論文 / 著書情報
Article / Book Information

論題(和文)	
Title	A language model for recognition of continuously uttered sentences
著者(和文)	今井 亨, 安藤 彰男, 古井 貞熙
Author	Toru Imai, Yohei Saito, Akio Ando, Sadaoki Furui
出典(和文)	, Vol. 21, No. 2, pp. 111-113
Journal/Book name	Acoustical letter, the Jornal of the Acoustic Society of Japan (E), Vol. 21, No. 2, pp. 111-113
発行日 / Issue date	2000, 3

J. Acoust. Soc. Jpn. (*E*) **21**, 2 (2000)

ACOUSTICAL LETTER

A language model for recognition of continuously uttered sentences

Toru Imai,* Yohei Saito,** Akio Ando,* and Sadaoki Furui**

* *NHK Science and Technical Research Laboratories,*
1-10-11, Kinuta, Setagaya, Tokyo, 157-8510 Japan

** *Department of Computer Science, Tokyo Institute of Technology,*
2-12-1, Ookayama, Meguro-ku, Tokyo, 152-8552 Japan

(*Received 28 September 1999*)

Keywords : Speech recognition, Language model
PACS number : 43. 72. Ne

1. Introduction

An N-gram language model is widely used for large vocabulary continuous speech recognition. It is usually trained from large corpus, where symbols of beginning-of-sentence and end-of-sentence are inserted at sentence boundaries.[1)] The symbols are effective when an utterance to be recognized is grammatically segmented like utterances in the training. However, continuously uttered sentences spoken in TV programs sometimes disable to be segmented correctly based on the length of pauses between utterances. Such an utterance results in low word recognition accuracy due to a mismatch of the language model between the training and the recognition.

In order to recognize even grammatically mis-segmented utterances, this paper proposes a new language model[2)] which replaces sentence boundaries and pauses with a breath symbol. The model is applied to speech recognition of a live broadcast from the National Diet.

2. Problems of continuously uttered sentences

Utterances spoken in TV programs need segmenting into sentences automatically before the recognition. When the segmentation is based on the length of pauses, it sometimes results in grammatically mis-segmented utterances that start from or end at the middle of a sentence, or include multiple sentences.

In general, the N-gram language model is trained from large texts with special symbols of beginning-of-sentence ⟨s⟩ and end-of-sentence ⟨/s⟩, which are also used as grammatical restriction in the recognition. For example, continuously uttered sentences to be recognized should be ideally segmented on a grammatical sentence basis as follows (in Japanese) :

⟨s⟩ w_1 w_2 'ari' 'mashita' ⟨/s⟩
⟨s⟩ 'watashi' 'wa' w_3 w_4 ⟨/s⟩

However, they may be grammatically mis-segmented as follows :

⟨s⟩ 'ari' 'mashita' 'watashi' 'wa' ⟨/s⟩

The utterance begins from the middle of a sentence and ends at the middle of the next sentence. It frequently happens when a speaker utters with thinking, for example in a live broadcast of the National Diet.

When the conventional N-gram language model is applied to speech recognition of such utterances, bigrams of P('ari'| ⟨s⟩), P('watashi'| 'mashita'), and P(⟨/s⟩| 'wa') are low and make word recognition accuracy worse.

3. Language model with a breath symbol ⟨br⟩

In order to recognize even grammatically mis-segmented utterances, this paper proposes a new language model[2)] with a breath symbol ⟨br⟩ instead of the sentence boundaries and pauses.

In the training of the N-gram language model, all punctuation marks in training texts are replaced with ⟨br⟩ instead of ⟨s⟩ and ⟨/s⟩. Although the replacement with ⟨br⟩ loses grammatical sentence boundary information from N-grams, it maintains the information of phrase boundaries. After the replacement the N-gram probabilities are calculated as usual.

In the recognition, a decoder works under the grammatical restriction where a sentence begins from ⟨br⟩ and ends at ⟨br⟩ instead of ⟨s⟩ and ⟨/s⟩. It is also allowed to get ⟨br⟩ in the middle of a sentence. The pronunciation of ⟨br⟩ is a silence.

For example the grammatically mis-segmented speech stated above is recognized as

⟨br⟩ 'ari' 'mashita' ⟨br⟩ 'watashi' 'wa' ⟨br⟩

in the decoder.

The proposed language model recognizes speech from a breath to a breath and is effective to avoid low word accuracy caused by grammatical mis-segmentation of utterances.

4. Experiment

4.1 Speech data

A recognition experiment on a live broadcast of the National Diet was performed against four types of evaluated speech.

(1) automatically segmented TV speech

The National Diet broadcasted on Oct. 13, 1997, was recorded and its audio data was automatically segmented into 53 utterances based on silences longer than 800 ms. The utterances were 1,493 words spoken by 5 male speakers. The segmentation was based on the power and zero crossings. Grammatically mis-segmented utterances were 71% of all the evaluated utterances.

(2) manually segmented TV speech

The same evaluated audio data was segmented manually and correctly sentence by sentence. The total utterance number was 54. The two types of speech (1) and (2) have the same contents with different boundaries.

