# STATISTICAL LANGUAGE MODELING WITH PROSODIC BOUNDARIES AND ITS USE FOR CONTINUOUS SPEECH RECOGNITION

*Keikichi Hirose†, Nobuaki Minematsu††, and Makoto Terao†††*

†Dept. of Frontier Informatics, School of Frontier Sciences, University of Tokyo
††Dept. of Inf. and Commu. Engineering, School of Inf. Science and Tech., University of Tokyo
†††Dept. of Inf. and Commu. Engineering, School of Engineering, University of Tokyo
Bunkyo-ku, Tokyo, 113-0033, Japan
{hirose, mine, terao}@gavo.t.u-tokyo.ac.jp

## ABSTRACT

A new statistical language modeling was proposed where word n-gram was counted separately for the cases crossing and not crossing accent phrase boundaries. Since such counting requires a large speech corpus, which hardly can be prepared, part-of-speech (POS) n-gram was first counted for a small-sized speech corpus for the two cases instead, and then the result is applied to word n-gram counts of a large text corpus to divide them accordingly. Thus, the two types of word n-gram model can be obtained. Using ATR continuous speech corpus by two speakers, perplexity reduction from the baseline model to the proposed model was calculated for the word bi-gram. When accent phrase boundary information of the speech corpus was used, the reduction reached 11%, and when boundaries were extracted using our formerly developed method based on mora-$F_0$ transition modeling, it still exceeded 8%. The reduction around 5% was still observed for sentences not included for the calculation of POS bi-gram and using boundaries automatically extracted from another speaker's speech. The obtained bi-gram was applied to continuous speech recognition, resulted in a two-percentage improvement of word accuracy from when the baseline model was used.

## 1. INTRODUCTION

In view of the importance of prosodic features in human speech perception process, a rather large number of research works have already been devoted to detect prosodic events and utilize them to facilitate speech recognition process. However, their results were little incorporated in speech recognition systems. Most successful case will be that of Verbmobil, but prosodic features were still utilized in rather limited ways [1]. The probabilistic factor in the human realization of the prosodic events may make it difficult to include them in the speech recognition process. However, we should also note that the positioning of prosodic boundaries is not a random process, and humans put boundaries only on possible locations, which mostly correspond to some linguistic boundaries.

As a method to utilize prosodic boundary information in speech recognition, we have developed a scheme to count prosodic boundary information in the n-gram based statistical language modeling, and realized an increased expressiveness [2]. The idea of incorporating prosodic boundary information into language modeling is motivated from the fact that the current statistical language modeling is only for written texts and prosodic features are tightly related to the structure of speaking. As outputs of human process of sound production, spoken sentences cannot be fully represented only by written language grammars. This consideration led us to an idea of separately modeling the word transitions for the two cases: one crossing and the other not crossing accent phrase boundaries. Here, "accent phrase" is a basic prosodic unit defined as a word or a word chunk corresponding to an accent component, which is also called as "prosodic word." The major difficulty along this line will be the collection of enough training corpora with prosodic information. In order to solve this problem, we first calculated POS bi-ram counts for a small speech corpus to find out the ratio of the two cases, and then applied the result to separate the word bi-gram counts of the text corpus into the two cases. Thus we can obtain two types of word bi-gram models. In the current paper, the detail of the proposed method is first explained and then its validity is shown as a reduction in perplexity from the baseline model (not counting prosodic boundaries). Speech recognition experiments are also conducted using the proposed model in several training situations and the results are compared with those obtained using the baseline model. Although the method can be applied for tri- or larger grams, in the current paper, it is restricted to bi-grams taking the size of the corpus into account.

