

An Optimized Multi-Duration HMM for Spontaneous Speech Recognition

Yuichi Ohkawa¹, Akihiro Yoshida², Motoyuki Suzuki³, Akinori Ito³, Shozo Makino³

¹Graduate school of Educational Informatics

²Graduate school of Information Sciences

³Graduate school of Engineering

Tohoku University, Japan

{kuri, akki, moto, aito, makino}@makino.ecei.tohoku.ac.jp

Abstract

In spontaneous speech, various speech style and speed changes can be observed, which are known to degrade speech recognition accuracy.

In this paper, we describe an optimized multi-duration HMM (OMD). An OMD is a kind of multi-path HMM with at most two parallel paths. Each path is trained using speech samples with short or long phoneme duration. The thresholds to divide samples of phonemes are determined through phoneme recognition experiment. Not only the thresholds but also topologies of HMM are determined using the recognition result.

Next, we parallelize OMD model with ordinary HMM trained by spontaneous speech and HMM trained by read speech in parallel. Using this ‘all-parallel’ model, 19.3% reduction of word error rate was obtained compared with the ordinary HMM trained with spontaneous speech.

1. Introduction

In spontaneous speech, many phonemes have shorter duration than that in read speech. These phonemes often cause misrecognitions on speech recognition result. An HMM has a capability to accept phonemes uttered in different length. However, the phonemes with short duration tend to have different distribution because of the two reasons. One is the influence of coarticulation. A phoneme with short duration is affected by its context more largely than a phoneme with longer duration. The other reason arises from the speech style. In spontaneous speech there are many vague pronunciations, in which many short phonemes are observed. Therefore it is difficult to model them using common models to phonemes with long duration. Okuda *et al.* [1] proposed multi-path HMM which is of two HMMs. One of these HMM is trained by phoneme samples with short duration and the others is trained by entire samples. However, their method to group phonemes into long and short duration samples is not guaranteed to be optimal because phoneme samples are grouped according to one threshold. Another approach treats with fast speaking rate through acoustic analysis. Nanjo *et al.* [2] proposed a method to change model topology and frame rate in order to avoid mismatch between HMM and short duration phoneme sample. In addition to these approach, Lee *et al.* [3] proposed multi-path HMM which exploits additional path for samples with low-likelihood. They first observed likelihood of each training sample using ordinary left-to-right HMM, and added another paths to express low-likelihood samples.

In this paper, we propose two new modeling scheme to create multi-path HMM. One is called multi-duration (MD) model, that is HMM with two parallel paths. One path is trained from

samples with long duration and the other from samples with short durations. Threshold to divide long and short duration is determined for each phoneme according to average phoneme recognition rate. The other model is called optimized-multi-duration (OMD) model. OMD model is basically two-path model, which is chosen among two-path models trained with various duration threshold through recognition experiment.

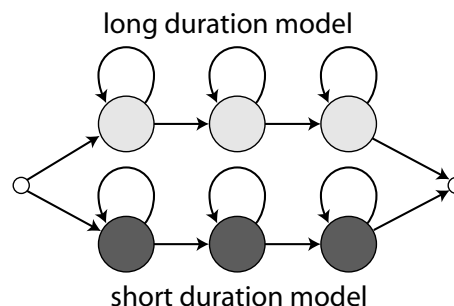


Figure 1: *Multi-duration model*

On the other hand, there may be a speaker whose speaking style is close to read speech. For such speaker, an acoustic model trained from read speech might be better. Therefore, we make a new model by connecting OMD and ordinary models in parallel. This model is evaluated through recognition experiments of spontaneous speech.

2. A Multi-Duration HMM

Figure 1 shows a simple multi-duration HMM. The model is of two parallel paths, one is trained from short duration speech samples and another is trained from long duration samples.

This model is similar to Okuda’s model[1]. The difference is how to split short and long phonemes. Okuda uses single threshold to divide all phonemes into both class, while we attempt to determine thresholds for each phoneme. The way to determine a threshold is as follows. First, single-path HMM is trained from spontaneous speech, and phoneme recognition experiment is carried out against development data. Second, one kind of phonemes are categorized into several duration class. Then phoneme recognition rates of each class are compared to average recognition rate. Finally, phoneme duration that gives average phoneme recognition rate is used as the threshold.

2.1. Recognition Systems and Speech Samples

In order to acquire the distribution of phoneme recognition correct per duration of phonemes, recognition experiment was carried out upon spontaneous speech. Julius speech recognizer[4] is used for the recognition. Syllable constraints are used as a language model. Acoustic analysis condition is shown in Table 1.

