2-2-3　人間の能力を超えた音声認識
2-2-3　Automatic Speech Recognition Beyond Human Capabilities

加藤 宏明　河井 恒　水上 悦雄
KATO Hiroaki, KAWAI Hisashi, and MIZUKAMI Etsuo

本稿では、機械による自動音声認識（ASR）性能の現状を人の音声書き起こし能力との比較において示す。ASR 評価用の日本語音声データセットを対象として、書き起こしの能力を、速さと正確さの両面で競わせた。人の書き起こしの単語精度を 1 ストロークごとに逐次モニタするツールを作成して精密かつ公平な評価を行った。速さでは、ASR は人の 2 倍あるいはそれ以上の性能であり、機械が人を大きく上回った。正確さでは、両者ともに平均単語精度が 97%を超えたが、ASR が誤認識する少数例にも正答する人の能力がまだわずかに機械を上回る結果となった。

In this paper, the current state of automatic speech recognition (ASR) performance was compared with human speech transcription capabilities. Both speed and accuracy were tested with a Japanese speech dataset for ASR evaluation. A tool was created to sequentially monitor the word accuracy of human transcription stroke by stroke, to ensure a precise and fair evaluation. In terms of speed, the ASR was more than twice as fast as the human transcribers; the machines greatly outperformed the humans. In accuracy, both had an average word accuracy of more than 97%, but the humans still slightly outperformed the machines, thanks to a few correctly answered instances that ASR misrecognized.

1　まえがき

自動音声認識（ASR: Automatic Speech Recognition）は音声を文字へ変換する技術であり、一般に速く正確であることが求められる。ASR の性能は、他章で述べるように、音声言語コーパスの量と質に大きく依存する。用途や環境により性能に対する要求水準が異なるため、絶対的な達成目標を決めることはできないが、『人の代わりが務まるかどうか』は一つの目安となる。どの程度役立つかも直感的に理解しやすい。そこで、ASR の現時点での到達点を正確に知り、更なる性能向上のためにコーパス構築に何が必要かを探るため、人の能力との精密な比較を試みた。

同様の比較は従来から試みられ [1]–[3]、最近では人と同等の性能に達したとの報告 [2][3] もあるが、未解明のポイントが少なくとも 2 つあった。まず、言語への依存性が不明であった。従来の対象は英語音声であり、英語以外の言語、特に、英語とは書字システムが大きく異なる日本語のような言語でも同様であるかどうかは未解明であった。次に、「人の能力」の定義が曖昧であり、機械と人との公平な比較になっていない可能性があった。従来報告では、性能を書き起こしの正確さのみで評価したが、速さも同様に重要である。機械が音声を 1 回だけ「聞いて」素早く文字へ起こすのに対し、人は繰り返し聞いて正解へ近づく。機械と同じ 1 回再生のスピード勝負で人がどこまで競えるかは未解明であった。以上を考慮した「厳正な競争」の結果を報告する。

なお、本稿は既発表の成果をもとに本研究報告読者向けに要点を絞ってまとめたものである。方法、結果等の詳細は別稿 [4] を参照されたい。

2　方法

2.1　音声素材

書き起こし対象となる音声には、NICT の多言語音声認識評価データセット（SPREDS2 : SPeech Recognition Evaluation Data Set 2）[5] の日本語セットを用いた。これは、全 1,000 発話から成る読み上げ音声のデータセットである。話者は、特に発話訓練を受けていない日本語を母語とする 15 歳から 60 歳の成人男女各 10 人（合計 20 人）であった。複数の話者間で原稿の重複があったため、ユニークなセットに原稿を絞った。書き起こし者の原稿への親密度を統制するためである。

表 1　音声素材の諸元

名称	SPREDS2 ja（一部）
発話タイプ	原稿読み上げ
収録環境	クリーン環境（SN 比 >15 dB）
収録場所	会議室
収録機材	iPhone 4/4 S/5/5 S/6（いずれか）
ドメイン	余暇，医療，防災，住居，消費，移動，情報通信、子育て，勤労，教養
デジタル化方式	標本化周波数 16 kHz 精度 16 ビット線形量子化

