音声入力　分散型音声認識　誤り訂正

Technology Reports

音声による文字入力の使いやすさ向上を目指した効率的な音声認識誤り訂正技術

携帯端末でのメール文章の作成や音声翻訳といった大語彙連続音声認識を用いたアプリケーションが導入されている．音声認識では認識誤りは避けられず，効率的に訂正する仕組みが重要である．このため，ユーザが誤り区間を指定し，その対象区間を含む部分的な音声特徴量を再認識することにより，連続音声認識の誤りを訂正する方法を提案する．提案手法では認識誤り単語の3～7割を訂正する効果が得られることを，実験により確認した．

先進技術研究所　中島 悠輔　張 志鵬　仲 信彦

1. まえがき

音声認識の性能が向上し，文章の作成や音声翻訳といった大語彙連続音声認識のアプリケーションが携帯端末で実用化されてきている．それに伴い，例えばFOMA 905iシリーズ以降，旅行会話日英／英日翻訳iアプリにて旅行会話文の音声入力が可能となり，らくらくホンプレミアム以降のらくらくホンシリーズでも，メール作成に対して音声で文字入力ができる「音声入力メール」サービスを提供するなど，ドコモの携帯端末において音声認識機能が拡充されてきている．しかしながら，入力音声が認識に用いる音響モデルや言語モデルと一致しないことに起因する認識誤りは避けられない．そのため，音声認識の利便性を向上させるには，認識性能の向上と同時に，認識誤りを効率的に訂正する仕組みも重要である．

音声認識誤りを訂正する代表的な方法の特徴を**表1**に示す．緒方ら[1]は初回認識時に単語ごとのNベスト[*1]を提示して，訂正が必要な場合にユーザが正解単語を選択する方法を提案しているが，訂正できるのは正解がNベストに含まれる場合に限られる．それ以外の単語も入力できる手法として，キー入力や再発声して認識し直す方法があるが，ユーザに操作の負担をかける．そこで，キー入力や再発声によるユーザの訂正操作を最大限不要としつつ，N

表1　代表的な認識誤り訂正方法

誤り訂正方法	訂正性能	初回認識に含まれない候補への変換	訂正対象区間	ユーザの操作負担	処理方法	処理量
Nベスト	高	無	単語	中	言語	軽
キー入力	高	有	単語	重	言語	軽
再発声	低	有	複数単語	重	音響＋言語	重
再認識（提案法）	中	有	複数単語	軽	音響＋言語	中

＊1　**Nベスト**：最も高い可能性をもつ仮説の系列．

ベストで対応できない訂正候補の単語が全文認識（初回認識）結果に含まれない場合も含めて訂正する手法を提案する．

提案手法では，ユーザが誤り区間を指定すると，部分再認識により認識誤りを訂正する[2]．本手法は，誤り区間の前後にある初回認識の正解単語列を再認識の探索時の拘束条件とすることで，訂正効果（誤り箇所の訂正による認識率向上の効果）を高めている．再認識の処理は，拘束条件を利用する以外，通常の大語彙連続音声認識と同様の処理のため，複数単語にまたがる訂正対象区間に対応できるのが特長である．再認識の対象区間を前後の正解単語列を含む区間に設定することで，全文を対象区間にする場合に比べて処理の負荷を軽減できる．また，初回認識と再認識を同一のモデルで行う必要はなく，例えば再認識専用のモデルを用いるといった構成も可能である．

また，訂正効果に加え，ユーザの操作に対する応答性を高める手法として，ユーザが誤り区間の始点を指定した時点で，再認識対象区間の終点を推定し，再認識処理を実行する方法も検討した．