Table 1: Acoustic analysis Conditions for recognition experiments

Sampling freq.	16kHz
Feature vector	12 MFCC + 12 Δ MFCC + Δ POW
Frame length	25msec
Frame shift	10msec
Acoustic model	PTM (64 mix)

We prepared four kinds of data for training, decision of the threshold, development and evaluation. The training data are 8100 sentences (2030 speakers) of dialogue in ATR spontaneous speech corpus. For decision of the duration threshold, other 6300 sentences (1740 speakers) of dialog in ATR spontaneous speech corpus are used. In addition, we also prepare about 1200 sentences (320 speakers) for evaluation and about 4900 sentences for development. The utterances for evaluation speech overlaps with utterances for the threshold decision.

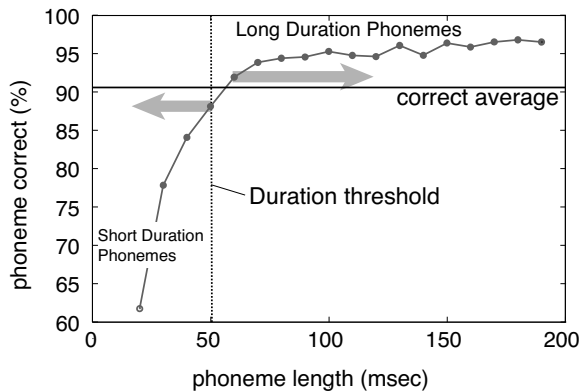
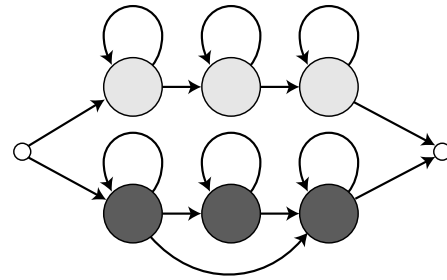


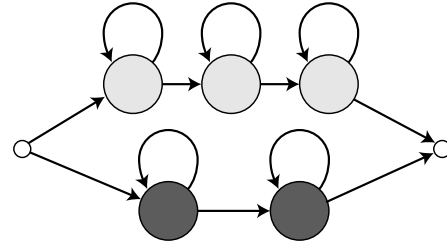
Figure 2: Example of decision of the duration threshold. (/a/)

2.2. Determination of duration thresholds of the Multi-Duration HMM

Figure 2 shows the example of decision whether an utterance is used for training of short or long duration model of multi-duration HMM of phoneme /a/. In this figure, the horizontal axis stands for duration class of samples and the vertical axis shows phoneme correct (%) of the samples in that duration class. The ‘average’ line shows phoneme correct of all /a/ samples. The duration threshold is determined as the maximum duration time when phoneme correct doesn’t exceed the average. Utterances longer than the threshold are used for training of long duration model and other utterances are used for training of short duration model. However, the consonants which have contracted sound /y/ part (e.g. “dy”, “gy”) and the long vowels are trained as normal single-path HMMs due to shortage of samples.



A. 3 states with a skip path.



B. 2 states.

Figure 3: Examples of topology of short duration path of MD HMM.

Table 2 shows the duration threshold of each phoneme decided by this method. The threshold differs according to kind of phoneme. For example, fricative sounds tend to have longer threshold.

Table 2: duration threshold of each phoneme decided by phoneme correct distribution

duration threshold	phoneme
20msec	u, g, r
30msec	o, w, y, p, t, b, d
40msec	i, e, k, n, f, z, q
50msec	a, m, N, h, ts
60msec	j
70msec	ch
80msec	sh
90msec	s

2.3. Model Topology of the Short Duration Path

Because a spontaneous speech has higher speaking-rate than read speech, phoneme duration often become short. Simple 3-state left-to-right HMM need at least 3 frames to accept utterance, but phonemes in this kind of speech sometimes become shorter than 3 frames. To treat this kind of short phoneme, we used 3 kinds of topology as short duration path of the multi-duration HMM. They are T-1) ordinary 3-state 3-loop, T-2) 3-state 3-loop with skip path (Figure 3-A.) and T-3) 2-state 2-loop (Figure 3-B.).

2.4. Evaluation through LVCSR

We evaluated performance of the multi-duration model through LVCSR. The proposed models were compared to two conventional models. One was 3 topology of the multi-path HMM

(MP), in which the short duration path was trained by shorter utterances than 30msec and the long duration path was trained by longer speech than common threshold. The other model was single-path HMM trained by spontaneous speech (SP).

