スポッティング法を基にした音声認識

Speech Recognition based on Word Spotting

室井　哲也*　望主　雅子*
Tetsuya MUROI　Masako MOCHINUSHI

要　旨

発話様式の自由度を確保するためにワードスポッティングを文音声に適用した．スポッティングは相補的音素認識モデルにより実行され，文テンプレートによりその駆動が制御される．

上記の音声認識技術と実フィールドでの実験結果について紹介する．

ABSTRACT

We propose a word spotting technique applied to sentence utterances. This technique employs (1) cooperative phoneme models to spot key words and (2) sentence templates to control the word spotting. Field tests results using the technique are also presented.

1．背景と目的

現在の音声認識を使用する場合には語い数や話者性，発話様式，環境騒音のレベルなどさまざまな制約がある．多くの研究機関でこれらの制約を緩めるための研究が行なわれているが，我々は，発話様式の自由度を向上させるための研究を行なっている．

通常，話し言葉では書き言葉と違って，せきばらいや不要語(えー，あのー)の付加，助詞(てにをは)の省略，倒置，言い直しなどの現象が数多く見られる．このため，同じ意図を持つ発話であっても，さまざまな様式で発声される．音声認識を意識しない自然な音声を認識するためには，上記の多様な音声を認識する必要がある．特に，音声認識を対話的に用いる場合には，発話様式によらず発話全体の意図を理解する必要があるので，発話様式の自由度を確保することがますます重要となる．

* 研究開発本部　情報通信研究所
Information and Communication R&D Center, Research and Development Group

ワードスポッティング手法は，発話を一字一句文字に置き換える方法ではなく，意味理解に必要なキーワードだけを抜き出して認識する手法なので，上記の目的に対応できる．

本稿では，音素を認識の単位としたワードスポッティングを文音声に適用する音声認識技術について紹介する．

2．技術

2-1　認識システムの構成

認識システムは，Fig.1に示すように特徴抽出部，音素認識部，スポッティング部，文認識部の4つのブロックから構成されている．

以下に各部の働きについて説明する．

2-1-1　特徴抽出部

特徴量は，次節の音素認識部で使用するLPCメルケプストラム，LPCスペクトラム，およびパワーである．これらの特徴量は，入力音声の1フレーム10msごとに算出される．

2-1-2 音素認識部

音素認識は，相補的音素認識モデル[1]を用いており，任意の音素pが入力音声の第mフレームから第i-1フレームに存在すると仮定したときのスコアsp(p,i,m)を

$$sp(p,i,m) = w_1 DST(p,i,m) + w_2 SPC(p,i,m) + w_3 POW(p,i,m) + w_4 DUR(p,i,m) \quad (1)$$

として計算する．

式(1)の右辺第1項のDST(p,i,m)は,当該区間をLPCメルケプトラムを特徴量とする継続時間制御型状態遷移(DST)モデル[2]でスコアを算出したものである．このモデルは，HMM[3]と計算機構が似ているが，統計量を使用していないので，学習データが1個のときでもモデルは生成できる．しかし，基本的には，学習データが増えるほど精密化される，という特徴を持つ．

一方，右辺の第2項以下は，それぞれ当該区間をスペクトラム(SPC)，パワー(POW)，継続時間(DUR)の観点で評価したものである．これらのスコアは，ヒューリスティックにもとづいたルールで評価されるため，一部の評価パラメータ(例えば，母音の共振周波数や各音素の平均継続時間など)を除いては，学習データ量に依存しない精度を持つ．

そこで，各評価関数に対する重みw_1〜w_4は，Σw_iを一定の値に保ちつつ，学習データが少ないうちは，w_1の値を小さくしておき，学習データが多くなるにつれて，w_1の値を大きくするように設定している．ただし，重みw_1〜w_4の初期値，上限値，下限値，Σw_iの値は，各音素ごとに異なる．例えば，/p/や/t/などの破裂音は，そのパワー形状に特徴があるため，POWに関する重みw_3の初期値，下限値共に他の音素より大きな値になっている．

2-1-3 スポッティング部

スポッティング単位は単語には限らない．文節のように途中に不要語やポーズが挿入されない単位であればスポッティングすることができる．この単位は，音節列を終端記号とする書き換え規則で記述された文法で定義される．