## 2. MODELING SCHEME

### 2.1. Outlines

As shown in Figure 1, the proposed modeling is based on separately counting inter-accent-phrase word transitions and intra - accent-phrase word transitions. Two types of n-gram language model (henceforth, LM) are selected and used according to the existence / absence of accent phrase (henceforth, AP) boundaries during the speech recognition process. In Japanese, an AP mostly coincide with a "bunsetsu," which is defined as a basic unit of grammar and pronunciation, and consists of a content word (or content words) followed or not followed by a function word (or function words). Japanese speakers sometimes omit particles in a "bunsetsu," and in that occasion mostly insert an AP boundary before the next "bunsetsu." Therefore, AP boundaries usually occur before content words. Tables 1 and 2 show how POS transitions differ when they are crossing and when not crossing AP boundaries. The ATR 503 sentence speech corpus and AP boundary labels for speaker MYI's utterances were used to obtain the tables [3]. The result clearly indicates the differences in POS transitions

and thus implies the differences in word transitions according to the existence/absence of AP boundaries.

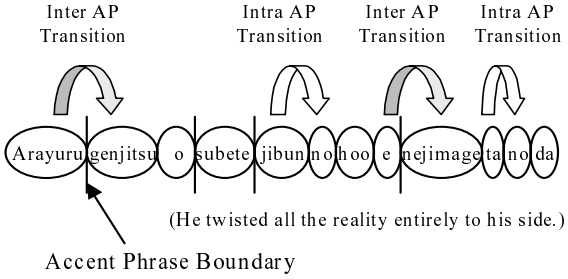


**Fig. 1**. Two types of word transitions. Transitions across AP boundaries and those not across are modeled separately.

**Table 1**. Probabilities (in %) of intra-AP POS bi-gram transitions.

| | | Transition to: | | | |
|---|---|---|---|---|---|
| | | Noun | Verb | Particle | Adverb |
| Transition from: | Noun | 8.9 | 5.2 | 67.5 | 0.1 |
| | Verb | 6.2 | 12.7 | 43.8 | 0.0 |
| | Particle | 6.8 | 47.9 | 36.9 | 0.3 |
| | Adverb | 2.5 | 15.0 | 60.0 | 0.0 |

**Table 2**. Probabilities (in %) of trans-AP-boundary POS bi-gram transitions.

| | | Transition to: | | | |
|---|---|---|---|---|---|
| | | Noun | Verb | Particle | Adverb |
| Transition from: | Noun | 71.1 | 13.4 | 1.4 | 2.8 |
| | Verb | 85.6 | 5.7 | 1.1 | 4.0 |
| | Particle | 51.1 | 34.3 | 0.2 | 6.1 |
| | Adverb | 59.6 | 28.8 | 0.0 | 1.4 |

### 2.2. Problem and solution

A large-sized text corpus, such as a newspaper corpus for one or more years, is required to train an LM. When training the two types of LM crossing and not crossing AP boundaries, we need a huge speech corpus with AP boundary information, which is not practical to prepare. As pointed out in Tables 1 and 2, differences in word transitions according to the existence/absence of AP boundaries can be well represented as POS transitions. Therefore, instead of directly constructing the two types of model, we first counted POS transitions for the two cases for a small speech corpus, and then divided word n-gram counts of the text corpus used for the training of the baseline LM accordingly. Figure 2 schematically illustrates this procedure to construct two types of LM. For instance, if 90 % of the noun to particle transitions occur inside an AP and the rest occur trans-AP boundaries, and if the bi-gram counts for the sequence "watashi (I) + ga (am)" are 1000, they are divided into 900 and 100 for the cases not crossing and crossing AP boundaries, respectively. As for the POS categories, 26 categories, such as nouns, pronouns, verbs, adjectives, adverbs, case particles, conjunctions, symbols, punctuation marks, and so on, are selected. From now on, LM before separation shall be called the baseline LM, and those separated using AP boundary information shall be called the proposed LM.

