Vol.2021-MUS-131 No.18
Vol.2021-SLP-137 No.18
2021/6/18

言い淀みラベル付けによる 非流暢発話の End-to-End 音声認識

堀井 こはる[1]　福田 芽衣子[2]　太田 健吾[3]　西村 良太[2]　北岡 教英[1]

概要：従来の ASR システムでは流暢な発話音声においては高い精度を発揮するが，高齢者音声等の非流暢発話では低くなってしまう．本研究では End-to-End 音声認識において，言い淀みをラベル付けして認識対象とすることによって，非流暢発話の精度がどう変化するか実験を行った．その結果，文誤り率はすべての評価データで改善し，モデルがラベルの意味を学習できていることが期待できる有効な結果を得られた．

キーワード：End-to-End 音声認識，高齢者音声，非流暢発話，言い淀み

End-to-End Speech Recognition of Non-fluent Speech by Hesitation Labeling

Koharu Horii[1]　Meiko Fukuda[2]　Kengo Ohta[3]　Ryota Nishimura[2]　Norihide Kitaoka[1]

Abstract: Conventional Automatic Speech Recognition (ASR) shows high accuracy for fluent speech, but low accuracy for non-fluent speech such as that of elderly people. In this study, we labeled hesitations and recognize these hesitations as well as characters using End-to-End (E2E) speech recognition and examine the accuracy for non-fluent speech. As a result, the sentence error rate was improved for all the evaluated data and it suggests that the model has possibility to capture the disfluency.

Keywords: End-to-End Speech Recognition, Elderly Speech, Non-fluent Utterance, Hesitation

1. はじめに

現在，自動音声認識（Automatic Speech Recognition：ASR）は広く普及しており，特にスマートフォンに搭載された Google アシスタントや Siri といった AI アシスタント等で非常に身近なものとなっている．ASR の研究は現在でも盛んに行われており，その精度は著しい向上を続けている．しかし，従来の ASR システムでは流暢な若年者の発話音声においては高い精度を発揮するが高齢者では低くなってしまう．高齢者の発話音声の ASR は介護・福祉の現場で非常に役立つことは明白であり，この精度を上げることは重要な課題である．

高齢者音声の特徴の一つに非流暢性がある．本研究では，高齢者音声や，高齢者以外でも自発的発話等言い淀みの多い非流暢発話を認識する際，非流暢性の特徴である言い淀みをラベル付けすることによって，モデルがその意味を学習できると仮定し，ASR の研究において最近の主流である End-to-End（E2E）型モデルのシステムにおいて，精度がどう変化するか実験を行った．

2. 関連研究

2.1 Transformer

今回の実験では Transformer で学習を行っている．過去に有力とされていた sequence-to-sequence（S2S）モデルはリカレントニューラルネットワーク（Recurrent Neural Network：RNN）で，これは複雑で逐次的な性質を持ち，

[1] 豊橋技術科学大学
Toyohashi University of Technology
[2] 徳島大学
Tokushima University
[3] 阿南工業高等専門学校
National Institute of Technology, Anan College

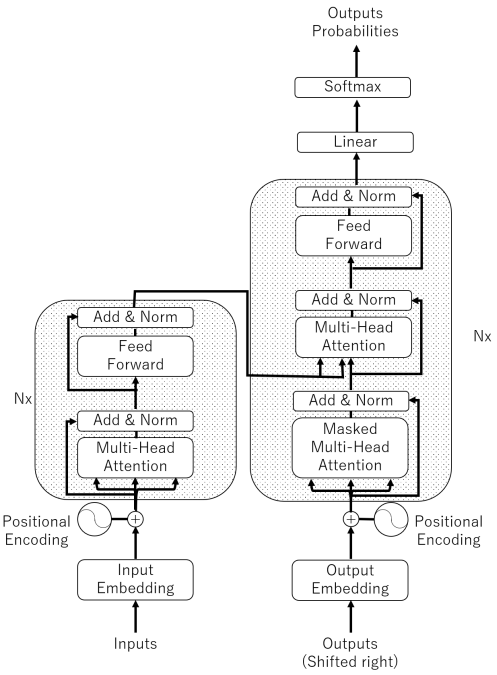

図 1 Transformer のモデル図
Fig. 1 The Transformer model architecture.

