論文 / 著書情報
Article / Book Information

論題(和文)	連続発話認識のための言語モデル
Title(English)	
著者(和文)	今井 亨, 斎藤 洋平, 安藤 彰男, 古井 貞熙
Authors(English)	Toru Imai, Yohei Saito, Akio Ando, Sadaoki Furui
出典(和文)	日本音響学会1999年春季講演論文集, Vol. , No. 2-1-6, pp. 63-64
Citation(English)	, Vol. , No. 2-1-6, pp. 63-64
発行日 / Pub. date	1999, 3

2—1—6 連続発話認識のための言語モデル*

○今井亨（NHK技研）、△斎藤洋平（東工大）、安藤彰男（NHK技研）、古井貞熙（東工大）

1. はじめに

大語彙連続音声認識の言語モデルにはn-gramが広く使われており、テキストの各文を文頭・文末記号で挟んで学習するのが一般的である[1]。認識対象の各発話が学習時と同様に文単位で区切られている場合には、この文法制約は有効である。しかし、テレビ番組などの連続発話を認識対象とする場合には、無音を基にして音声区間を自動的に切り出す必要があり、切り出された各発話が必ずしも文法的な意味での「文」になっていないことがある。この場合、学習時と認識時で言語モデルの不整合が起こり、認識率が低下する。

本稿では、連続発話から文法的に誤って分割された発話を認識するために、文頭、読点、文末をすべて一つの息継ぎ記号で置き換えた言語モデルを提案する。これを国会中継番組の音声認識に適用し、その有効性を示す。

2. 連続発話認識の問題点

放送音声の認識を番組単位で行うような場合、自動的に音声を適当な長さに分割する必要がある。無音の長さを基準にして音声を分割すると、切り出された発話は、必ずしも文法的な意味での文とはならず、文の途中で始まっていたり、複数の文を含んでいたり、文の途中で終わっているようなことがある。

言語モデルは一般に、文頭記号を<s>、文末記号を</s>として学習されており、認識時にはこの制約下でデコードを行う。例えば認識対象の連続発話を、理想的には

<s> w_1 w_2 あり ました </s>
<s> 私 は w_3 w_4 </s>

のように文法的な文単位で分割してデコードしたい。しかし実際には、無音の長さによっては、

<s> あり ました 私 は </s>

というように、文法的に誤って分割された発話ができてしまうことがある。特に、国会中継における答弁のように、考えながらしゃべっているような場合には、このようなことがよく起こる。

こうした発話を認識する時、従来のbigramを適用すると、文頭、文中、文末において、P(あり|<s>)、P(私|ました)、P(</s>|は)などが低い値を示し、認識率が低下してしまう。

3. 息継ぎ記号
を用いた言語モデル

連続発話から文法的に誤って分割された発話を認識するために、文頭、読点、文末をすべて一つの息継ぎ記号
（breathの意）で置き換えた言語モデルを提案する。

言語モデルの学習時には、学習テキスト中のすべての読点、および各文の境界の句点を
で置き換える。<s>と</s>は用いない。同一記号
への置換により、文頭や文末それぞれに出現しやすい単語の情報が言語モデルから失われるが、文や句の境界、あるいは息継ぎの前後に現れやすい単語かどうかが学習されることになる。学習テキストを
で置換した後は、従来と同様の手順でn-gramを得る。

認識時には、デコーダーは<s>で始まって</s>で終わる文法制約ではなく、
で始まって
で終わる文法制約を使う。もちろん、文中の単語として
を採用することも許す。
の発音としては、無音(sil)を用いる。

例えば前述の例は、デコーダーでは

 あり ました
 私 は

として認識されることになる。

提案する言語モデルは、息継ぎから息継ぎまでを認識対象としてとらえるので、連続発話の分割誤りによる従来の確率値の低下が軽減され、特に考えながら連続してしゃべるような場面での認識に有効であると期待できる。

4. 実験

4.1 評価音声

提案する言語モデルの有効性を確認するために、国会中継の答弁の音声認識実験を、次の4種類の評価音声に対して行った。

(1) 自動分割TV音声

1997年10月13日にNHKで放送された衆議院予算委員会の中継を収録し、男性話者5名の答弁の一部を800ms以上の長さの無音で自動的に分割した（計53文1,493単語）。音声区間の検出には、パワーと零交差数を用いた。この評価音声における文法的に誤った分割は、全発話境界のうち71%