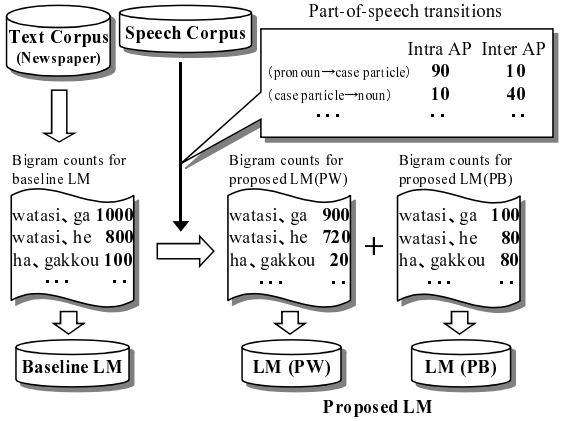


**Fig. 2**. Method of constructing two types of bi-gram LM by dividing bi-gram counts of the baseline LM. In the figure, PB and PW respectively indicate models crossing boundaries and those not crossing.

## 3. PERPLEXITIES

### 3.1. Outlines

The baseline bi-gram LM was trained for Mainichi Newspaper Corpus '97. The vocabulary size was 20 k words. Chasen version 2.02 [4] was used for the morpheme analysis and the Good Turing discounting was adopted for the bi-gram calculation.

As for the speech corpus for the POS bi-gram calculation, utterances by speakers MYI and MHT included in ATR continuous speech corpus of 503 sentences were selected [3]. The perplexity of the proposed model was calculated also for the ATR continuous speech corpus by changing two types of LM according to the existence/absence of the AP boundaries. Text closed and open experiments and speaker closed and open experiments were conducted.

### 3.2. Using AP boundaries labeled in the corpus

Experiments were conducted for speaker MYI's utterances using AP boundary labels attached to the speech corpus. The boundary information was used in the training of proposed model and also for the perplexity calculation. Table 3 shows the result when all of 503 sentences were used both for training and perplexity calculation. The total perplexity reduction from the baseline LM to the proposed LM was reached 11.0 %, indicating the validity of the proposed method. When the perplexity reductions were viewed separately for the cases of intra AP and trans AP boundaries, they were about 9 % and 15 %, respectively. The proposed method is valid for the both cases, but the effect is larger for the bi-gram across AP boundaries.

Table 4 shows the result of text-open experiments, where the corpus was divided into 453 and 50 sentences, and used for train-

ing and perplexity calculation, respectively. The cross validation scheme (ten combinations of training and evaluation data) was adopted to obtain a reliable result. Close to 9 % perplexity reduction was still observed on average.

**Table 3**. Perplexities for baseline and proposed LM's for speaker MYI speech when the AP boundary labels of the corpus are used. Trained and evaluated for all the 503 sentences.

| | All | Intra AP | Inter AP |
|---|---|---|---|
| Baseline model | 117.0 | 25.56 | 2664 |
| Proposed model | 104.1 | 23.32 | 2253 |
| Reduction rate | 11.0% | 8.76% | 15.4% |

**Table 4**. Perplexities for baseline and proposed LM's for speaker MYI speech when the AP boundary labels of the corpus are used. Trained for 453 sentences and evaluated for the rest 50 sentences. The cross validation scheme is used.

| | All | Intra AP | Inter AP |
|---|---|---|---|
| Baseline model | 117.4 | 25.66 | 2752 |
| Proposed model | 107.1 | 23.93 | 2408 |
| Reduction rate | 8.77% | 6.74% | 12.5% |

### 3.3. Using automatically detected AP boundaries

Experiments were further conducted in the more realistic situation, where the AP boundaries were detected automatically. The detection was done by a method formerly developed by one of the authors. It is based on modeling an AP $F_0$ contour as an HMM of mora unit $F_0$ contours [5]. For the 503-sentence corpus by speaker MYI, the detection rate and insertion error rate were 57 % and 24 %, respectively. Here, a morpheme boundary obtained by Chasen is assumed to be an AP boundary when one of the detected AP boundaries drops in the +/- 40 ms period of the morpheme boundary in question.

Table 5 shows perplexities when all the 503 sentences are used for the training and perplexity calculation, while Table 6 shows perplexities averaged over 10 sets of 50 sentences when the remaining 453 sentences are used for training in each set. In short, Tables 5 and 6 show the results corresponding to the cases of Tables 3 and 4, respectively. Although the rates of perplexity reduction from the baseline LM to the proposed LM came smaller, they still indicated the validity of the proposed method.

