second, labeled Japanese continuous speech database for a conference registration task was divided into independent sets: adaptation data, development data, and test data; these data sets consisted of 749, 354 and 276 phrases, respectively. We used two speakers' utterances (MHT and MAU).

4 Context-independent Non-uniform Unit HMM

To confirm the capacity of the long-unit generated by using the proposed algorithm, classification tests and tests using non-uniform unit HMMs were carried out.

In the retraining procedure, the training section was restricted to getting a higher performance with manually given phoneme boundaries (Ariki, 1994).

A context-independent phoneme HMM, constructed by three states and 3 diagonally Gaussian mixtures, was used as the initial HMM. The feature was a 34-dimensional vector consisting of 16 cepstral coefficients, 16 Δcepstral coefficients, logarithmic power and Δlogarithmic power. The analysis conditions are listed in Table 1.

Table 1: Analysis conditions

pre-emphasis	$1\text{-}0.98z^{-1}$
sampling frequency	12kHz
window length	20.0ms (Hamming window)
window shift	5ms
LPC analysis order	16
LPR cepstral order	16

Forty-eight (MHT) and 61 (MAU) long units were generated by repeating the algorithm 100 times. The averaged unit lengths of the long-unit HMMs were 3.0 (MHT) and 2.8 (MAU) phonemes.

The average sample numbers of retrained long-unit HMMs for the long-unit retraining procedure are listed in Table 2.

Table 2: Average retraining sample numbers (context-independent non-uniform unit HMMs)

speaker	accept	reject
MHT	55.4	20.8
MAU	74.5	20.3

"Accept" means the average retraining sample number of accepted long-unit HMMs, or in other words the likelihood of the long-unit HMM is higher than the concatenated phoneme HMMs'; "reject" means the average retraining sample number of rejected long-unit HMMs, or in other words the likelihood of the long-unit HMM is lower than the concatenated phoneme HMMs'. Table 2 shows that the long-unit correlates closely with the number of retraining samples.

Next, the long-unit classification test was carried out on the test data sections where long-units were applicable. The error rates are listed in Table 3.

Table 3: Error rates of classification test (context-independent non-uniform unit HMMs)

speaker	concatenate	non-uniform	multi
MHT	5.63%	6.54%	5.17%
MAU	5.51%	4.26%	3.38%

"Concatenate" means a concatenated phoneme model retrained using the same data that was used in retraining long-units; "non-uniform" means the non-uniform unit HMM. "Multi" means a multi-model containing both non-uniform unit HMMs and concatenated phoneme HMMs.

In classification the likelihood of each HMM is calculated. For example, the likelihood of all non-uniform unit HMMs "masu, de, kai, ..." and that of all concatenated phoneme HMMs "m-a-s-u, d-e, k-a-i, ..." were both calculated at the "masu" section of the test data. The HMM (either "masu" or "m-a-s-u") achieving the higher likelihood was recognized as correctly classified.

Table 3 shows that the error rate of "multi" was lower than those of "concatenate" and "non-uniform".

To evaluate the performance of the long-unit models, we checked the ratio of the number of correctly classified sections where the long-units achieved their highest likelihood to the number of correctly classified section where the phoneme HMMs did. The results in Table 4 show that the long-unit HMMs accounted for about 80% of all sections recognized correctly.

Table 4: Ratio of correctly classified sections (context-independent non-uniform unit HMMs)

speaker	accept	reject
MHT	77.4%	22.6%
MAU	79.5%	20.5%

This shows that long-unit HMMs expected to represent the long-span spectral characteristics can be generated by this algorithm.

Next, preliminary phrase recognition tests using HMM-LR (Singer, Takami and Matsunaga, 1994) were carried out. In this test we used long-unit HMMs that were applied to intra-word phoneme sequences. Accordingly, 24 (MHT) and 27 (MAU) long-unit HMMs were used as non-uniform unit HMMs.

The number of model parameters of the phoneme HMMs and the non-uniform unit HMMs (long-unit HMMs + phoneme HMMs) are listed in Table 5.

Table 5: The number of model parameters (context-independent non-uniform unit HMMs)

speaker	phoneme	non-uniform unit
MHT	390	1170
MAU	390	1260

The error rates are listed in Table 6.