[書き換え規則の例]

何曜日 = 日|月|火|水|木|金|土 + 曜日
曜日 = ようび
日 = にち
月 = げつ
・・・・・・・

また，個々の音節ごとに音素のネットワークで記述されたテーブルを持っているので，音節列は音素のネットワークに展開される．つまり，文法上は陽になっていないが，音節列(例えば「にち」)を左辺，音素ネットワークを右辺とする規則が存在する．

スポッティング部は，書き換え規則の最上段の左辺の記号(上記の例では「何曜日」)をスポッティングするように文認識部から命令を受け，認識結果(例えば「日 + 曜日」)とそのスコア，始終端のフレーム番号を返答する役割を持っている．

始端が第Mフレーム，終端が第I-1フレームで左辺の記号Wが認識されたときのスコアsw(W,I,M)は，右辺の記号列を$w_1 w_2 ... w_L$とするとき，次式で計算される．

$$sw(W,I,M) = \Sigma_j spw(w_j,i_j,m_j) \quad (2)$$

ただし

$$m_1 = M \quad (3)$$

$$m_j = i_{j-1} (j>1) \quad (4)$$

$$i_L = I \quad (5)$$

spw()は，左辺記号が音節列でない場合，式(2)の形式で計算される右辺の記号w_jに対するスコアsw()であり，左辺記号が音節列である場合，式(1)で説明した音素のスコアsp()である．

スポッティング部では，式(2)で定義される左辺記号Wのスコアを最大にする右辺記号列$w_1 w_2 ... w_L$と，始端M，終端I-1の組み合わせを動的計画法を用いて求める．また，認識速度を向上させるためビームサーチ法を採用している．

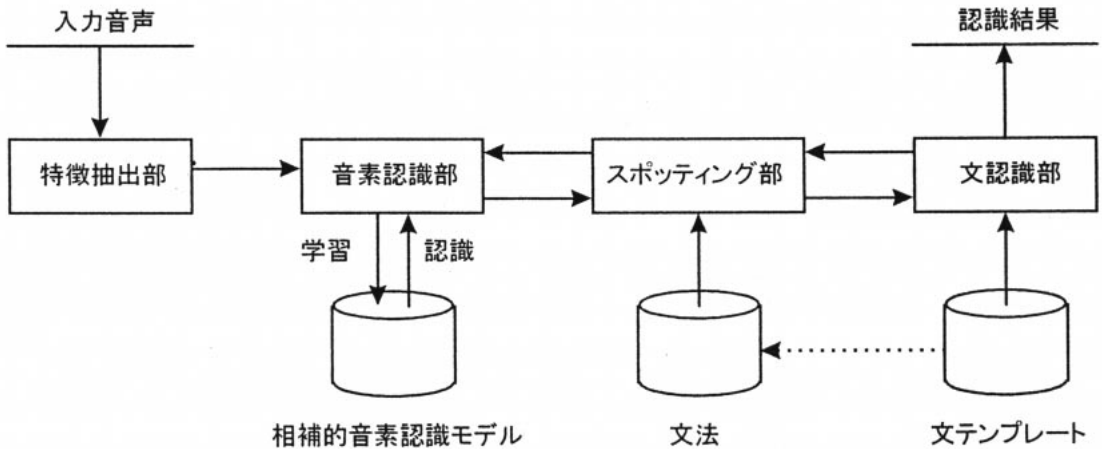

Fig.1 Block diagram of speech recognition system

2-1-4 文認識部

話し言葉の文章を認識するために文テンプレート[4]と呼ぶデータベースを用いて認識制御を行なっている．文テンプレートは，単語の意味を用いて，単語の共起関係や出現順序などの文型を規定している．

文テンプレートの記述に従って，書き換え規則の左辺記号と入力音声中の範囲を限定して，スポッティング部を駆動し，スポッティング結果の右辺記号列，存在位置，スコアを得る．次に，得られた結果と文テンプレートの記述から新たに認識すべき左辺記号と解析範囲を決定する，という動作を繰り返す島駆動型の認識方式である．なお，話し言葉に特有な不要語やポーズの挿入を許すために，スポッティングされた左辺記号の存在位置は，式(4)のように前の単位の終端と後の単位の始端が重なっている必要はない．