本稿では，提案する部分再認識による認識誤り訂正方法と実験により確認した効果について解説する．

2. 部分再認識による音声認識誤り訂正

2.1 音声認識処理

音声認識処理の概要を**図1**に示す．認識デコーダは，未知の入力音声Xに対する単語列$\hat{w}$を探索する．探索は，音響モデルによる確率$P(X|w)$と言語モデルによる確率$P(w)$の積（対数スケールでは和）が最大となる単語列（最尤解）を探索する．音響モデルは音声特徴量[*2]の時系列信号の確率モデルである．言語モデルは与えられた単語列(Nグラム[*3])に対してその出現確率を与えるモデルである．

$$\hat{w} = \arg\max_{w} P(X|w)P(w) \qquad (1)$$

ただし，最尤解が必ずしも正解とは限らず，誤認識となる場合もある．

また実際の認識処理は，単語数や音素[*4]数が数千以上と膨大になることが多く，探索の過程ではその何乗かの認識結果の候補（仮説）が生成されるため，効率的な探索のために，有望でない仮説の枝刈りが行われる．したがって，必ずしも全仮説における最尤解が出力されるとは限らない．

2.2 誤り区間の再認識

提案する再認識を用いた認識誤り訂正のシステム構成例を**図2**に，ユーザインタフェース例を**図3**に示

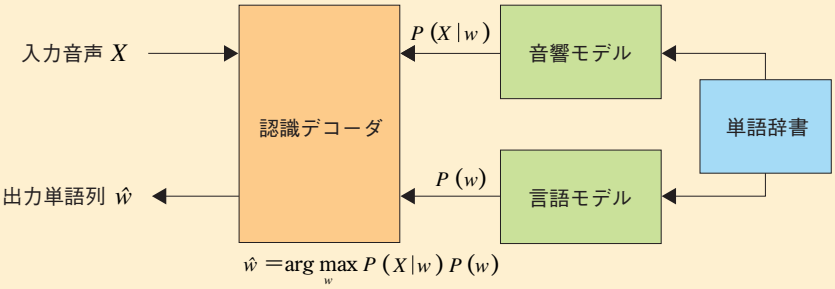

図1　音声認識処理の概要

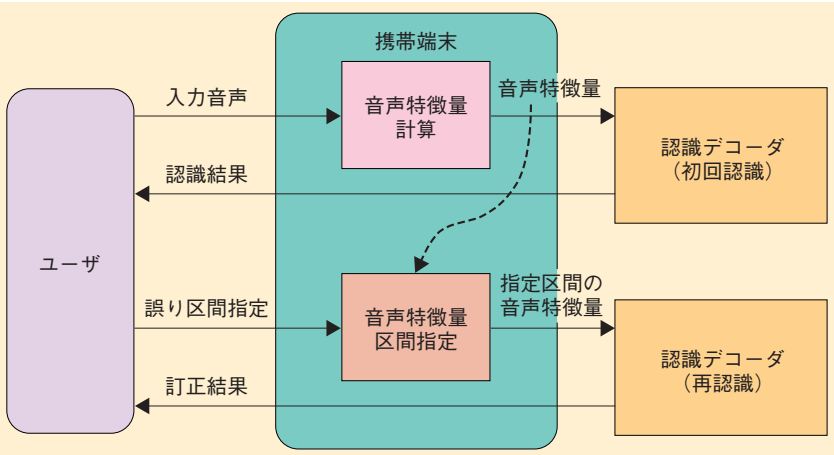

図2　システム構成例

＊2　**音声特徴量**：短時間ごとに切り出された音声信号から抽出される特徴ベクトルの時系列．
＊3　**Nグラム**：N個の単語の連鎖．
＊4　**音素**：言語における意味の弁別に用いられる最小の音の単位．

す．まず，表示された認識結果の誤りを訂正するために，ユーザが実行する編集操作により訂正対象区間を確定する．編集操作としては，例えば削除のためのキー入力や（図3③），タッチパネルで対象区間を指でなぞる動作がある．訂正対象区間が確定されると，訂正対象区間とその前後の単語列を含む区間を再認識対象区間とし，その音声特徴量を再認識用の認識デコーダへ入力する（**図4**）．認識デコーダは，訂正対象区間の前後の単語列を探索時の拘束条件として，再認識対象区間の音声特徴量に対して認識処理を実行する．訂正対象区間の再認識の結果は，初回認識結果の対応する部分と置き換えられ，訂正結果としてユーザに表示される（図3④）．