(3) automatically segmented read speech

(4) manually segmented read speech

The audio of the broadcast from the National Diet includes background noises, reverberation of the hall, and the influence of specific utterance styles of the Dietmen. An acoustic model of the recognition makes word accuracy worse if any countermeasures are not taken for the acoustic condition. To evaluate the proposed language model without such an acoustically bad influence, another male speaker read the same contents as (1) and (2) including fillers in a quiet room.

4.2 Language models

Training texts for the language model were got from transcriptions of the National Diet of 120 days excluding the evaluated data. Manually transcribed texts, which were 16% of all the training texts, include filler words but others from official gazettes or the Internet do not. The total number of words and sentences were 5.7 M and 168 K respectively. We got the following three types of language models for comparison.

(1) baseline language model (base-LM)

The symbols of beginning-of-sentence ⟨s⟩ and end-of-sentence ⟨/s⟩ were inserted between all the sentence boundaries of the training texts. All punctuation symbols were removed.

(2) language model with punctuation (punc-LM)

Symbols of ⟨s⟩ and ⟨/s⟩ were inserted same as (1) but punctuation symbols in the middle of sentences were kept.

(3) language model with a breath symbol (br-LM)

This is the proposed language model explained in Chapter 3.

All the above language models of bigrams and trigrams were trained with the back-off smoothing using the CMU-Cambridge Toolkit.[3] The vocabulary size was 20 K words. Test set perplexity of the language models is illustrated in Table 1. The test set was processed in the same way as each language model. The perplexity of the base-LM and punc-LM for the manually segmented test data was smaller than that for the automatically segmented one. There was no difference in perplexity for the br-LM between manually and automatically segmented test data. In all the language models, the br-LM showed the smallest perplexity regardless of the way of segmentation. Actually in the example above, the bigram of P('watashi'| 'mashita') in the base-LM was small and 0.00056. Also in the punc-LM, P(',' | 'mashita') and P('watashi'| ',') were 0.019 and 0.022 respectively, because the punctuation mark ',' is not allowed in the middle of the evaluated utterance in the recognition, even if the utterance is grammatically mis-segmented. In the proposed br-LM, P(⟨br⟩|'mashita') and P('watashi'| ⟨br⟩) increased to 0.47 and 0.025 respectively.

Table 1 Test-set perplexity.

Language model	Automatically segmented speech		Manually segmented speech	
	Bigram	Trigram	Bigram	Trigram
base-LM	112.5	87.1	93.2	70.3
punc-LM	83.7	65.0	73.4	56.5
br-LM	63.3	47.7	63.3	49.9

Table 2 Word recognition accuracy.

Language model	Automatically segmented speech		Manually segmented speech	
	TV	Read	TV	Read
base-LM	44.4%	74.9%	45.7%	76.9%
punc-LM	45.6%	79.2%	46.6%	80.2%
br-LM	46.4%	80.8%	46.6%	80.0%

4.3 Recognition experiment

Speech recognition was performed with a decoder of the word dependent N-best search in the first pass and rescoring by trigrams in the second pass.[4] An acoustic model was triphone-HMMs trained from continuous speech database of ATR and ASJ.

Result of the speech recognition for the four types of evaluated speech is illustrated in Table 2. The br-LM obtained better word accuracy than the base-LM and punc-LM for the automatically segmented speech. For the manually segmented speech, the br-LM obtained close word accucary to the punc-LM and better accuracy than the base-LM. This tendency is seen in both the TV speech and read speech. Word accuracy for the TV speech is worse than that for the read speech due to the bad acoustic condition. In the base-LM and punc-LM, the manually segmented speech showed better word accuracy than the automatic one because the training and test data were segmented in the same way. The br-LM showed smaller differences between the two segmentation ways than the other LMs. It should be concluded that the br-LM takes small influence of the segmentation.

The live broadcast of the National Diet achieved lower word accuracy, even in the case of read speech, than other test data like broadcast news.[4] The reason is considered to be higher perplexity of the language models due to less training data of the National Diet transcriptions.

5. Conclusion

This paper proposed the new language model with a breath symbol in order to recognize grammatically mis-segmented utterances also. Its effectiveness was shown in the speech recognition experiment on the broadcast of the National Diet.

References

1) R. Rosenfeld, "The CMU statistical language modeling toolkit and its use in the 1994 ARPA CSR evaluation," Proc. The Spoken Language Systems Technology Workshop, 47–50 (1995).
2) T. Imai, Y. Saito, A. Ando, and S. Furui, "A language model for recognition of continuously uttered sentences," Proc. Spring Meet. Acoust. Soc. Jpn. 2-1-6 (in Japanese) (1999).
3) P. Clarkson and R. Rosenfeld, "Statistical language modeling using the CMU-Cambridge toolkit," Proc. Eurospeech-97, 2707–2710 (1997).
4) T. Imai, K. Onoe, A. Kobayashi, and A. Ando, "A decoder for broadcast news transcription," Proc. Autumn Meet. Acoust. Soc. Jpn. 3-1-12 (in Japanese) (1998).