E-034

漸次的な発話理解のための単語部分木を出力する音声認識システム
Incremental Speech Recognition System Outputting Word Tree for Partial Utterance Understanding

高橋 伸弥† 森元 逞† 吉村 賢治† 乙武 北斗†
Shinya Takahashi Tsuyoshi Morimoto Kenji Yoshimura Hokuto Ototake

1. はじめに

近年の計算機技術の進歩に伴い，音声対話システムが実用化されるようになってきた．しかし現状におけるほとんどの音声対話システムでは，音声認識，言語理解，応答生成の各処理が順に行われるため，入力に対して応答するまでに時間がかかるという問題がある．これは，発話の入力と並行して逐次的に音声認識処理を行うものの途中では極めて多数の候補が発生することから，発話終了時に最尤スコアの単語列を決定する必要があるためである．その結果，後段の言語処理にとっては，発話終了時まで処理を開始することができないことになる．

一方，人と人との対話においては，聞き手は聞き取った発話を漸次認識，理解することにより，発話の途中においても，適切な「あいづち」や「うなづき」などをバックチャネルとして相手（話し手）に返すことができる．また話し手は，聞き手からのこのようなバックチャネルによって相手が本当に聞き取れたか，理解できたかを確認しながら，話を展開していく．

我々は，このような人に近い音声認識，言語理解システムを構築することを目標としている．これまでも，認識の途中段階で結果を出力する方式がいくつか提案されている[1–3]が，これらはいずれも音声認識結果として 1 つに絞り込めた（またはその可能性が非常に高い）部分を確定部分として早期に出力しようというものである．これに対し，我々の目標は，前述したように，後続の言語処理も含めて漸次理解を行うことができるシステムの開発である．このようなシステムでは，候補の絞り込みはむしろ言語処理で実現した方が良いと思われる．なぜなら，言語処理の方が意味や文脈などの知識を援用してより適切な絞り込みができ，また言語的な判断も加えて適切なバックチャネルを返すなどの処理が実現できるためである．そこで，このような漸次理解を可能とするような対話システム構築の第一段階として，音声認識の途中において，認識結果が確定しなくても，その時点までに認識された単語を漸次単語木の形で出力するシステムを開発した．

システムは，HTK の HVite[4]をベースとして，そこに上記のような漸次出力機能を組み込んだ．300 文程度の旅行会話文から生成した FSA 言語モデルを用いて動作確認実験を行い，本手法の有効性を検討した．

2. 漸次音声認識の処理方式

2.1 音声認識の原理

一般的な音声認識システムでは，①音響的な特徴を定義した HMM，②認識対象となる単語とその発音を定義した単語辞書，③単語間の接続制約等を定義した言語モデル，の 3 種の情報が用意される．言語モデルとしては種々のモデルが提案されているが，本システムでは FSA 言語モデルを用いる．これは，単語の接続関係を有限オートマトン(FSA)として定義したものである．

HVite ではシステムが起動されると，その初期設定において，言語モデル，単語辞書，HMM から認識用ネットワークがメモリ上に作成される．認識では HMM に定義された状態遷移確率に従って状態間の遷移が行われるが，同時に記号の出力確率を用いて入力音声とのマッチングが行われる．HMM の最終状態に到着すると直ちに後接する単語ノードに遷移し，先頭音素に対応した HMM ノードの開始状態に遷移する．状態遷移確率，記号の出力確率から計算されたスコアは，処理中の状態に対応して用意されたトークンに格納される．また単語ノードに言語的なスコアが定義されていれば，単語が認識された時点でスコアに加算される．以上の処理を発話の終わりまで行い，最終的に得られた候補のうち，最もスコアの良いものを認識結果として出力する．

2.2 言語モデルの木構造への展開

FSA 言語モデルは，ある単語ノードで合流することを許したモデルである．このため，複数のパスを経由して単語に到着する場合には，パスごとにスコアや到着時刻（フレーム番号）が異なるのが一般的である．そこで文献[1]と同様に，FSA 言語モデルから木構造への展開を行い，パスごとの情報（認識されたフレーム番号，パスのスコア）を格納しておくようにする．なおこの展開は，認識時に動的に行うことにより，不要な展開をしないようにしている．