Table 3 shows the word accuracy of the recognition result upon 1200 sentences evaluation set. 79.4% of word accu-

Table 3: *Word accuracy of LVCSR with each acoustic model (%)*

MD model			MP model			SP model
T-1	T-2	T-3	T-1	T-2	T-3	
78.7	79.4	79.2	79.2	79.4	78.9	78.7

racy was obtained by recognition with both the multi-duration HMM and the multi-path HMM. This performance is 0.7% of improvement against normal spontaneous HMM. This result showed that the proposed model outperformed ordinary single-path model, but its performance was not better than multi-path model with fixed threshold, which meant the threshold determination of MD model was not optimum.

3. An Optimized Multi-Duration HMM

The results of section 2.4 suggest that the threshold of phoneme duration decided by average phoneme correct recognition is not optimum. Moreover, optimum topology for each phoneme could vary from phoneme to phoneme. In this section, we propose a new scheme to choose optimum duration threshold and topology phoneme by phoneme through phoneme recognition experiment. First, multi-path models trained using various thresholds and topologies are prepared. Then phoneme recognition experiments are carried out using each model. Next, for a certain phoneme, the model that gives the highest accuracy for that phoneme is chosen. Finally, a set of HMM is created by gathering the chosen HMMs for each phoneme. We call this set of HMM ‘optimized multi-duration HMM(OMD-HMM)’.

3.1. Determination of Optimum Duration Threshold

As candidates of OMD model, we prepared the following 22 kinds of acoustic model, which have various threshold.

- 3 multi-duration HMM with three kinds of topology described in the previous section.
- 3 multi-duration HMM with 3-states topology, the threshold increased by 10msec, 20msec and 30msec evenly.
- 3×5 multi-path HMM with three kinds of topology, the common threshold is 30msec, 40msec, 50msec, 60msec and 70msec.
- The single-path HMM

We chose the optimum acoustic model for each phoneme from 22 kinds of model according to results of phoneme recognition of the development speech. The development speech is about 4900 sentences, which is not used for decision of threshold, training and evaluation.

3.2. Evaluation of the Optimized Multi-Duration HMM

Table 4 shows performances of the optimized multi-duration HMM (OMD) evaluated by LVCSR. The best accuracy is shown in the column of other methods.

Table 4: *Word accuracy of continuous recognition using each acoustic model (%)*

acoustic model	OMD	MP	MD	SP
word accuracy	80.8	79.4	79.4	78.7

The proposed model gave 80.8%, which was 2.1 point better than the single-path model trained by spontaneous speech, and 1.4 point better than the multi-path HMM.

4. Utilization of the read-speech model

4.1. Influence of disfluencies

Spontaneous utterances often have interjections and other disfluencies which don’t appear in read speech. As distribution of phoneme duration of utterances with some kind of disfluencies is more diverse than that of phonemes in read speech, the existence of interjection might show that the utterance is fluent or not. If an utterance is fluent, an acoustic model trained by read speech might work well.

To confirm this hypothesis, we carried out recognition experiments against 400 utterances with interjections and 800 utterances without interjections drawn from the evaluation set. The result of the experiment are shown in Table 5. In this experiment, read speech model (RD) is single-path HMM with 3 states, trained by about 9800 sentence (2030 speakers) of phonemically balanced sentences in ATR spontaneous speech corpora.

Table 5: *Word accuracy with each acoustic model (%)*

utterance	OMD	MP	MD	SP	RD
interjection	75.7	74.4	74.8	73.0	69.9
no interjection	86.3	84.9	84.6	84.8	84.5

The optimized multi-duration HMM gave best performance, whether utterances have interjections or not. The performance of single-path HMM trained by read speech was as high as other models trained by spontaneous speech.

4.2. The Speaker-Dependency of the Optimum Model

Next, speaker-by-speaker performance was observed to investigate if the optimum model differs from speaker to speaker. Table 6 shows the rate of speaker whose utterances was recognized better using single-path model trained by spontaneous speech (SP) or read speech (RD). The result shows that there are about quarter speakers to whom the SP or RD model gives higher performance than the proposed model.

Table 6: *Rates of speaker whose utterances was recognized better with SP or RD model than with OMD model.*

	rate
SP>OMD	25.6%
RD>OMD	27.4 %

5. A Poly Optimized Multi-Duration HMM

The results of the previous section show that there are many speakers whose speech style is close to read speech. Therefore, we propose a poly optimized multi-duration HMM, which combines the optimized multi-duration HMM, the single-path model and the read speech model altogether. An example of a poly optimized multi-duration HMM is shown in Figure 4.