再認識処理において，探索する仮説の両端を訂正対象区間前後の正解単語列で拘束することで訂正効果が期待される．探索する仮説の例を**図5**に示す．多くの認識デコーダでは，探索の過程で信頼度の低い候補を枝刈りする．初回認識（図5(a)）では信頼度が比較的高い仮説「で形」や「人か」を残し，正解の「で肩」を含む仮説は，他の候補と比べて信頼度が低いために，この時点で枝刈りされる．再認識（図5(b)）では，誤り区間の前の正解単語「で」と「だ」で探索する仮説を拘束することで，「人」に連なる仮説は探索から除かれ，「肩を組ん」を通る仮説が生き残る．

また，再認識で用いる音響モデルや言語モデルの選択により，訂正効果の向上を図ることができる．分散型音声認識（DSR：Distributed Speech Recognition）[*5]の場合，計算量の大きい初回認識はサーバで実施し，認識区間が短くなることで計算量を限定できる再認識はクライアント（端末）で実施することが考えられる．このように，サーバでは大規模で一般的なモデルを用い，クライアントでは小規模でもユーザに適応したモデルを用いることで，訂正効果が期待できる．また，初回認識と再認識で異なるモデルだけでなく，同一のモデルを用いる場合でも，前述のように，正解単語列で探索する仮説の両端を拘束する手法で，訂正効果が期待できる．初回認識と同一のモデルで訂正できれば，より小規模で簡易なシステム構築が可能になる．特に，DSRの場合，初回認識で用いた音声特徴量や認識結果をサーバで保持しておき，再認識