2.3 パスの管理とスコアの計算

処理の進行に伴って認識パスが指数的に増大する理由は，言語モデルでのパスの組み合わせが増大することに加え，連続音声認識では可能性のある単語境界すべてについて認識を試みる必要があるためである．そこで，パス情報を 2 段階で管理することとし，1 段目では単語境界が異なれば別パスとして管理するが，2 段目では単語境界が異なっても同じ単語列として管理することとする．またパスごとの最良スコアを求めるために，1 段目のパスのスコアの変化からその極大値（複数）を求め，その時点で各単語が認識されたとして最良スコアとフレーム番号を求める．なおこのため，出力するのはスコアが最大になった時点より若干遅れることになる．

このとき，あるフレームにおいて，1 段目のスコアが特定の閾値を下回ったものは破棄すると同時に，スコアの上位 N ベスト候補を記録しておく．次に，各 N ベストに対応する 2 段目のパスについて，スコアの最大値の検出を行い，最大値が検出されれば，出力を行う．ここでスコアが単調増加している場合は，まだ増加する可能性があるため，最大値点の取り出しは行わないこととする．

† 福岡大学工学部

Copyright © 2013 by Information Processing Society of Japan and
The Institute of Electronics, Information and Communication Engineers
All rights reserved.

なお，パスのスコアは先頭のフレームから単純に累積した値を用いるのではなく，累積スコアをフレーム数で割ったフレームあたりのスコアを用いる．このとき，実際のスコアのフレーム変化は短時間での細かな増減があるため，移動平均法を用いてスコアの平滑化を行った．

2.4 継続時間のスコアへの取り込み

我々が用いた HMM[5]では，各音素の継続時間は陽には定義されていない．そのため，この HMM をそのまま用いてもスコアの時間的変化がかなり小さく，最大値をうまく検出できないことが多い．そこで n モーラの単語の継続時間の分布は正規分布 $N(n\mu, n\sigma^2)$ であると仮定し，この確率から求めたスコアを全体のスコアに加える．なお，μ，σ^2はあらかじめ学習用の音声コーパスより求めておく．

3. 動作例と考察

本提案方式の基本的な動作を確認する実験を行った．我々の最終的な目標は，後段に適切な言語処理機能を配置し，発話の進行に伴って漸次的に内容を理解し，適当なタイミングであいづちなど返すことができるようなシステムを開発することである．従って，将来的には発話も漸次的に行われたもの（「…はですね，…」「…えーと，…」などのような発話）を対象とすることを考えている．しかしまずは本システムの動作を確認するため，漸次的ではない発話，具体的には旅行ガイドブックに表れるような旅行会話文を対象とし，認識実験をおこなった．ガイドブック等から収集した 300 文から[6]の方法により FSA 言語モデルを作成し，また各文の音声発話を収録した．実験における諸元は表 1 に示した通りである．また短い単語の挿入誤りに対処するため，1 単語あたり一定値のペナルティを加えることとした．なお実際のペナルティ値は実験的に決定した．

図 1 に逐次認識結果の例を示す．各行は左から，新規（N）／更新（U）の別，処理時のフレーム番号，順位，最大値ポイントのフレーム番号，単語列，スコアを示している．最下段が発話完了後に最尤確定された認識結果である．図に示した「地下鉄-の-路線図-を-もらえ-ます-か」という発話は，かなりうまく動作した例である．認識途中では間違った単語列も認識されたが，発話が進むにつれ，正しい単語列に徐々に収束した．正解の先頭単語「地下鉄」は 88 フレームあたりで絞り込まれており，「地下鉄-の-路線図」までは 133 フレームあたりで確定している．また出力された正解単語（<s>，</s>を除く）についての最良時点（フレーム番号）と，発話完了後に最尤確定された単語の認識時点との誤差の絶対値は，単語あたりの平均で約 32 ミリ秒であった．また確定するまでには，さらに 5 フレームの遅れが加わるから，誤差と合わせて最大 82 ミリ秒遅れることになる．一方，「美術館-めぐり-の-ツアー-は-あり-ませ-ん-か」という入力発話ではあまりうまく動作しなかった．途中フレームで「美術館-めぐり-の」までは確定したが，それより後ろの部分は発話全体の認識が終わるまで確定しなかった．これは文末に近づくにつれ，長さの短い単語が続き，また発話の明瞭性等が低下したのが原因であろうと思われる．ただし発話終了後の最尤確定処理では正しい認識結果が得られた．