発話数	745
音声区間の合計時間長（時間）	0.95
発話ごとの平均･最大･最小時間長（秒）	4.6, 15.6, 1.3
同平均・最大・最小語数	14.8, 34, 4
同平均・最大・最小モーラ数	31.2, 74, 8

その結果、745 発話が残った。音声素材の諸元を表 1 に示す。

2.2　実験参加者

プロの書き起こし者 3 人と校正者 1 人が実験に参加した。いずれも日本語音声書き起こし作業者として 3 年以上の経験を有し、かつ実収録時間で 50 時間以上の日本語音声の書き起こし実績を持ち、その間作業の速度と正確さにおいて優秀と認められた者であり、実験実施時点でピーク時と同等の能力を維持していた。校正者は機械では原理的に生じ得ないタイプミスを修正する役割で加わった。従来報告の中では Saon ら [3] が採用した体制である。

2.3　実験準備：書き起こし制御・測定ツール

図 1 の左側に人の右側に機械の実験の流れを示す。人の側の実験の流れを制御し、必要なデータを測定・記録するために、専用のツールを作成した。このツールは、使用者の求めに応じて音声ファイルをあらかじめ決められた順序で再生・停止し、書き起こしの内容を記録するとともに、書き起こし時にタイプされたすべてのキーストロークの打鍵時刻を、音声の 1 回目の再生開始時点を起点とした経過時間としてミリ秒単位で記録した。加えて、使用者が校正者の場合には、音声の 1 回目の再生開始と同時に校正対象のテキストを修正可能な状態で表示した。

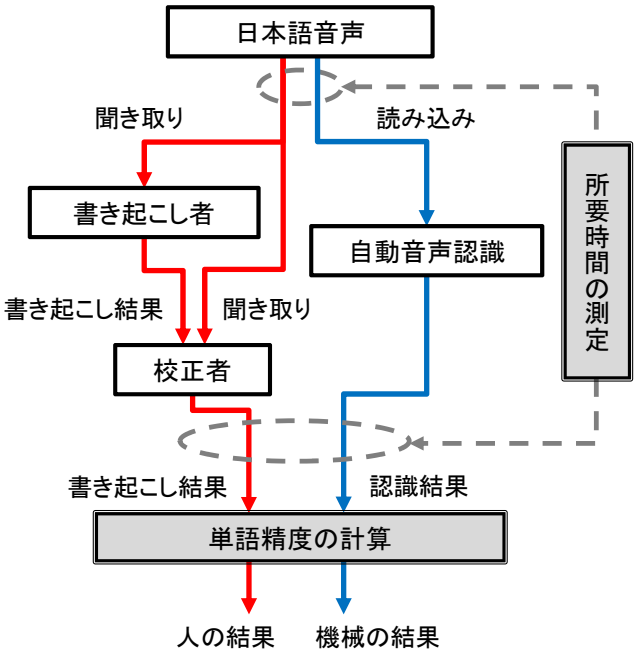

図 1　人と機械の文字起こし比較実験の流れ

2.4　実験実施

音声素材のうち、ランダムに選んだ 45 発話を練習用とし、残り 700 発話を本実験用とした。参加者は PC（HP, p6-2410 jp）上で動作する書き起こし制御・測定ツールにて書き起こしの練習を行い、ツールに十分慣れた後、1 時間以上の休憩をおいて本実験に臨んだ。ランダムな順序で再生される対象音声をヘッドホン（Sony, MDR-CD900 ST）で聞き、PC 付属のキーボードで書き起こした。日本語入力システムは OS（Windows 10 Pro）に組み込みの IME を用いた。IME の学習履歴は各実験セッション開始時に初期化され、自動修正機能は使われなかった。書き起こし者は各音声ファイルの 1 回目の再生で最大限書き起こした後、必要ならば 2 回以上聞き取り、間違いがないと判断した時点で終了した。3 人とも 700 発話を書き起こした。校正者は同じ方法で書き起こし者の結果を修正した。

