連続音声認識による読影レポート作成の試み

高原　太郎　　中島　美佳　　似鳥　俊明　　蜂屋　順一

杏林大学医学部放射線医学教室

Japanese Radiological Report Creation with Continuous Speech Recognition

Taro Takahara, Mika Nakajima,
Toshiaki Nitatori, and Junichi Hachiya

Ten Japanese radiological reports consisting of 1381 characters(681 words)were created by two board-certified radiologists who used conventional typing and a continuous speech-recognition system called AmiVoice(Advanced Media, Inc., Tokyo, Japan). The two radiologists had not had any special training prior to their use of the continuous speech-recognition system. The model of speech-to-text analysis was generated from 22,589 radiological reports(5.7MB). Dedicated pronunciations for loan words(i.e., English words)were registered by a board-certified radiologist in consideration of variations in Japanese pronunciation. Misrecognition occurred in 40 of 1362 words, corresponding to a 97.1% rate of accuracy of recognition. The average speech recognition time per report was 31.3 sec, and the additional time required for corrections was 25.0 sec. The total speech input time of 56.2 sec was much less than the conventional input time of 142.8 sec for typing. Continuous speech recognition is faster than typing, even considering the additional time required for corrections, and is acceptable in view of the overall reduction in report turn-around time.

Research Code No.: *220.2*

Key words: *Continuous speech recognition, Radiology, Dictation, Report creation*

Received Dec, 6, 2001; revision accepted, Dec 28, 2001
Department of Radiology, Kyorin University

別刷請求先
〒181-8611　東京都三鷹市新川 6丁目20番 2 号
杏林大学医学部放射線医学教室
高原　太郎

はじめに

読影レポートの作成は，放射線科における診断業務の中核をなす作業である．従来は手書きで行われてきたが，パーソナルコンピュータの発達や，DICOM Viewerなどの読影端末が普及してきたことに伴い，レポートもデジタル化する必要性が高まってきている．読影レポートのデジタル化において最も時間と経費の負担がかかる部分は文字の入力である．時間を節約するために欧米で用いられてきた方法には，トランスクライバーの導入が挙げられる．また，経費を節約するための方法としては各診断医が自分でタイプ入力をする方法が採られてきた．前者は人件費が毎年固定して発生することに，また後者はレポート作成速度に劣る点や手間がかかることに難点がある．最近，コンピュータによる音声認識技術が発達してきており，これを用いることで人件費の発生しないシステムを構築出来る可能性がある[1)]．今回われわれは，連続音声認識を入力手段とした読影レポートの作成を試み，その認識率と入力速度をタイプ入力と比較検討し，有用性を評価した．

方　法

1)ソフトウエア

連続音声認識システムとしては，アドバンスト・メディア社の「AmiVoice」を用いた．このソフトウエアの特徴は不特定話者タイプであることである．従って，初めて使用する際の，いわゆる「ならし」(enrollment)は必要がない．使用したコンピュータは，Pentimum Ⅲ 500MHzのCPU，256MBのメモリを搭載したノートパソコンである．入力マイクはコード部分にON/OFFスイッチの装備されたものを用い，発話時にはONに，また発話終了時にOFFにした．

2)言語解析モデル

連続音声認識システムが使用する辞書(言語解析モデル)は，当科で１年余に渡り蓄積された，主にCTとMRIのレポートから成る22,589件のデータ(文書量5.7MB)から，患者固有のデータ(IDおよび氏名)を除いたものを自動学習させて構築した．これらレポートは，当科に所属している放射

Table 1 Pronunciation registration considering variation of Japanese speaker.