Table 6: Error rates of phrase recognition test (context-independent non-uniform unit HMMs)

speaker	concatenate	non-uniform
MHT	8.7%	6.2%
MAU	8.7%	7.2%
average	8.7%	6.7%

Table 6 shows that the error rate of the non-uniform unit HMMs was lower than that of the concatenated phoneme HMMs. It also shows that the non-uniform unit HMMs achieved a 23% error reduction over the phoneme HMMs.

5 Context-dependent Non-uniform Unit HMM

We generated context dependent non-uniform unit HMMs. A 600-state 1-mixture HMnet

was used as the initial HMM. This HMnet is a context-dependent phoneme HMM generated by the Successive State Split algorithm (Takami and Sagayama, 1992). HMnet is a highly generalized form of the HMM and incorporates context-dependent variations of phones and state sharing among different allophones.

In the long-unit candidate selection procedure, the candidate phoneme sequence included the preceding/succeeding context sets having the same expressions as those in the HMnet.

In the retraining procedure, the initial long-unit model was obtained by concatenating the state sequences of the HMnet so as to take account of the preceding/succeeding phoneme contexts. For retraining, the Baum-Welch algorithm was applied. Because the amount of retraining data is known to be sparse in generating context-dependent non-uniform unit HMMs, the Vector Field Smoothing (VFS) algorithm (Ohkura, Sugiyama and Sagayama, 1992) was also used.

The VFS algorithm is a speaker adaptation algorithm for handling sparse adaptation data, and the algorithm estimates the vector field, by considering the correspondence between feature parameters before and after adaptation, with both interpolation and smoothing processes.

The number of generated long-unit HMMs is listed in Table 7.

Table 7: The number of generated context-dependent non-uniform unit HMMs

speaker	Baum-Welch	VFS
MHT	46	47
MAU	62	71

Figure 2 shows the average retraining sample data of the accepted or rejected long-unit HMMs under each condition.

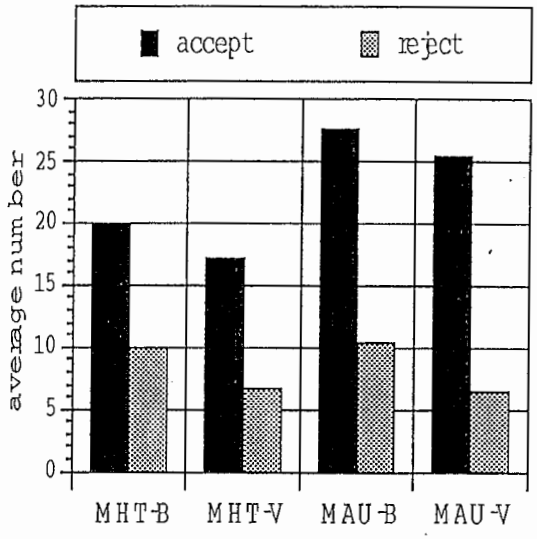

Figure 2: The average number of retraining samples (context-dependent non-uniform unit HMMs)

MHT and MAU denote the speaker indices. "-B" means the Baum-Welch algorithm and "-V" means VFS algorithm in the retraining procedure. In Figure 2 the accepted long-unit HMM has more retraining sample data than the rejected HMM.

Next, long-unit classification tests were carried out in the same way as in Section 4. Figure 3 shows the error rate for each condition.

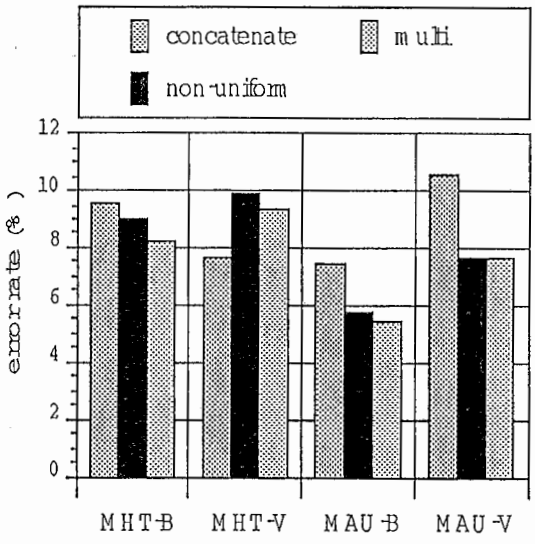

Figure 3: Error rates of classification tests (context-dependent non-uniform unit HMMs)

Except for MHT-V, the error rate in us-

ing a non-uniform unit HMM is lower than the case of using concatenated phoneme HMMs. We can see that the error rate using Baum-Welch is lower than that using VFS.