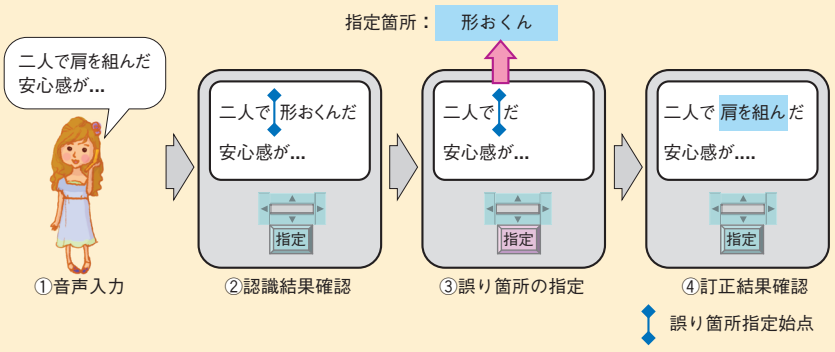

図3　ユーザインタフェース例

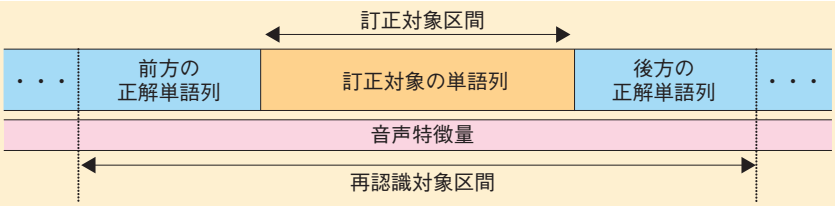

図4　再認識対象区間

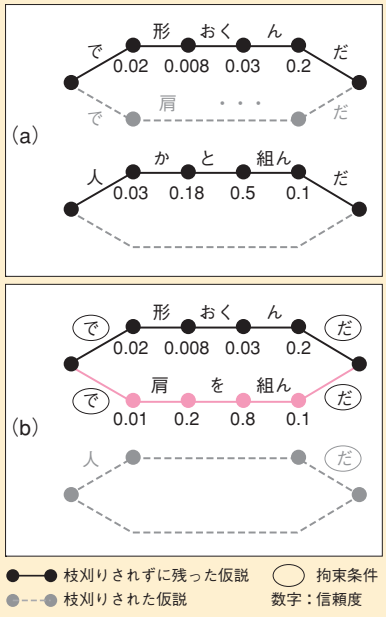

図5　誤り区間の前後の単語で拘束した仮説の例

＊5　**分散型音声認識（DSR）**：入力音声からの音声特徴量抽出処理は移動端末で行い，音声特徴量からの認識結果変換処理をサーバで行う音声認識．

で利用できれば，クライアントとサーバとの間の伝送量の削減と，それによる応答の高速化が期待できる．

2.3 再認識対象区間の自動推定

ユーザの操作に対する応答性を高める手法として，誤り区間の始点を指定した時点で，再認識対象区間の終点を推定する方法を提案する．再認識対象区間の終点を推定できれば，ユーザが訂正対象区間の終点を指定する前に再認識処理を開始できる．ユーザが訂正対象区間の終点を指定した時点で，あらかじめ再認識した結果のうち訂正対象区間に該当する部分を表示することで，ユーザの操作に対する応答性を高めることができる．

再認識訂正対象区間の終点は，再認識の拘束条件となるため，初回認識結果が正解であることが望ましい．推定する方法として次の3つが考えられる．

(a) 信頼度

音声認識処理の過程で得られる単語ごとの信頼度を再認識対象区間の終点の判定に利用する．ここでは，信頼度があるしきい値以上になる単語を再認識対象区間の終点とみなし，拘束条件とする（**図6**(a)）．

(b) 単語数

通常の大規模連続音声認識ではNグラムが利用されているため，誤認識が複数の単語にまたがって連鎖して発生する場合も多い．そこで，再認識対象区間を単語数により区切り，再認識対象区間の終点の指定に利用する（図6（b））．

(c) ポーズ（無音）

句点「。」や読点「、」で表されるポーズの部分は，それ以外の部分に比べると，認識誤りが発生しにくい．訂正対象区間の始点に後続する最も近いポーズをショートポーズ(SP)，訂正対象区間の始点を含む文の終端をロングポーズ（LP）とし，再認識対象区間の終点の指定に利用する（図6（c-1），(c-2)）．

3. 実験

認識誤りを部分再認識で訂正する効果を実験により確認するため，再認識対象区間の両端を正解単語で拘束した実験を行った．また，再認識対象区間の終点自動推定についても効果を確認するため，再認識対象区間の始点は正解単語にし，終点は自動推定し実験を行った．

3.1 実験条件

実験において，初回認識と再認識で音響モデルは共通のものを用い，言語モデルによる効果を確認した．言語モデルは，入力音声に近い言語モデルとして新聞から学習した2万語と，汎用的で語彙数の大きい言語モデルとしてウェブから学習した6万語を用いた．入力音声には，新聞記事読上げ音声データベース（JNAS）から選定した100文章（男女各6人，1,504単語）を用いた[3]．認識処理には研究開発で一般的な大語彙連続音声認識エンジン「Julius」[4]

図6 再認識対象区間の自動推定の例

を用いた．音声特徴量は，標準設定のMFCC（Mel-Frequency Cepstrum Coefficient）[*6]が12次元，ΔMFCCが12次元，Δ対数パワー[*7]が1次元の計25次元を用いた．認識結果には，単語トライグラムによる逆向き探索を用いた第2パスの最尤解を採用した．音響モデルは，16混合性別非依存の状態数2,000のトライフォン[*8]のHMM（Hidden Markov Model）[*9]と状態数129のモノフォン[*10]を用いた[5]．

3.2 実験結果

(1) 再認識対象区間の両端を指定した場合

再認識対象区間の両端を正解単語で指定した場合の再認識訂正の検証を実施し，その結果を**表2**に示す．初回認識の結果は，言語モデルが新聞の場合の単語認識率が94.8％，ウェブの場合は86.0％となった．誤り単語数は，言語モデルが新聞の場合78単語，ウェブの場合211単語となり，再認識による訂正の対象とする．

