2—Q—4　　語彙を限定しない単語認識方式の検討*

○平林 晃　伊藤克亘　△ 田本真詞　△ 田中穂積 (東京工業大学)

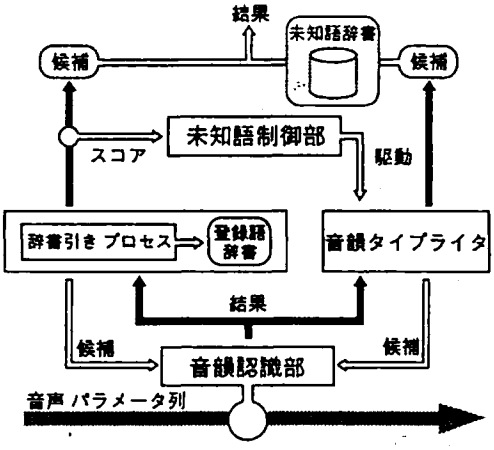
1　はじめに

近年、さまざまな連続音声認識システムが構築されているが、いまだ、極めて制限された領域でしか用いられていない。これらのシステムで用いられている手法のほとんどは、認識できる語彙を大規模にすると、精度や効率が失われてしまう。

我々は、これまでに連続音声認識において未知語を検出し、その音韻系列を推定する手法を提案している [1]。従来の未知語が扱えないシステムでは、認識用辞書に登録されていない語を認識できない。本システムでは、辞書引きプロセスと並行して、音韻タイプライタを駆動し、未知語についても音韻列を生成する。本論文では、この未知語処理を用いた大語彙認識方式について検討する。この方法では、認識すべき語彙が、あまり発話されない語とよく発話される語に分割できる場合に、前者だけで認識用の辞書を作成する。後者については、未知語として検出し、音韻系列を推定する。検出された未知語については推定されたものと同じ、もしくは、似た音韻系列の語を本来の辞書から検索する。この手法では、大抵の頻出語だけからなる発話は、効率のよい認識ができるうえに、頻出語以外の場合でも、若干効率は落ちるが認識は可能なため、全体として性能が向上することが期待できる。また、未知語の音韻系列の推定までできるので、語彙を限定しない認識も可能になる。

本論文では、頻出語の辞書が 2000 語程度の大きさであるとした場合の提案した方式の有効性について検討する。

2　未知語処理の概要



本システムでは、図に示すように、音韻タイプライタと辞書引きプロセスを並行して駆動し、辞書を用いた認識をすると同時に未知語を検出・推定する。

辞書引きプロセスは認識用辞書の制約のもとに認識を進め、音韻タイプライタは認識用辞書によらずに言語モデルのみの制約を用いて認識を進める。しかし、この制約だけでは、音韻タイプライタは、非常に多くの候補を生成するので、入力が未知語でない時にはあまり候補を生成しないように、処理の効率化を図らなければならない。本システムでは、辞書引きプロセスの候補のスコアから閾値を決定し、音韻タイプライタの候補のうちスコアが閾値を越えるものを残す。すると、認識用辞書に登録されている語の認識時は、辞書引きプロセスの生成する候補のスコアが大きく、閾値が大きいので音韻タイプライタは、ほとんど候補を生成しない。逆に、辞書に登録されていない語ならば、辞書引きプロセスのスコアは相対的に小さいので閾値も小さくなり、音韻タイプライタの生成する候補数が多くなる。

3　認識実験

認識候補のスコアは、次式で求められる。

$$
\begin{aligned}
P_{total} &= P_{hmm} + w_{gram}P_{gram} + w_{dur}P_{dur} \\
P_{hmm} &= \frac{\sum_{t=1}^{N_{frame}} \log P_{hmm_t}}{N_{frame}} \\
P_{gram} &= \frac{\sum_{p=1}^{N_{phone}} \log P_{gram_p}}{N_{phone}} \\
P_{dur} &= \frac{\sum_{p=1}^{N_{phone}} \log P_{dur_p}}{N_{phone}}
\end{aligned}
$$

ここで、N_{frame} はその時点でのフレーム数、N_{phone} はその認識候補の音韻数、P_{hmm_t} は各フレームでの音韻モデルと入力を照合して得られるスコア、P_{gram_p} は各音韻の音韻 N-gram モデルのスコア、P_{dur_p} は音韻の継続時間長に関する統計モデルのスコアである。また、w_{gram} は N-gram モデル、w_{dur} は継続時間モデルの重みである。

3.1　音声資料

実験に用いた HMM の訓練用音声資料は ATR 研究用日本語音声データベース [2] から成人男性 1 名 (MAU) が発声したものを用いた。テキストデータは 5240 単語で、奇数番のデータをモデル生成用、偶数番を認識用とした。これらの資料を標本化周波数 16 kHz に変換した。分析のフレーム周期は 5 ms で、14 次のメルケプストラム係数と、その時間方向の変化量、パワーの時間方向変化量の合計 29 個のパラメータを 1 つのコード帳にベクトル量子化した。コード帳のサイズは 1024 である。継続時間長は正規分布で近似した。

3.2　言語モデル

音声認識では、音韻連鎖の統計モデルが音韻認識の精度向上のために有効であることが示されている [3, 4]。音韻 N-gram モデルとは、長さ $N-1$ の任意の音韻連鎖 $p_1,p_2,\cdots,p_{N-1}$ の後に任意の音韻 p_N の続く条件つき確率 $P(p_i|p_{i-(N-1)},p_{i-(N-2)},\cdots,p_{i-1})$ の集合である。N-gram モデルでは、音韻連鎖 $P = p_1,p_2,\cdots,p_{n-1},p_n$ の生成確率は、次のように近似される。

$$
P(P) = \prod_{i=1}^{n} P(p_i|p_{i-(N-1)},p_{i-(N-2)},\cdots,p_{i-1})
$$

*Processing Technique of Unknown Word for Speech Recogniton By HIRABAYASHI Akira, ITOU Katunobu, TAMOTO Masafumi, TANAKA Hozumi (Tokyo Institute of Technology)