All the above results are obtained when training and perplexity calculation are done for the same speaker's utterances. One possible drawback of the proposed method is the speaker dependency; AP boundary positions may differ between speakers. In order to check this point, similar experiments were conducted using utterances by both of speakers MYI and MHT. Tables 7 and 8 show the results when the training was done for speaker MYI's utterances and perplexity calculation was done for speaker MHT's utterances. Perplexity reduction rates are still around 6 %. It is interesting that the result in Table 8 is even better than that in Table 6.

Table 9 shows the result when the AP boundary labels of the corpus were used for the training instead. The perplexity increased a lot for the proposed LM. This result indicates that, in both phases of training and recognition, we should use the AP's detected in the same criterion even if they contain errors.

**Table 5**. Perplexities for baseline and proposed LM's for speaker MYI speech when the automatically detected AP boundaries are used. Trained and evaluated for all the 503 sentences.

| | all | Intra AP | Inter AP |
|---|---|---|---|
| Baseline model | 117.0 | 57.13 | 1344 |
| Proposed model | 107.4 | 53.75 | 1133 |
| Reduction rate | 8.24% | 5.92% | 15.7% |

**Table 6**. Perplexities for baseline and proposed LM's for speaker MYI speech when the automatically detected AP boundaries are used. Trained for 453 sentences and evaluated for the rest 50 sentences (cross validation).

| | All | Intra AP | Inter AP |
|---|---|---|---|
| Baseline model | 117.4 | 57.32 | 1436 |
| Proposed model | 111.7 | 55.12 | 1331 |
| Reduction rate | 4.84% | 3.84% | 7.26% |

**Table 7**. Perplexities for baseline and proposed LM's when the automatically detected AP boundaries are used. Trained for 503 sentences by speaker MYI, and evaluated for 503 sentences by speaker MHT.

| | all | Intra AP | Inter AP |
|---|---|---|---|
| Baseline model | 117.0 | 46.40 | 1932 |
| Proposed model | 109.0 | 44.04 | 1701 |
| Reduction rate | 6.84% | 5.09% | 12.0% |

**Table 8**. Perplexities for baseline and proposed LM's when the automatically detected AP boundaries are used. Trained for 453 sentences by speaker MYI, and evaluated for 50 sentences by speaker MHT (cross validation).

| | All | Intra AP | Inter AP |
|---|---|---|---|
| Baseline model | 117.4 | 46.54 | 2082 |
| Proposed model | 110.4 | 44.41 | 1893 |
| Reduction rate | 5.96% | 4.58% | 9.10% |

**Table 9**. Perplexities for baseline and proposed LM's when the AP boundary labels of the corpus are used for training and those automatically detected are used for perplexity calculation. Speaker MYI's 503 sentences are used for both training and perplexity calculation.

| | All | Intra AP | Inter AP |
|---|---|---|---|
| Baseline model | 117.0 | 57.13 | 1344 |
| Proposed model | 141.2 | 69.18 | 1601 |
| Reduction rate | -20.7% | -21.1% | -19.2% |

## 4. SPEECH RECOGNITION EXPERIMENTS

Continuous speech recognition was done using the proposed LM and the results were compared with those obtained using the baseline LM. All the 503 sentences of ATR speech corpus were used for the POS bi-gram calculation. The proposed LM was arranged for several cases: speakers MYI and MHT, and labeled and detected AP boundaries. The recognition was conducted for the 503 sentences using the 1st pass of the Japanese speech recognition engine, JULIUS, where beam search process went on frame-synchronously [6]. The 2nd pass was not included in the experiments, because tri-gram model was not dealt with in the current paper. Acoustic models are Japanese male tri-phone models of 3000 states with 16 mixtures. The grammar scale factor was kept to 8 throughout the experiments. Table 10 summarizes other conditions of recognition experiments. Denoting number of words correctly detected as $Ncor$, and insertion, deletion, insertion numbers as $Nsub$, $Ndel$, $Nins$, respectively, the word accuracy rate can be defined as follows:

$$WAR = \frac{Ncor - (Nsub + Ndel + Nins)}{Ncor}.$$

Table 11 shows $WAR$'s for the two speakers' utterances when the baseline and proposed LM's were used. Around 2 % improvements are still obtained when the proposed LM was constructed using different speaker's utterances. Figure 3 shows how the proposed LM recovered recognition errors for an utterance "saisho hayai teNpode makikoNde $\cdots$(First, it was involved in a high-pace $\cdots$)."