初回認識と再認識の言語モデルが共に新聞（新聞ー新聞）を用いた結果，前後の正解単語による拘束なしの場合の単語訂正率は28.2％（対象78単語），拘束ありの場合は41.0％であった．再認識は拘束なしの場合でも，入力音声が誤り区間とその前後の正解単語を含む区間（再認識対象区間）に設定されている点，また，再認識対象区間外からの拘束の影響がない点で，初回認識と異なる．そのため，再認識では初回認識と異なる候補が最尤解として出力されたと考えられ，誤り区間前後の正解単語による拘束がなくても，一定の訂正効果が確認できた．また，拘束ありの場合の訂正効果が拘束なしの場合に比べて高いことを確認した．さらに，図5に示したように，正解の候補であっても初回認識で枝刈りされる場合があるが，再認識で枝刈りされずに残る効果を確認した．

初回認識にウェブ，再認識に新聞の言語モデルを用いた場合（ウェブー新聞），初回認識の言語モデルでは未知語で，再認識の言語モデルでは未知語でない単語6個のうち4個が訂正された．また，未知語の前後の単語も含めた誤認識を訂正できることを確認した．例えば，「0ー五と」と音声入力（正解）した区間が，初回認識では未知語「ー」（ヨミは「タイ」）を含むため「食べたい事」と誤認識されたが，再認識により「レイー五と」となり，「ー五と」の部分が訂正された（**図7**）．

初回認識と再認識の言語モデルが共に新聞（新聞―新聞）で前後の正解単語による拘束ありの場合，初回認識と再認識の結果を合わせた単語認識率は96.9％となる．ウェブー新聞の単語認識率95.9％を超える結果が得られた．ただし，ウェブー新聞の場合の単語認識率95.9％は，新聞の初回認識の単語認識率94.8％を上回った．初回認識と再認識が共に入力音声に近い言語モデルで認識した場合の認識率が他の場合と比べて高くなり，また，再認識だけでも入力音声に近い言語モデルを用いれば，初回認識に入力音声に近いモデルを用いた場合と同等以上の認識率が得られる．初回認識でも再認識でも，入力音声に近いモデルを用いることが効果的と考えられる．

一方，言語モデルが初回認識と再認識共にウェブの場合（ウェブーウェブ）の単語訂正率は29.9％（対象211単語）と，一定の訂正効果

表2　再認識結果

初回認識	再認識		単語認識率 (N=1504)	単語訂正率
言語モデル	言語モデル	拘束		
新聞			94.8%	
	新聞	無	96.3%	28.2% (N=78)
		有	96.9%	41.0% (N=78)
ウェブ			86.0%	
	ウェブ	有	90.2%	29.9%(N=211)
	新聞	有	95.9%	71.1%(N=211)

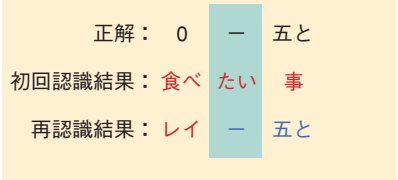

図7　未知語を含んだ区間の結果例

＊6 **MFCC**：メル周波数ケプストラム係数．人間の聴覚を模した音声特徴量係数の系列．
＊7 **Δ対数パワー**：パワーの1次差分．
＊8 **トライフォン**：3つ組の音素．
＊9 **HMM**：隠れマルコフモデル．確率モデルの1つ．
＊10 **モノフォン**：前後の音素を考慮しない音素モデル．

はみられた．初回認識と再認識の言語モデルが入力音声と一致しない場合でも，再認識による訂正の効果を確認できた．すなわち，入力音声と言語モデルの整合性に関係なく，再認識による訂正の効果を確認することができた．