*A Language Model for Recognition of Continuously Uttered Sentences.
By Toru Imai (NHK Sci. & Tech. Res. Labs.), Yohei Saito (Tokyo Institute of Technology), Akio Ando (NHK Sci. & Tech. Res. Labs.), and Sadaoki Furui (Tokyo Institute of Technology)

である。
(2) 手動分割TV音声
分割誤りがない理想的な音声を評価するために、同じ音声を人手で文法的な文の区切りで分割した（計54文）。自動分割TV音声と手動分割TV音声は発話の分割の仕方が違うだけで、全体の音声はまったく同じである。
(3) 自動分割読み上げ音声
(4) 手動分割読み上げ音声
国会中継の音声には背景雑音、議事場の残響、議員特有の発話スタイルなどの影響があり、これらへの対策がなされていない音響モデルでは認識率が低下する。そこで、こうした音響的な劣化要因を取り除いて言語モデルの改善を評価するために、評価音声とまったく同じ内容（間投詞なども含む）を比較的静かな部屋で男性話者が読み上げ、各TV音声とまったく同じ分割を行った。

4.2言語モデル

言語モデルの学習テキストには、評価用TV音声の収録日を除く約120日間の予算委員会、本会議の書き起こしテキスト（16%は音声からの人手による書き起こしで間投詞あり、残りは官報やインターネット上のもので間投詞なし）を用いた。総学習単語数、文数はそれぞれ5.7M、168Kである。これから、次の3つの言語モデルを作成した。
(1) ベースライン言語モデル（base-LM）
コーパスの文頭・文末にそれぞれ<s>と</s>を挿入し、句読点を除去して学習。
(2) 読点付き言語モデル（punc-LM）
<s>と</s>は(1)と同様で、読点「、」も1単語として学習。
(3) 息継ぎ記号を用いた言語モデル（br-LM）
以上3つのタイプについて、語彙サイズ20Kでbigramとtrigramを作成した。
評価音声を自動分割した場合と、人手で文法的な文単位に分割した場合について、各言語モデルのテストセット・パープレキシティを調べた。その結果を表1に示す。ここで評価テキストは、それぞれの言語モデル構築用学習テキストと同じ処理で文頭・文末、句読点の処理を行っている。

表1　テストセット・パープレキシティ

言語モデル	自動分割音声		手動分割音声	
	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

音声の分割の仕方の違いを比較すると、base-LMとpunc-LMでは、やはり学習時と同様の分割である手動分割の方がパープレキシティは小さい。一方br-LMでは、分割の違いによるパープレキシティの変化はほとんどない。各言語モデルの中では、音声の分割の仕方によらず、br-LMのパープレキシティが最も小さかった。実際、前述の例はbase-LMにおいてP(私|ました)=0.00056と小さく、punc-LMでもP(、|ました)=0.019とP(私|、)=0.022であったが、br-LMではP(
|ました)=0.47とP(私|
)=0.025に上昇した。

4.3 認識実験

音声認識は、第1パスでbigramを用いた単語依存N-best探索を行い、第2パスでtrigramを用いたリスコアリングを行った[2]。音響モデルには、ATRと日本音響学会の連続音声データベースで学習されたトライフォンHMMを用いた。
4種類の評価音声に対する各言語モデルでの音声認識結果を表2に示す。

表2　単語正解精度

言語モデル	自動分割音声		手動分割音声	
	TV	読み上げ	TV	読み上げ
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%

言語モデルの違いを比較すると、自動分割音声ではbase-LMとpunc-LMよりもbr-LMは高い認識率を示し、手動分割音声ではbr-LMはpunc-LMとほぼ同等でbase-LMよりも高い認識率を示した。この傾向は、TV音声と読み上げ音声のどちらにも同様にみられる。読み上げ音声の認識率が言語モデルによらずTV音声よりも高いのは、発声様式や雑音の影響と思われる。base-LMとpunc-LMで音声の分割の仕方の違いを比較すると、言語モデルの学習時と同様の分割である手動分割の方が認識率は高い。br-LMは他の言語モデルよりも認識率の変化が小さく、連続発話の分割の仕方の影響を受けにくい言語モデルであると言える。

5．まとめ

連続発話から文法的に誤って分割された発話を認識するために、息継ぎ記号を用いた言語モデルを提案した。国会中継の答弁に対して認識率の向上が得られ、その有効性を確認した。

参考文献

[1] Ronald Rosenfeld, Proc. of the Spoken Language Systems Technology Workshop, pp.47-50 (1995.1).
[2] 今井他、音講論集、3-1-12 (1998.9).