2.5　評価指標

書き起こしの正確さの指標は下記の式で定義される単語精度とした。

単語精度 = 1 −（Sub + Del + Ins）/ NW

ここで、

NW: 発話の対象部分に含まれる語の数。

Sub: 置換。正解と異なる語の数。

Del: 削除。正解にあり結果にない語の数。

Ins: 挿入。正解になく結果にある語の数。

速さの指標は 1 回目の再生開始からの所要時間とした。所要時間は、対象音声の時間長の違いによる統計値への影響を吸収するため、音声の継続時間長で正規

化した単位(Real Time Factor、以下 RTF)で示す。たとえば、長さ 2 秒間の音声の書き起こしに 4 秒を要したとすると、所要時間は 2.0 RTF である。

2.6　機械による文字起こし

同じ音声素材を ASR システムに入力し、人の書き起こしの場合と同じ方法で各評価指標を得た。ASR システムには、VoiceTra® の音声認識部を用いた。VoiceTra の日本語音声認識は、2020 年時点で一般に利用可能な商用 API サービスとおおむね同等の認識精度であることを確認している。

機械における音声入力は、ASR システムの性能評価における標準的な方法を踏襲するため、音響的に一旦再生することなく、計算機上で音声ファイルをディスクから直接読み込んで行った。したがって、所要時間は、実際に再生された音声を読み込みつつ処理した場合と同じになるよう補正した。使用計算機の CPU は Intel® Xeon® Gold（6152 CPU、動作周波数 2.1 GHz）であった。

3　結果と考察

3.1　再生回数による結果の違い

人と機械の単語精度と所要時間との関係を図 2 に示す。700 音声の平均値である。人の場合は、さらに書き起こし者 3 人を平均した結果である。比較のため、図の横軸は最も長いものに合わせて延長した。これらの最終値を表 2 に示す。まず、1 回のみ再生の条件では、機械は人よりも明らかに早く処理を終え、平均的な速さは人の 2 倍以上であった。正確さにおいても、機械が処理を終えた時点では人は機械にはるかに及ばず、時間をかけても機械をしのぐことはなかった。次に、人に複数回の再生を許した条件では、音声の再生開始から約 2 RTF までの間は 1 回再生の結果に重なるが、その後、時間の経過とともに単語精度が上昇し、最終的にはすべての書き起こし者が機械を上回った。

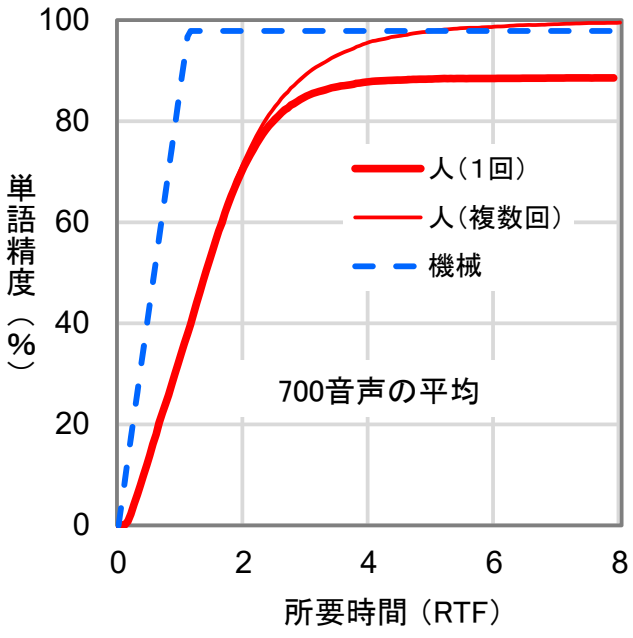

図 2　人と機械の文字起こし比較実験の結果

表 2　人と機械の文字起こし比較実験の結果

	単語精度（%）	所要時間（RTF）	平均再生回数
機械	97.88	1.11	1
人（1 回）	88.55	2.26	1
人（複数回）	99.71	4.42	2.46