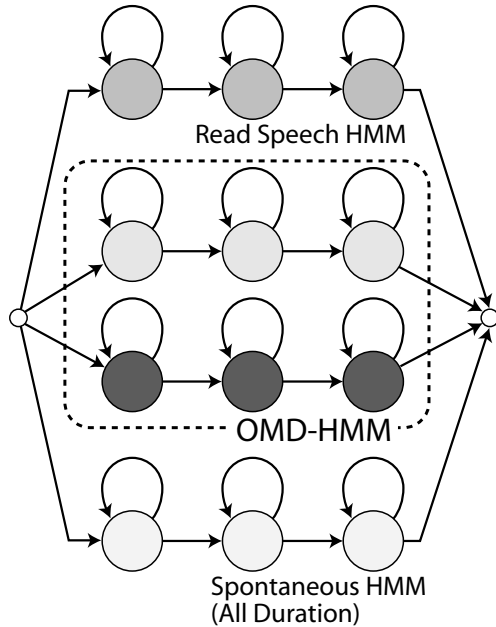


Figure 4: An example of Poly Optimized Multi-Duration HMM.

Table 7 shows the results of continuous recognition of about 1200 sentences of evaluation speech. Where, the read speech model (RD) is single-path HMM with 3 states. The optimized multi-duration HMM (OMD) has 2 paths. The multi-path HMMs combined with OMD and RD (or SP) models have 3 paths. Proposed poly optimized multi-duration HMM (POMD) combined with OMD, RD and SP models has 4 paths.

Table 7: word accuracy of continuous recognition with each model(%).

acoustic model	Word accuracy
RD	77.0
SP	78.7
OMD	80.8
OMD + SP	81.9
OMD + RD	82.6
POMD	82.8

82.8% word accuracy is obtained by LVCSR of 1200 spontaneous utterances of evaluation speech with proposed poly optimized multi-duration HMM. This performance is 4.1 point improvement against the single-path spontaneous HMM.

5.1. Comparison with Same Scale as Single-Path HMM

The parameters of each parallelized model is larger than that of normal spontaneous model. For example, the POMD-HMM have 4 times as large as 1-path HMM. Therefore the comparison

between parallelized OMD-HMM and simple SP model may be unfair because larger models can represent acoustic events in detail. So we prepared acoustic models with 192 and 256 mixtures, which are 3 and 4 times as large as the SP models respectively. These models are trained by both spontaneous and read speech. Table 8 shows performances of parallelized OMD-HMM and larger mixture model of single-path SP+RD HMM evaluated by LVCSR.

Table 8: word accuracy of continuous recognition with each model(%).

acoustic model	Word accuracy
1-path 192-mixture	81.3
3-path 64-mixture (OMD+RD)	82.6
1-path 256-mixture	82.3
4-path 64-mixture (POMD)	82.8

81.3% of word accuracy was obtained by recognition with 192-mixture PTM model trained by read and spontaneous speech. This model was the same scale as parallelized OMD+RD HMM. This performance is 1.3 points worse than the result with OMD+RD HMM. Then, 82.3% of word accuracy was obtained by recognition with the 256 mixtures model, which is same scale as POMD-HMM. This result is 0.5 point worse than performance of recognition with POMD-HMM. These results show that acoustical features are represented well by parallelized OMD-HMMs compared with single-path HMM of the same scale.

6. Conclusion

In this paper, we proposed an optimized multi-duration HMM, which is a multi-path HMM of which each path are trained long duration and short duration samples selected by the optimum threshold. An LVCSR of spontaneous speech with proposed model gave 2.1 point improvement of word accuracy.

We also proposed poly optimized multi-duration HMM, which is the combination of the optimized multi-duration HMM, the single-path model trained by all duration samples and the read speech model. 19.3% reduction of word error rate of continuous recognition is given by the poly optimized multi-duration HMM against conventional single-path spontaneous model. These parallelized optimized multi-duration HMM also gave better performance than single-path HMM, which is as large as the proposed model.

7. References

- [1] K.Okuda, T.Kawahara, and S.Nakamura, "Speaking rate compensation based on likelihood criterion in acoustic model training and decoding", Proc. ICSLP, pp.2589–2592, 2002.
- [2] H.Nanjo and T.Kawahara, "Speaking-rate dependent decoding and adaptation for spontaneous lecture speech recognition", Proc. IEEE-ICASSP, pp.725–728, 2002.
- [3] A. Lee, Y. Mera, K. Shikano, H. Saruwatari, "Selective multi-path acoustic model based on database likelihoods", Proc. ICSLP, FrB59p.10, 2002.
- [4] T. Kawahara, *et al.*, "Free Software Toolkit for Japanese large vocabulary continuous speech recognition", Proc. ICSLP, pp.476–479, 2000.