並列計算が不可能である．Transformer の場合，逐次性のない attention 機構のみに基づいているため，モデルは単純かつ並列化可能で計算が高速である [1]．また機械翻訳 [1]，ASR[2] タスクの実験で RNN より良い精度を達成している．Transformer の構造を**図 1** に示す．Transformer は encoder-decoder 構造を持ち，エンコーダは入力シーケンス $x_1,\cdots, x_n$ を連続シーケンス $z_1,\cdots, z_n$ に変換する．z が与えられると，デコーダは一要素ずつ記号の出力シーケンス $y_1,\cdots, y_m$ を生成する．エンコーダとデコーダの両方で，積層型の self-attention とそれぞれで全結合された feed-forward ネットワークを使用している．エンコーダは Multi-Head Attention（MHA）層と全結合型 feed-forward 層といった 2 つのサブレイヤーを持つ 6 層の同一レイヤーで構成されており，サブレイヤーそれぞれで残差結合とレイヤー正規化を行う．出力は

$$LayerNorm(x + Sublayer(x)) \tag{1}$$

で，$Sublayer(x)$ はサブレイヤー自身で実装された関数である．すべてのサブレイヤーは 512 次元の出力を行う．デコーダも同じく 6 層の同一レイヤーの積層で最後に残差結合とレイヤー正規化を行うが，デコーダでは二つのサブレイヤーに加え、エンコーダの出力に対し MHA を行う層を持つ．

self-attention は intra-attention とも呼ばれ、sequence 表現を計算するために、一つの sequence 中の異なるポジションを関連付ける attention 機構である．attention 関数は query と key-value のペアを出力にマッピングするもので，出力は value の加重和として計算する．ここで query，key，value はすべてベクトルである．各 value の重みは query と対応する key の互換性関数によって計算される．

embedding 層で、入力・出力トークンを 512 次元のベクトルに変換する．positional encoding は embedding を入力するための関数で，シーケンスの順序を利用するためシーケンス中のいくらかのトークンの位置情報を入力し，

$$PE_{(pos,2i)} = \sin(pos/10000^{2i/512}) \tag{2}$$

$$PE_{(pos,2i+1)} = \cos(pos/10000^{2i/512}) \tag{3}$$

といった出力を得る．ここで pos は位置，i は次元である．これにより，学習時に遭遇したものより長いシーケンスをモデルが推定できるようになる [1].

2.2 ESPnet2

ESPnet[3] は主に E2E の ASR に焦点を当てて開発されたオープンソースの音声処理ツールキットであり，柔軟なモデルの記述・拡張を実現する．レシピ（シェルスクリプトで書かれた実行ファイル）はディープニューラルネットワーク（Deep Neural Network：DNN）と隠れマルコフモデル（Hidden Markov Model：HMM）のハイブリッドモデルである DNN-HMM の音声処理ツールキット Kaldi の方式に基づいており，再現実験を行うために必要な全ての手順が一括で実行できるようになっている．最も ASR システムの評価に使われるタスクの一つである，Librispeech コーパスの ASR ベンチマークでは ESPnet は様々な ASR ツールキットの中でも高い性能を誇ることが証明されている [2].