Words	Pronunciation 1 original	Freq.	Pronunciation 2 by radiolotist	Pronunciation 3 by radiologist
ENHANCEMENT	えんはんすめんと	407		
ENHANCING	えんはんしんぐ	24		
ENHANCMENET		4	綴り間違い→削除	
ENHANCMENT		10	綴り間違い→削除	
ENHENCEMENT		6	綴り間違い→削除	
ENTERITIS		25	えんてらいてぃす	
ENTEROCOLITIS	えんてろこらいてぃす	2		
ENTRY	えんとりー	23		
EOSINOPHILIC	えおしのふぃりっく	7		
EPI	いーぴーあい	95		
EPIDERMOID	えぴでもいど	10		
EPIDURAL		12	えぴでゅーらる	
EPIGASTRALGIA		18	えぴがすとらるじあ	
EPIPLOIC	えぴぷろいっく	2		
EPISODE	えぴそーど	15		
EQUIVOCAL	えくいぼーかる	67	えくいぼーかる	いくいぼーかる
ERALY		15	綴り間違い→削除	
ERBD		3	いーあーるびーでぃー	
ERC		4	いーあーるしー	
ERCP	いーあーるしーぴー	16		
EROSION		36	えろーじょん	いろーじょん
EROSIVE		6	えろーじぶ	いろーじぶ
ERP	いーあーるぴー	1		
ERROR	えらー	3		
ESOPHAGEAL	いそふぁじーる	2	えそふぁじある	いそふぁじある

Pronunciation 1 original: Primary pronunciation registered by Advanced Media using an English dictionary.
Pronunciation 2 and 3 by radiologist: Pronunciation registered by a radiologist considering variation of Japanese speaker.
Freq,: Frequency in 22,589 radiologiral reports.

線科医がパーソナルコンピュータを用いてタイプ入力したもので，今回の連続音声認識による実験を行う以前に作成されたものであり，入力に際しての注意点やルールは特に設けず，自由に入力されたデータである．

3)外来語学習

認識率を向上させる目的で，約2,700語の外来語について，放射線科領域で普通に用いられる読みを，日本人の発音のバリエーションを考慮して設定した(Table 1)．例えば“erosion”は「えろーじょん」という読みで呼ばれる場合と，「いろーじょん」という読みで呼ばれる場合がある．このため，両者のいずれを発音しても“erosion”と認識されるように読み仮名を振った．さらに，自動学習させた外来語には，読影者が誤った綴りで入力したものが含まれている．これらについてはリストから削除を行った．

4)入力対象

入力対象としたレポートは，辞書学習には使用していない新しいレポート10件を用いた．いずれもMRIに関するもので，合計の文字数は1,361文字(単語数689語)である．内訳は脳神経領域5件，腹部領域5件である．代表的なものを以下に示す

■脳神経領域レポート

「2000年5月の前回MRI所見と比較した．

左内頸動脈後交通動脈分岐部動脈瘤クリッピング術後．左島皮質や両側放線冠から半卵円中心に前回同様梗塞巣がみられる．新たな梗塞巣や脳室拡大の出現はない．前回に続いて左内頸動脈後交通動脈分岐部動脈瘤クリッピング術後の変化がある．描出範囲内に明らかな限局性の狭窄性変化や動脈瘤はみられない．」

■腹部領域レポート

「前回MRI所見(2000年3月27日)と比較しました．

肝門部の総胆管癌に関しては著変ありません．(造影後脂肪抑制T1強調画像でIVCからのアーチファクトがきわめて病変に似て見えますのでご注意下さい．)今回の検査では左水腎症を生じています．これはおそらく後腹膜リンパ節への転移によるものと思われます．」

Table 2　Input speed of continuous speech recognition versus conventional typing

	Recognition (seconds)	Check (seconds)	Correction (seconds)	Speech (seconds)	CPM(S) (characters)	Typing (seconds)	CPM(T) (characters)	CPM(S) CPM(T)
Radiologist A	35.1	11.0	12.8	58.9	138.6	82.2	99.3	1.40
Radiologist B	27.4	12.1	14.0	53.5	152.6	203.5	40.1	3.81
Average	31.3	11.6	13.4	56.2	145.6	142.9	73.7	1.98

Recognition: Average speech recognition time per report. Check: Average check time for detection of mis-recognition per report. Correction: Average correction time for mis-recognition per report. Speech: Average overall speech input time per report. CPM(S): Character per minutes in speech recognition. Typing: Averate typing time per report. CPM(T): Character per minutes in typing. CPM(S)/CPM(T): Relative speed ratio in CPM.