Figure 4 illustrates the ratio of the number of correctly classified sections where the long-units achieved a highest likelihood to the number of the correctly classified section where the phoneme HMMs did.

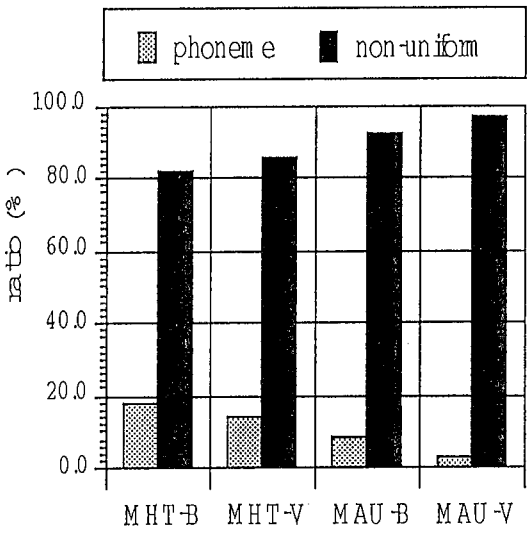

Figure 4: Ratio of correctly classified sections (context-dependent non-uniform unit HMMs)

As the figure shows, more non-uniform unit HMMs than concatenated phoneme HMMs were used in the correctly classified sections. This shows that non-uniform unit HMMs provide more precise modeling than phoneme HMMs.

Next, preliminary phrase recognition tests using HMM-LR were carried out with context-dependent non-uniform unit HMMs and an HMnet retrained by the VFS algorithm. Long-unit HMMs applied to intra-word phoneme sequences were used in the same way as in Section 4. Accordingly, 49 (MHT) and 96 (MAU) long-unit HMMs were used as non-unit HMMs.

The number of model parameters of the HMnet and non-uniform unit HMMs are listed in Table 8.

Table 8: The number of model parameters (context-dependent non-uniform unit HMMs)

speaker	phoneme	non-uniform unit
MHT	600	1112
MAU	600	1588

The error rate results are listed in Table 9.

Table 9: Error rates of phrase recognition test (context-dependent non-uniform unit HMMs)

speaker	concatenate	non-uniform
MHT	8.7%	7.3%
MAU	9.4%	7.3%
average	9.1%	7.3%

Table 9 shows that the error rate of non-uniform unit HMMs is lower than that of HMnet. The non-uniform unit HMMs achieved a 19.8% error reduction over the HMnet.

6 Conclusion

In this paper we proposed a non-uniform unit HMM that represents highly co-articulated speech and linguistic/acoustic characteristics of the target speech.

A non-uniform modeling algorithm that automatically generates long-unit models using text and speech data was introduced for model selection, taking into account the robustness of the unit and the amount of speech data.

In tests, long-unit models generated with the proposed algorithm showed higher potential than the conventional phoneme models. However, in preliminary phrase recognition tests the number of model parameters of the non-uniform unit HMM was larger than that of phoneme HMMs (Table 5 and Table 8). Accordingly, an investigation is needed on an evaluation method

able to cope with changes in the number of model parameters.

After comparing results of phrase recognition tests on context-independent and context-dependent non-uniform HMMs, the former was found to achieve higher recognition performance than the latter. There are various reasons for this. First in LR-parsing, the search space of context-dependent non-uniform unit HMMs grows larger than that of context-independent non-uniform HMMs. In these tests however, the same beam width was used. Another reason is the use of a different retraining algorithm in these tests. The Baum-Welch algorithm and the restricted training section method were used for context-independent non-uniform unit HMMs. On the other hand, the VFS algorithm was used to retrain the context-dependent non-uniform HMMs.

In the future, investigates are necessary on appropriate retraining algorithms for the non-uniform unit HMMs generation.

Finally, in this recognition system (HMM-LR), it was found that if the model's unit size is changed, we must manually rewrite grammars to adapt to the new unit size. Therefore, we must also investigate an algorithm for automatic grammar modification.

参考文献

[1] R. Schwartz et al., "Context-Dependent Modeling for Acoustic-Phonetic Recognition of Continuous Speech", Proc. of ICASSP-85, 1985.

[2] R. Moore, "Context Adaptive Phone (CAP) Modeling for Vocabulary-Independent Automatic Speech Recognition", Proc. of IEEE ASR Workshop, 1993.