**Table 10**. Speech recognition conditions.

| | |
|---|---|
| Sampling frequency | 16 kHz |
| Analysis window | 25 ms Hamming |
| Frame shift | 10 ms |
| Pre-emphasis | $1 - 0.97z^{-1}$ |
| Feature vector | 12 MFCC + 12 $\Delta$MFCC + $\Delta$power |

**Table 11**. Word accuracy rates using the baseline and the proposed LM's. The proposed models are calculated using AP boundaries obtained in various conditions.

| Speaker | Language Model | WAR (%) |
|---|---|---|
| MYI | Baseline model | 49.75 |
| | Proposed model, obtained using labeled boundaries of MYI speech | 51.32 |
| | Proposed model, obtained using detected boundaries of MYI speech | 50.66 |
| | Proposed model, obtained using detected boundaries of MHT speech | 50.80 |
| MHT | Baseline model | 51.66 |
| | Proposed model, obtained using detected boundaries of MHT speech | 52.93 |
| | Proposed model, obtained using detected boundaries of MYI speech | 52.91 |

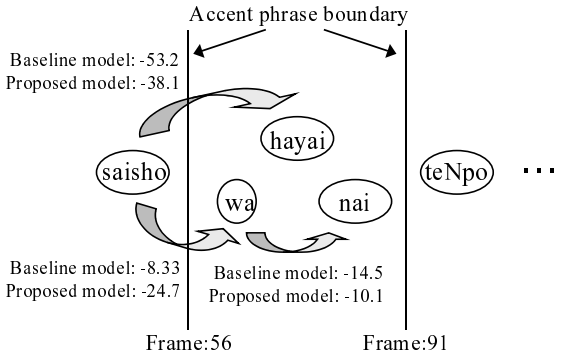


**Fig. 3**. Comparison of language model likelihood for baseline and proposed LM's when recognition errors were recovered by the proposed LM.

## 5. CONCLUSION

A new method was developed to include AP boundary information into n-gram language modeling. Its validity was proved through perplexity reduction from the baseline. Slight improvements in recognition when the proposed LM was used were also shown through the experiments. Further investigations are necessary to extend the method to tri-grams and to improve the AP boundary detection.

The work is partly supported by Grant in Aid for Scientific Research of Priority Areas (#746).

## 6. REFERENCES

[1] E. Nöth, A. Batliner, A. Kießling, R. Kompe, and H. Niemann. Verbmobil: The use of prosody in linguistic components of a speech understanding system. *IEEE Trans. Speech & Audio Processing*, 8(5):519–532, 2000-9.

[2] K. Hirose, N. Minematsu, and M. Terao. N-gram language modeling of japanese using prosodic boundaries. *Proc. Speech Prosody 2002, Aix-en-Provence*, 1:395–398, 2002-4.

[3] http://www.red.atr.co.jp/database_page/digdb.html. Speech Corpus Set B.

[4] http://chasen.aist nara.ac.jp/.

[5] K. Hirose and K. Iwano. Detection of prosodic word boundaries by statistical modeling of mora transitions of fundamental frequency contours and its use for continuous speech recognition. *Proc. IEEE ICASSP'2000, Istanbul*, 3:1763–1766, 2000-6.

[6] T. Kawahara et. al. Evaluation of japanese dictation toolkit -1999 version-. *Technical Report, Spoken Language Information Processing Group, Information Processing Society of Japan*, 2000(54):9–16, 2000.