5)入力方法

入力(読影)者は放射線科専門医 2 名である．それぞれのタイプ速度は結果に示すが，各々ワープロ検定で 1 級および 3 級に相当する速度である．

音声入力は，印刷されたレポートを発話入力し，その所要時間を「発話入力時間」とした．発話時にはリアルタイムで入力・表示されている様子は見えないようにした．実際の読影業務においては，入力結果が正しいかどうかを検証して修正する時間も必要である．そこで，入力結果を確認する時間(確認時間)と修正する時間(修正時間)をそれぞれ測定した．これは，他者によりワープロ上にコピーされたものを，ワープロ上で確認，修正することにより行った．これらの合計を「音声入力時間」とした．一方タイプ入力においては，印刷されたレポートをタイプ入力し，その所要時間を「タイプ入力時間」とした．タイプ入力では，タイプ中にミスを生じるが，これを修正して完全な文章になった時間を測定した．

6)評価

以上のようにして計測した時間を連続音声認識とタイプ入力で比較した．さらに，音声認識においては，単語単位での誤りの数を基に認識率を計算した．今回用いたシステムは，英単語で始まる文の文頭を大文字にする機能が装備されてない．このため，大文字にならない部分は誤りとは判定せず，文字列が合っているかどうか(つまり音声を正しい文字列として認識できるかどうか)を判断基準とした．

結　果

1)発話入力時間および認識率

読影者Aの発話入力時間の合計は351秒であった．従って 1 文字および 1 単語あたりの速度はそれぞれ0.258秒および0.509秒であった．誤認識単語は689語中24語であった．従って認識率は96.5%であった．読影者Bの発話入力時間の合計は274秒であった．従って 1 文字および 1 単語あたりの速度はそれぞれ0.201秒および0.398秒であった．誤認識単語は，689語中16語であった．従って認識率は97.7%であった．A，B 2 者の平均認識率は97.1%と高率であった．

2)入力速度

各読影者における入力速度をTable 2に示す．A，B 2 者の平均発話入力時間は31.3秒，平均修正時間は11.6秒，平均確認時間は13.4秒であった．平均音声入力時間は56.2秒であった．これに比べ平均タイプ入力時間は142.8秒であった．

各々の結果を 1 分当たりの文字数(character per minutes, 以下CPM)で表示すると，A，B 2 者の平均タイプ入力速度は69.8CPMであるのに対し，音声入力速度は145.6CPMと1.68倍高速であった．読影者Aはタイプ入力が高速で，99.3CPMであった．これはワープロ検定 1 級(80CPM以上)に相当する．しかしAにおいても音声入力は1.40倍高速だった．また読影者Bのタイプ入力速度は40.1CPMで，これはワープロ検定 3 級に相当し，平均的なものと考えられるが，この場合音声入力は3.81倍高速だった．

考　察

・連続音声認識の特徴と認識率：

連続音声認識は，単独の音声を変換するのではなく，前後の文脈(通常 3 語)の連携の特徴からもっとも出現確率の高い単語を予測する方法である[2]．たとえば，「右_上肺野_に_腫瘍_を_認める．」という文章と，「この_方法_は_主要_な_もの_で_ある．」という 2 つの文章に出現する「しゅよう」は前後の文脈の特徴により前者では「腫瘍」，後者では「主要」であることを出現確率から推定する．このため，従来の単語ベースの変換に比し，理論的にははるかに高い変換精度が得られる．しかし連続する単語群の組み合わせ頻度を学習するには膨大なデータの蓄積を必要とする．たとえば，新聞記事読上げの音声認識用の言語解析モデルを作成するためには，学習テキスト量として数百MB必要であるとされている．このため従来は読影レポートの学習には同様の大きなデータ量が必要と考えられていた．

一方，今回われわれが用いた読影モデルは5.7MBと極めて小さい．しかし認識率は97.1%と高率であった．これは，読影レポートが，語彙数(専門用語の数)は多いものの，文章の構造は比較的単純であることと関係すると考えられる．読影においては「(場所)に(性状)な(構造)が(個数)認められる．」といったような文が多数を占める．したがって各施設毎の語彙と，文脈の前後にでてくる組み合わせを学習させることにより，比較的小さなデータにおいても高