ESPnet2[4] は ESPnet の弱点を克服するべく開発された次世代の音声処理ツールキットで，利便性と拡張性を高めるため，ESPnet から様々な拡張が行われている．ASR モデルは RNN，Transformer，Conformer の三種類から選べる．こちらが今後開発のメインになることが考えられるため，本研究では音声処理ツールキットとして，ESPnet2 を採用する．

2.3 高齢者音響モデルによる大語彙連続高齢者音声認識

2002 年に高齢者音声認識を新聞記事読み上げ音声コーパス（Japanese Newspaper Article Speech：JNAS，平均年齢 28.6 歳）で学習した HMM 音響モデルと大規模な高齢者（60〜91 歳）音声データベース（200 文章× 301 人）で学習した HMM 音響モデルを用いて音声認識実験を行い，その結果を比較する研究が行われている [5]．ここでは大語彙連続音声認識システム Julius[6] を用いて認識実験を行った．言語モデルは新聞記事から作成された語彙サイズ 2 万語の N-gram モデルで，評価データは学習用の話者とは異なる 46 人の高齢者話者による合計 200 文の新聞記事読み上げ文である．HMM はそれぞれモノフォン，トライフォンと，モノフォンモデルの各状態が持つ数十

個のガウス分布集合をトライフォンの対応する状態に割り当て，重みのみを変えて共有することで合成する PTM（Phonetic Tied Mixture）[7] というモデルの三種類，さらに性別非依存（Gender Independent：GI）モデルと性別依存（GenderDependent：GD) モデルを構築した．

実験の結果，PTM モデル（GI）で JNAS で学習したモデルと比較して 4.6%の単語認識率の改善が得られた．この結果から高齢者音声認識において高齢者音声で学習したモデルを使うことが有効であると示された．

2.4 海外における高齢者適応

ポルトガル語において ASR モデルを高齢者音声に適応する研究が行われている [8]．この研究では一般的なコーパスで学習されたベースラインモデルの音響モデルをポルトガルの Microsoft 言語開発センター（Microsoft Language DevelopmentCenter ：MLDC）の収集した高齢者音声コーパスで再学習することによって高齢者適応を行った．認識実験には HMM と多層パーセプトロン（MultiLayer Perceptrons：MLP）のハイブリッド型 ASR エンジン Audimus[9] を用いた．

ベースラインは音響モデルと言語モデルを用いており，音響モデルは、最初にポルトガルのテレビ番組から収集された 46 時間の書き起こし付きニュースデータを用いて教師あり学習し、2 回目に 1000 時間のポルトガル語ニュースで教師なし学習したもので，言語モデル（Language Model：LM）は，7 億語の新聞テキストコーパスで学習したバックオフ 4-gramLM と、約 53 万語のニュースの書き起こしコーパスで推定したバックオフ 3-gram LM を含む，いくつかの LM により推測された 4-gram LM である．高齢者コーパスは，60〜100 歳の 1,038 人の話者が，数字のみから長い文章まで様々なテキストからそれぞれ平均 12 分 160 文を読み上げた，約 150 時間の読み上げ音声から構成されており，60-65 歳の話者が最も多く、64 時間分を占める．評価用データは、コーパス中の約 10%の発話をランダムに選択したもので，話者オープンである．

ベースラインの単語誤り率（Word Error Rate：WER）は評価データ全体で 35.3%，年齢別では 60〜65 歳の話者で 29.1%，66〜70 歳で 28.1%，71〜75 歳で 36.1%，76〜80 歳で 45.1%，81〜85 歳で 41.0%，86〜90 歳で 54.9%と話者の年齢が高くなるにつれて増加した．60〜65 歳から 80〜85 歳までの 5 段階の年齢層でそれぞれ 6 時間のデータで，86〜90 歳ではデータが少ないため，2 時間のデータで適応モデルを構築した。その結果，全体の WER は 60〜65 歳のデータによる適応モデルでは 31.5%，66〜70 歳モデルで 31.4%，71〜75 歳モデルで 31.1%，76〜80 歳モデルで 30.0%，81〜85 歳モデルで 30.0%，86〜90 歳モデルで 33.4%とすべてのモデルでベースラインと比較して改善し，試験データの話者年齢に合わせた適応モデルを使用することで，改善につながることが確認できた．