また，スポッティング部を駆動する際の左辺記号はなるべく長い方が良い．なぜならば，スポッティング性能自体が安定する(2-3-1参照)上に，次に解析すべき範囲をより狭くできるからである．そこで，スポッティング部へは，自立語単体だけでなく助詞や助動詞がついた文法を与えている．ただし，助詞や助動詞を常に正しく発声する必要はなく，結果的に自立語だけがスポッティングされることはあり得る．逆に，2-3-2の実験のように長い複合語は，その内部に不要語やポーズが含まれる可能性がある．この場合は，複合語の構成要素を個別にスポッティングするようにしている．

2-2 学習

本稿で紹介している音声認識方式は，特定話者方式であるために，話者ごとに音声の学習が必要となる．この学習手順について，簡単に説明する．学習は，システムが提示する言葉を順に読み上げていくことで自動的に実行される．

まず，「あ行，か行，さ行，た行，ぱ行の音節」「あん，えん，おん」を学習する．この段階では，音素の中で出現頻度が高くかつ継続時間が長い母音，摩擦音，撥音のモデルを構築することを目的としている．この時点で上記の音素のDSTモデルとSPC,POW,DURの認識ルールが確立されるので，認識動作が可能となる．

次に，単語，短文からなる209発声から，個々の音素モデルの学習を行なう．これは連結学習[5]と呼ばれる方法で，通常の認識と全く同様にスポッティングを行なって，その照合経路のバックトラックを行なって，音素の存在区間を推定し，当該区間の特徴量をもとに音素認識モデル(主にDSTモデル)の学習を行なう．

本方式では，音声学的な意味ではなく，音響的に1つのまとまった特徴を持つ区間を「音素」と定義しているため，音素の数は307個となる．学習リストはなるべく多くの音素が含まれるように設計してあるが，学習終了の段階でも1割程度の未知音素が含まれている．

2-3 認識実験

2種類の認識実験の結果について述べる．音声試料の諸元をTable 1に示す．

Table 1 Specification of speech samples

実験1	話者	男女各5名
	収録場所	一般オフィス
	発声方法	読み上げ
	評価文数	100
実験2	話者	女性1名
	収録場所	お客様相談室
	発声方法	復唱
	評価文数	293
マイクロフォン		接話型指向性
収録器材		PC＋サウンドボード
分析周波数		16kHz

2-3-1 都市名の認識実験(実験1)

ここでは，「○○まで」「××から○○まで」という2種類の文型で都市名を認識する実験を行なった．都市名は電子協提案の100都市名を用いた．なお，25%の評価文の先頭に故意に不要語を付加している．認識結果はTable 2の通りである．

ここで，抽出率は正解数／発声された単語数，適合率は正解数／スポッティングされた単語数の割合を示している．また，上段は，付属語「まで」「から」を含む文法を用いてスポッティングした場合の値であり，下段は，単に都市名だけの文法でスポッティングした結果である．

スポッティングは，2-1-3で述べたようにスコアだけでなく，存在位置も同時に決定する方式である．Table 2の上段の実験では，付属語「から」「まで」は，それぞれ100都市全てに付加されるので，識別性能(スコアの相対的な差)には本来影響を与えない．つまり，Table 2の上段と下段の認識性能の差は，存在位置の推定の正確さの差であると見ることができる．100都市名は2音節(千葉，野田など)22個，3音節(別府，桐生など)46個と短い単語が多く認識が難しいが，付属語付きの長い単位でスポッティングすると認識性能が安定することがわかる．

Table 2 Results of recognition tests (100 city-names)

	抽出率(%)	適合率(%)
付属語付き	95.5	97.3
付属語なし	78.7	73.1

2-3-2 実フィールドでの実験(実験2)