い認識率を有するシステムを構築できる可能性を示唆するものと思われる．

・英単語の日本語発音に関するバリエーションの考慮：

一方，今回の実験においては，単語レベルの「読み」に関しても工夫を行った．日本人の読影者による英単語の日本語発音にはかなりのバリエーションが存在するため，単に英語辞書を参考にした「英語読み」を用意しても良い結果が得られないと考えたためである．例えば，“capsular”という単語を，「きゃぷすらー」と読み仮名を振ったとしても，その他のバリエーション，例えば「かぷすらー」，「かぷしゅらー」，「きゃぷしゅらー」などに対しては正確な変換が期待できないと考えられる．

今回は2,700語について個別に登録を行ったが，これにより認識率がさらに向上したものと思われる．さらに多くの単語読みを登録することでより信頼性の高いモデルが構築可能と思われる．

・認識率と修正時間

小野木[3]らは，いちはやく連続音声認識を用いた実験を行い先駆的な結果を残しているが，同法の問題点として，発話入力時間が極めて短い反面，修正時間が長くかかることを挙げている．すなわち，40語の文章の入力を10回試行した平均値において，発話入力時間は19秒と高速であったが，確認・修正時間は48秒を要し，音声入力時間は67秒であった．これはキーボード入力の60秒よりも遅かった．このため修正方法が改善されればさらに高速になるであろうと推察している．このときの認識率は90％であった．しかしわれわれの実験結果では，修正に要する時間を考慮してもなおタイプ入力より高速であった．今回われわれの用いたシステムの認識率は97.1％であった．これは誤変換率に直すと，2.9％ということになる．一方認識率90％のシステムでは誤変換率は10％である．従って，誤変換率は1/3以下に減少したことになる．今回のわれわれのシステムでは，修正方法は通常のワープロ上で行ったが，このような通常の修正方法であっても，誤認識率さえ低ければ修正の必要がなくなり，結果として入力速度が大幅に向上することを示唆するものと思われる．また，今後，音声による修正や，次変換候補の自動表示ができれば，修正時間自体も短縮される可能性はあると思われる．

・実際の読影業務との関連

連続音声認識は，「連続」して認識することで高い変換効率を得る方法である．一方実際の読影では所見を考えながら入力するため，言い澱みなどの発生により認識率が下がる懸念がある．しかし，連続音声認識による代筆機能自体は，現時点の性能ですでにタイプ入力を遙かに凌駕している．このため今後，誤変換文字の修正方法などが改善されることなどにより，急速に臨床現場に普及し得る潜在的な能力があるものと思われる．

トランスクライバーによる入力に比較すると，連続音声認識の97.1％という認識率はまだ劣っていると思われる．また，AmiVoiceに関して言えば，現在のところ，レポーティングシステムの一機能として販売されているため，連続音声認識単体での価格を算出することができない．このためトランスクライバーによるシステムと比較して所要時間・経費の点で優れているかどうかについては結論できない．しかしながら連続音声認識には1)人件費のような固定のコストが発生しないこと，および2)読影医が直接修正を行うため，入力・修正後にはリアルタイムでレポートが完成しており，トランスクライバーシステムに存在する録音(読影者)→運搬→タイプ(トランスクライバー)にかかるタイムラグがないこと，の2点の特長がある．また今回用いた連続音声認識システムは不特定話者タイプであり，読影者を選ばない．これらを考慮すると，すでにトランスクライバーシステムに比肩すべきレベルに達しているものと思われる．また，トランスクライバーがおらず，読影医本人がパーソナルコンピュータを用いてタイプ入力をする病院や，遠隔診断における在宅労働においては，連続音声認識を導入することにより大幅な所要時間の削減が可能であると思われる．

結　語

連続音声認識による入力速度は，現時点の性能ですでにタイプ入力より高速である．従って，実際の読影業務への応用が可能な技術と思われる．

文　献

1) Ernst R, Carpenter W, Torres W, et al: Combining speech recognition software with Digital Imaging and Communications in Medicine(DICOM)workstation software on a Microsoft Windows platform. J Digit Imaging. 14(2 Suppl 1): 182–183., 2001

2) Rose RC, Juang BH: Hidden Markov models for speech and signal recognition. Electroencephalogr Clin Neurophysiol Suppl. 45: 137–152, 1996

3) 小野木雄三：連続音声認識を利用したレポートシステム．新医療 28: 109–111, 2001