日本語で語彙の限定がない場合には、音韻パープレキシティ（等価音韻数）がかなり大きな値となるので、音韻タイプライタが生成する候補に対し、全く何の制約も与えないと、日本語らしくない音韻系列がたくさん生成される。N-gram を日本語の音韻列データから生成すれば、出現頻度の高いパターンほど N-gram の確率が大きくなるので、この値はより日本語らしい音韻系列ほど大きな値を持つと考えられる。本実験では、音韻連鎖の統計モデルとして trigram ($N = 3$) を用いる。trigram を生成するのに使用したデータは、1982 年の日本経済新聞の記事 37 日分の 15,207 文、音韻数で 1,385,082 である。

3.3 認識実験

認識実験に用いた単語の総数は 2620 、そのうち辞書登録語は 1100 単語、未知語は 1520 単語である。

3.3.1 未知語処理を行なわなかった場合

辞書登録語のみの認識率 ($w_{gram} = 0$, $w_{dur} = 0.15$)。

認識率 (%)		
top1	top3	top5
92.5	97.2	98.6

この結果は、音韻連鎖の統計モデルを用いずに、音韻認識の性能と認識辞書による制約のみで認識した結果で、本システムの基本的な認識性能を示している。

3.3.2 未知語処理を行なった場合

未知語のみの認識率と検出率 (w_{dur}=0.15) を以下に示す。

w_{gram}	検出率 (%)			認識率 (%)
	top1	top3	top5	
0.10	50.9	52.0	52.0	95.8
0.15	51.0	52.3	52.3	93.9

言語モデルの重みを変化させた場合の全体の認識率を次に示す。

w_{gram}	検出率 (%)			認識率 (%)		
	top1	top3	top5	top1	top3	top5
0.10	78.8	85.5	86.0	53.4	60.5	61.0
0.15	78.8	85.3	85.8	54.8	62.1	62.6

ここで、検出率、認識率はそれぞれ以下の式で表される。

検出率 = (辞書登録語の正認識数
+ 未知語の検出数) / 認識対象の総数
認識率 = (辞書登録語の正認識数
+ 未知語の認識の総数) / 認識対象の総数

音韻タイプライタが生成する候補は辞書引きのような制約がなく、自由度が高いので、辞書引きプロセスの生成する候補と比較する前に、重みを音韻タイプライタのスコアにかけて、自由度の違いによるスコアの差を補正する。全単語についての認識率がどのように変化するかを以下の表に示す ($w_{gram} = 0.10$)。

重み	検出率 (%)			認識率 (%)		
	top1	top3	top5	top1	top3	top5
1.05	80.5	87.9	88.2	55.3	62.9	63.3
1.00	80.5	87.4	87.9	55.4	63.0	63.4
0.90	81.3	85.7	86.1	57.6	62.9	63.4
0.80	73.0	76.8	77.0	50.6	57.7	58.1
0.70	61.6	62.9	62.9	40.1	47.2	47.6

結果を登録語と未知語に分けて観察すると、重みを減少させた場合、辞書登録語の認識率は 0.90 の時に最大となる。音韻タイプライタの誤った候補が第一位となっていた場合、スコアを減じることにより誤った候補の順位が下がり、かわりに辞書引きによる正しい候補が一位になることがある。このため、誤認識した単語の個数は減少する。一方で、音韻タイプライタによって正解が生成されていた語の数も減少する。0.90 前後で音韻タイプライタの誤りを訂正する効果が最大になるものと思われる。

認識可能な語数のうち、未知語処理なしのときには正解で、未知語処理をしたときに不正解になった語数 (false alarm) の割合を以下に示す ($w_{gram} = 0.10$)。

重み	1.05	1.00	0.90	0.80	0.70
false alarm (%)	11.5	10.8	5.5	2.7	1.1

4 まとめ

本論文で提示した未知語の処理方法によって、認識辞書 2178 単語、未知語 1520 単語において未知語を含めた認識率は最大で 57.6 % 、検出率は 81.3 % となった。また、未知語処理の精度を評価する false alarm のその時の値は 5.5 % であった。

今後の課題としては、認識率を向上させ、この方式を文音声データに適用し、その有効性について検討したい。

謝辞

本稿で使用した日本経済新聞の記事に関するテキストデータベースは、NTT 情報通信処理研究所メッセージシステム研究部から提供して頂きました。また、音声データは、ATR 研究用日本語音声データベースを使用させて頂きました。貴重なデータを使用させて頂いたことを深く感謝致します。また、論文作成に御協力頂いた電子技術総合研究所の速水悟氏に心から感謝の意を表します。

参考文献

[1] 伊藤克亘, 速水悟, 田中穂積 : "連続音声認識における未知語の扱い" 電子情報通信学会, (1991-12)

[2] 武田一也, 匂坂芳典, 片桐滋, 桑原尚夫 : "研究用日本語音声データベースの構築", 音響学会誌 Vol.44 No10, pp747-754, (1988-10)

[3] 荒木哲郎, 村上仁一, 池原悟 : "2 重音節マルコフモデルによる日本語の文節音節認識候補の曖昧さの解消". 情報処理学会論文誌, Vol.30,No.4,pp.467-477,4 1989.

[4] 村上仁一, 荒木哲郎, 池原悟 : "2 重マルコフモデルを使用した単音節音声入力の改善". 音響学会音声研究会資料, No.SP88-29, pp.63-70, 1988