[3] J. Takami and S. Sagayama, "A Successive State Splitting Algorithm for Efficient Allophone Modeling," Proc. of ICASSP92, pp. 573-576 (1992).

[4] K. Ohkura, M. Sugiyama and S. Sagayama, "Speaker Adaptation Based on Transfer Vector Field Smoothing with Continuous Mixture Density HMMs", Proc. of ICSLP-92, 1992.

[5] M. Tamoto, K. Itoh and H. Tanaka, "A Tree-based Stochastic Phoneme Sequence", IEICE Technical Report, Vol.92, No.23, March 1992 (In Japanese).

[6] Y. Ariki and K. Doi, "Phoneme Recognition Improvement by Restricting Training Section In Concatenated HMM Training" ICASSP-94, 1994.

[7] H. Singer, J. Takami and S. Matsunaga, "Non-Uniform Unit Parsing for SSS-LR Continuous Speech Recognition", ICASSP-94, 1994.

コンテキスト依存不定長モデル作成ツール

以下のプログラムはコンテキスト依存不定長モデルの作成のためのものである。これらは、鷹見 元研究員作成の SSS-ToolKit.ver2 を元に作成されており、そちらも参考にされたい。なお、それぞれのソースファイルは

/data/atrh32/researchers/matumura/SSS-ToolKit.ver2/Src.concatenate

にあります。不定長モデル作成のシェルは、その下の

Exe_sh.CDHMM/Make_new_CDHMM.normal.sh

自由発話データに対する不定長モデル作成のシェルは、

Exe_sh.Spontaneous/Exe/_sh.Spontaneous/Make_new_CDHMM.normal.sh

です。参考にして下さい。

1. 音素モデルの連結

テキストデータから得られた不定長単位候補の音素系列を、あらかじめ用意した音素モデルを連決することにより、不定長 HMM として作成するプログラムである。

使用方法 Exe.initial_HMnet [option]

オプションは以下のものが有効です。（不要なオプションもありますが、それらについては、ここで取り上げません。）

-il 連結に使う音素 HMnet の全リスト
-of 作成された不定長 HMM のファイル名
-nml 不定長単位候補のリストファイル
-mn 連結に使う音素 Hnet の混合数

・不定長単位候補のリストファイルのフォーマット

m 26 34 a 97 108 ma 26 108
ma 26 108 s 388 149 mas 26 149
...

アルファベットは、モデルシンボルを表し、数字は連結に使う HMnet における各モデルシンボルのパスの最初と最後の状態番号である。例えば一列目は
状態番号 26 がら始まり 34 で終わるパスを持つモデル「m」と、
状態番号 97 がら始まり 108 で終わるパスを持つモデル「a」を連結し、
状態番号 26 がら始まり 108 で終わるパスを持つモデル「ma」

が不定長モデル単位候補である、という意味である。

2. 不定長モデルの再学習

1. で作成した不定長 HMM を VFS algorithm を用いて再学習するプログラムである。SSS-Toolkit.ver2 の Exe.adapt_HMnet と同じである。

使用方法 Exe.adapt_HMnet [option] < speech sample

音声サンプルのフォーマット、オプションおよびその使い方は SSS-Toolkit.ver2 に準ずる。（ユーザーズマニュアルを参照）

3 . 不定長モデルの尤度計算

さまざまな長さを持つ不定長モデルの、与えられた音声サンプルデータに対する尤度を求めるプログラムである。不定長モデル作成アルゴリズムでは、一つの不定長モデルの正解カテゴリのデータに対する尤度を求めるために用いた。

使用方法 Exe.recognize_new_model [option] < speech sample > likelihood

オプションは以下のものが有効です。

-if 不定長モデル
-nl 正解カテゴリのリスト (SSS-Toolkit.ver2 の Exe.recognize_HMnet のオプション -an に相当)
-fn モデルの要素数
-mf 中心要素番号
-ob 正解カテゴリに対するモデル連結のリスト
-re ダミー

・正解カテゴリのリストの例

1 mas
1 masu
...

左から、右側のカテゴリの個数、正解カテゴリ。

・正解カテゴリに対するモデル連結のリストの例

2 ma s
1 masu
...

左から、連結するモデル数、連結するモデルの種類

・プログラムの出力の例

Total_length 1505
answer: masu
masu 3.321655e+03
...
answer: masu
masu 3.266709e+03
...

上から、音声サンプルの全フレーム長、正解カテゴリ、正解カテゴリに対する尤度。