次に，本稿で紹介した認識システムを実フィールドで実験した結果について述べる．実験した場所は，当社のお客様相談室で，電話による問い合わせを受け付ける機関である．ここでは，商品の価格や使用方法などの問い合わせにオペレータが応対しているが，顧客の質問の商品を「はい，○○ですね，ありがとうございます」のように復唱して確認している．この復唱から商品名をスポッティングする実験を行なった．実験は，女性オペレータが1996年4月から1996年12月の間に実際の顧客と応対する中で行われた．なお，本稿で説明する実験結果は,このとき同時に収録された最大4秒の音声データをもとに認識実験をしたものである．

認識対象となる商品名は，「リポートTS41」「マイリポートNT350」「雄弁家」などの141種類である．ただし，「41」は「ヨンジュウイチ」「ヨンイチ」の2種類の読みを用意する，など1つの商品でも複数の読みを用意した．また，「リポート」「TS」「41」の間には，格助詞「の」やポーズ，「えー」などの不要語が入りうること，「リポート」は省略される場合もあることから，「リポート」「TS」「41」は，それぞれ個別にスポッティングする，というように1つの商品名を1〜4のスポッティング単位で認識するようにした．認識結果をTable 3に示す．なお，認識率は，個々のスポッティング単位が認識された割合ではなく，1発話中から正しく商品名が特定できた割合を表している．まれに例外(リポートTS41の発声に対し，リポートがリジェクトされてもTS，41が認識できれば特定できる)はあるが，1発話に含まれる全てのスポッティング対象が正しく認識された割合と考えて良い．

Table 3で示した標準の学習方法とは，2-2で示した方法である．また，適応学習1は，標準の方法で学習された音素モデルをタスクで使用する言葉55発声で追加学習したもの，適応学習2は，適応学習1の音素モデルをさらに実稼動時の58発声で追加学習したものである．

このタスクで頻出する外来語(上記の例では「TS」など)の認識精度を上げるために，適応学習1を行なったが，あまり認識率は改善されていない．その原因は，適応学習1で発声された音声と実稼動中に発声された音声の発話速度の変動によるものと考えられる．例えば，適応学習1の段階で発声された「マイリポート」が680msであるのに対し，同日に発声された実稼動時には520msと，約76%の継続時間になっている．

適応学習2は， 発話速度の変動を，実際の稼動時に発声された音声を用いて学習させる目的で行なった．ここで用いた58発声は，稼動実験開始から1週間の間に収録された発声で，評価用の発声からは除外した． Table 3からわかるように，この学習の効果は非常に大きい．実際の応対中に話題となった商品であるため発声された商品名も偏っているが，自動的に頻度の高い商品名だけを学習していることになり，効率的な学習方法である．

Table 3 Results of recognition tests (field tests)

学習方法	認識率(%)
標準	75.8
適応学習 1	78.5
適応学習 2	91.8

3．成果

相補的音素認識モデルと文テンプレートを用いてスポッティングを行なう音声認識システムについて述べた．この音声認識システムは現在，実フィールドで稼動テストを実施している．

認識実験の結果から，

(1) 長い単位でスポッティングする方が認識性能が良いこと

(2) 認識の際に発声した音声で学習すると効果的であることがわかった．

4．今後の展開

スポッティングをベースとする認識方式は，発話様式の自由度を確保できるメリットと裏腹に，解析空間が広くなるため，認識速度，語い数の制約が強くなる．今後は，フレーム同期型のスポッティング方式など実時間処理可能な解析方式，発話データの解析による文テンプレートの精密化などについて検討する．

謝辞

認識実験に協力いただいたお客様相談室の皆様に深謝いたします．

参 考 文 献

1) 室井 :相補的な音素認識モデルを用いたワードスポッティング,日本音響学会平成5年春季講演論文集 ,1-4-1 (1993) pp.1-2

2) 室井,米山 :継続時間制御型状態遷移モデルを用いた単語音声認識,電子情報通信学会論文誌D-II Vol.J72-D-II No.11 (1989) pp.1769-1777

3) 中川 :確率モデルによる音声認識,電子情報通信学会 (1988) pp.29-87

4) 望主,室井 :文テンプレートによる発話文認識,音声情報処理 7-20 (1995) pp.129-134

5) K.F.Lee,H.W.Hon :Large-Vocabulary speaker independent continuous speech recognition using HMM, ICASSP'88 (1988) pp.123-126