表 3　人が正解した機械の誤認識例
（下線部で人と機械が異なる）

	人の書き起こし（正解）	機械の認識結果（誤認識）
①	送信者にユーザーが含まれている場合、オンにします。	～本にします。
②	カニですね、どの辺が、カニに見えるのですか？	～仮に見える～
③	まずは型名を教えてください。	まずは片目を～
④	根津には、東京メトロという地下鉄を利用して行きます。	ネズミは、～
⑤	プロテスタントでは、牧師と言うのですが、たぶんそうすると思います。	～ボックスと言うのですが～

3.2　機械が人に及ばない部分

時間とともに人と機械の単語精度が逆転した背景には、人は容易に排除できるが機械にとってはそうではない誤認識候補の存在があった。表 3 に例を示す。機械の誤認識部分は、いずれも音響的に正解に近く文法的にも正しいが、人ならばまずおかさない誤りであるようだ。①と②は発話の残りの部分との意味的な整合性からあり得ない候補である。③と④は拍数やアクセントといった韻律要素が違うので候補から除外される。⑤はその両方を含む。

4 あとがき

1回の再生のみの条件を公平な比較とみなせば、音声の文字起こし能力において、機械は速さと正確さの両面で既に人を大きく凌駕（りょうが）すると言って良い。しかし、聞き直しを許し、十分な時間を与えれば、人が最終的に優勢となった。機械が人に劣る部分は数値的にはわずかであるが、人がおかさないような誤りを含むので、コミュニケーションに及ぼす影響は小さくないと思われる。この残された部分は、開発者にはコーパス設計を含む今後の研究開発の課題を与えるとともに、利用者には、機械との付き合いでストレスをためないための知恵を与えることになる。

謝辞

人の書き起こし実験の準備、自動音声認識実験の実施、データ処理で協力いただいた研究室メンバーに感謝します。

【参考文献】

1 R. P. Lippmann, “Speech recognition by machines and humans,” Speech Communication, vol.22, Issue 1, pp.1–15, 1997.
2 W. Xiong, J. Droppo, X. Huang, F. Seide, M. Seltzer, A. Stolcke, D. Yu, and G. Zweig, “Achieving human parity in conversational speech recognition,” arXiv:1610.05256v2, 2017.
3 G. Saon, G. Kurata, T. Sercu, K. Audhkhasi, S. Thomas, D. Dimitriadis, X. Cui, B. Ramabhadran, M. Picheny, L.-L. Lim, B. Roomi, and P. Hall, “English conversational telephone speech recognition by humans and machines,” Proc. Interspeech, pp.132–136, 2017.
4 加藤 宏明, 河井 恒, 水上 悦雄, “人と機械の文字起こし能力比較：自動音声認識は人間を超えたか？,” 日本音響学会春季研究発表会講演論文集, pp.1373–1376, 2021.
5 https://ast-astrec.nict.go.jp/release/SPREDS2/download.html

加藤 宏明（かとう ひろあき）
ユニバーサルコミュニケーション研究所
先進的音声翻訳研究開発推進センター
先進的音声技術研究室
主任研究員
博士（工学）
音声言語コーパス、聴覚・音声コミュニケーション
【受賞歴】
1995年　日本音響学会 第12回粟屋潔学術奨励賞

河井 恒（かわい ひさし）
ユニバーサルコミュニケーション研究所
先進的音声翻訳研究開発推進センター
先進的音声技術研究室
室長
博士（工学）
音声情報処理
【受賞歴】
2015年　電気通信普及財団 第31回（2015年度）電気通信普及財団賞（テレコムシステム技術賞）
2014年　電子情報通信学会 2014年度論文賞
2010年　情報処理学会 2010年度喜安記念業績賞

水上 悦雄（みずかみ えつお）
ユニバーサルコミュニケーション研究所
先進的音声翻訳研究開発推進センター
先進的音声技術研究室
主任研究技術員
博士（理学）
音声言語コーパス、コミュニケーション科学、対話研究
【受賞歴】
2008年　社会言語科学会 2007年度徳川宗賢賞萌芽賞