4. まとめ

入力発話の終了を待たず，認識した結果を漸次的に部分木として出力する音声認識システムの処理方式について報告した．また実際に認識実験を行った結果について報告した．今後は発話自体も漸次的に行われたものについて認識実験を行う予定である．また，漸次理解を行うことのできる言語処理部についても開発を進めていく．

表 1 音声認識実験の諸元

HMM	4 ミクスチャのトライフォンモデル
言語モデル	旅行会話文 300 文から作成した FSA 言語モデル 単語ノード数=878 平均ブランチ数=1.53
テスト発話	上記会話文の音声発話を収録（男性 3 名が各々別文を発話）
認識条件	平滑化のための移動平均数=10 N ベスト数=3，フレーム幅=5

```
N: 22 rank=1 18: <s>-: -58.33
N: 31 rank=3 ,26: <s>- 1 -: -65.66
N: 50 rank=3 ,45: <s>-スカート-: -62.86
N: 51 rank=2 ,47: <s>-貸し-: -62.86
N: 61 rank=1 ,57: <s>-地下鉄-: -60.97
N: 72 rank=2 ,67: <s>-地下鉄-の-: -61.15
N: 75 rank=3 ,65: <s>-ランプ-: -65.97
N: 88 rank=2 ,83: <s>-地下鉄-の-ある-: -62.83
N: 100 rank=2 ,97: <s>-地下鉄-の-バス-: -62.79
N: 109 rank=3 ,106: <s>-地下鉄-の-保険-: -62.13
N: 121 rank=1 ,117: <s>-地下鉄-の-路線図-: -60.23
N: 121 rank=2 ,117: <s>-地下鉄-の-路線図-: -60.23
N: 129 rank=3 ,119: <s>-地下鉄-の-話せる-: -62.78
N: 131 rank=3 ,121: <s>-地下鉄-の-シングル-: -62.58
N: 133 rank=1 ,123: <s>-地下鉄-の-路線図-を-: -60.45
N: 139 rank=2 ,129: <s>-地下鉄-の-路線図-は-: -61.05
N: 153 rank=3 ,146: <s>-地下鉄-の-路線図-を-２-: -61.92
N: 154 rank=2 ,151: <s>-地下鉄-の-路線図-は-買え-: -61.63
N: 155 rank=1 ,151: <s>-地下鉄-の-路線図-を-もらえ-: -60.59
N: 178 rank=1 ,174: <s>-地下鉄-の-路線図-を-もらえ-ます-: -60.95
N: 178 rank=2 ,168: <s>-地下鉄-の-路線図-は-閉まっ-: -61.93
N: 178 rank=3 ,168: <s>-地下鉄-の-路線図-を-お願い-: -61.97
N: 195 rank=1 ,192: <s>-地下鉄-の-路線図-を-もらえ-ます-か-: -60.92
Recognition result:
<s> 地下鉄 の 路線図 を もらえ ます か </s>
== [197 frames] -61.0463 [Ac=-11326.1 LM=-700.0] (Act=4190.1)
```

図 1　認識結果の例

参考文献

[1] P. F. Brown, *et.al.*, “Partial Traceback and Dynamic Programming,” *Proc. of ICASSP-82*, pp.1692-1632 (1982)

[2] 今井他，“最ゆう単語列逐次比較による音声認識結果の早期確定”,信学論(D-II), Vol.J84-D-II, No.9, pp.1942-1959 (2001)

[3] 中川，小林，“連続単語認識における部分単語列の早期検出”，音講論集，3-1-8，pp.97-98 (1998)

[4] S. Young *et.al.*, “The HTK Book (for Ver. 3.0)”

[5] T. Kawahara *et.al.*, “Recent Progress of Open-Source LVCSR Engine Julius and Japanese Model Repository -- Software of Continuous Speech Recognition Consortium,” *Proc. of ICSLP-2004* (2004)

[6] T. Morimoto, S. Takahashi, “Automatic Construction of FSA Language Model for Speech Recognition by FSA DP-Matching,” Lec. Notes in Electrical Engineering, Vol.6, Springer (2008)

Copyright © 2013 by Information Processing Society of Japan and
The Institute of Electronics, Information and Communication Engineers
All rights reserved.