⑵ 再認識対象区間を終点自動推定した場合

再認識対象区間の始点を正解単語にし，終点は自動推定した場合の効果を検証する実験を行った．ウェブの言語モデルを用いて初回認識した単語認識率は82.4％であり，誤り区間は56区間，誤り単語数は264単語であった．誤り区間の判定において，認識結果の形態素解析[*11]は形態素解析システム「ChaSen」，正誤判定は大語彙連続音声認識エンジン「Julius」付属のスクリプトを用い，自動化している．再認識区間の終点を信頼度，単語数，SP，LPのそれぞれで決定し，再認識を行った（**表3**）．再認識には新聞の言語モデルを用いた．なお，信頼度のしきい値および単語数は一番認識率が高くなるものを使っており，それぞれ91.4％と91.9％である．SPやLPで区切った場合がそれぞれ93.3％，93.6％となり，信頼度や単語数で区切る場合に比べて訂正率が高い．これは，信頼度や単語数で区切る場合は，区間終点の単語列が正解であるとは限らず，一方，SPやLPで区切る場合は，正解である可能性が高くなる傾向にあり，再認識対象区間に複数の誤り区間が含まれることがあるものの，終点を指定する方法に訂正率が近いためと考えられる．

表3　再認識対象区間の終点を自動推定した場合の再認識訂正の結果

終点自動推定法	単語認識率 (N=1504)	単語訂正率 (N=264)
信頼度	91.4%	51.1%
単語数	91.9%	54.0%
SP	93.3%	61.9%
LP	93.6%	63.6%

4. あとがき

本稿では，ユーザが指定した訂正対象区間の前後の正解単語を含む区間の音声特徴量を，正解単語を探索時の拘束条件に用いて再認識することによる音声認識誤りの訂正方法を提案した．実験により，拘束条件を用いた場合の効果を確認し，入力音声と言語モデルの整合性に関係なく，再認識による訂正の効果を確認した．特に，初回認識と同一の言語モデルの場合でも3～4割の単語を訂正できることは，既存の音声認識システムに大きな変更を加えることなく性能を改善できる可能性を示しているといえる．

さらに本稿では，訂正対象区間の始点が指定された時点で再認識対象区間の終点を自動的に決定し，再認識処理を開始できる手法を提案し，終点を指定する場合に近い効果が得られることを確認した．

本稿では，初回認識と再認識で共に最尤解をユーザに提示する場合の訂正効率向上を検討したが，Nベストを提示した場合の訂正法や，キー入力による訂正法との組合せ方式も，今後の検討が望まれる．さらに訂正効果を向上する手法や，音声認識システムへの導入の検討を進めていく．

文　献

[1] J. Ogata and M. Goto：“Speech Repair：Quick Error Correction Just by Using Selection Operation for Speech Input Interfaces，”in Proc. Interspeech 2005，pp.133-136，2005.

[2] Z. Zhang，Y. Nakashima and N. Naka：“Error Correction with High Practicality for Mobile Phone Speech Recognition”，in Proc. International Workshop of Mobile HCI 2008.

[3] K. Itou，M. Yamamoto，K. Takeda，T. Takezawa，T. Matsuoka，T. Kobayashi，K. Shikano and S. Itahashi：“JNAS：Japanese Speech Corpus for Large Vocabulary Continuous Speech Recognition Research，”J. Acoust. Soc. Jpn.(E)，Vol.20，No.3，pp.190-206，1999.

[4] A. Lee，T. Kawahara and K. Shikano：“Julius --- an Open Source Real-Time Large Vocabulary Recognition Engine，”in Proc. EUROSPEECH，pp.1691-1694，2001.

[5] 武田 一哉，峯松 信明，伊藤 彰則，伊藤 克亘，宇津呂 武仁，河原 達也，小林 哲則，清水 徹，田本 真詞，荒井 和博，山本 幹雄，竹沢 寿幸，松岡 達雄，鹿野 清宏：“大語彙日本語連続音声認識研究基盤の整備 —汎用音素モデルの作成—，”情処学研報，97-SLP-18-3，1997.

＊11 **形態素解析**：文を構成する最小の意味単位である形態素の列に分割する作業．