2.5 高齢者音声の非流暢性

高齢者における認知症の割合は 2012 年時点で 65 歳以上で 15%，85 歳以上では 50% を超えると推計されている [10]．よって認知症に多いとされている特徴は高齢者の音響的特徴とも捉えられる．

認知症は主に 4 つのタイプに分類でき，認知症のうち 44% がアルツハイマー型認知症（Alzheimer-Type Dementia：ATD），21% がレビー小体型認知症（Dementia with Lewy bodied：LDB），15% が前頭側頭葉変性症（Fronto Temporal Lobar Degeneration：FLTD），10% が脳血管性認知症（Vascular Dementia：VaD）であると言われている．最も多い ATD の，米国国立老化研究所・アルツハイマー協会（National Institute on Aging-Alzheimer's Association workgroups：NIA-AA）による診断基準の一つに言語機能の障害があるという項目がある．具体的には，発話の途中で一般的な単語を思いつかない，ためらう，発話，つづり，書字を誤るという障害である．また，FTLD では言語障害が現れやすく，その一種である進行性非流暢性失語では，主な症状に吃音がある [11]．その他のタイプでも，認知症全体の中核症状として失語があり，失語もいくつかの種類があるが，脳血管性認知症に多いとされている運動性失語の最大の特徴に非流暢性発話がある [11], [12]．

3. 提案手法

2.5 節で示したように，高齢者には認知症が多く，その特徴として発話の非流暢性がある．今回認識する超高齢者音声にも吃音等，その特徴が見られ，また，実際に認知症傾向のある話者が含まれている [13]．吃音の症状として認められるのは連発，引き伸ばし，難発 [14] であるが，本研究では，言葉の一部の音を引き伸ばす引き伸ばしや，言葉がなかなか出ない，語中の途切れといった症状である難発については考慮せず，吃音の最もメジャーな症状の一つである連発と，吃音には含まれないが，吃音のある人にしばしば見られる特徴である語や句の言い直し [14] のみに着目し，これらを「言い淀み」として扱う．

言い淀みは高齢者音声に顕著に含まれるが，高齢者以外でも吃音のある人や，自発的発話では発生しやすく，また，これらは文字に起こしたデータを利用する場合に必要のない情報である．言い淀みをラベルとして付加して，その学習に成功した場合，モデルが言い淀みという音響的特徴を理解すると共に，ASR においてノイズである情報の学習を避けられるため，高齢者音声やそれ以外の ASR でも高精度化が期待できる．

「私は」というフレーズを例に三つの特徴についての説明とそれが現れた時の発話を**表 1** に示す．

これらを一種類の言い淀みラベル「@」としてラベル付

表 1　言い淀みの具体例
Table 1　Concrete Example of hesitation.

特徴	説明	発話例
連発	不随意に同じ単語もしくは単語の一部分を繰り返す，代表的な吃音の症状.	「わわ私は」「私私は」
言い淀み	同じ単語もしくは単語の一部分を二度繰り返していて，連発と異なり，一度目の発音の直後にショートポーズや無音区間が入る.	「わ…私は」「私…私は」
言い直し	語句を読み間違えた後に言い直す．読み間違えているためラベル部分は正解とは発音が異なる.	「わし私は」

けを行う．今回認識したデータは文の読み上げ音声である．実際の読み上げる見本となった文，実際の発音，ラベルを付けた文，ラベルを付けた文からラベルを削除した文の例を以下に示す．ラベルを付けた文からラベルを削除した文は実際の発音よりも読み上げ見本と近いものとなっている．

読み上げ見本： 自分の 好きな ジャズピアニストの 演奏への 震えるような 感動を 綴る 人も いた
実際の発音： 「自分の好きなジャ ジャム ジャズピアニストの 演奏への 震えるような 感動を 綴る人もいる」
ラベル付加文： 自分の好きな@@ ジャズピア@ニストの 演奏への 震えるような 感動を 綴る人もいる
ラベル削除文： 自分の好きな ジャズピアニストの 演奏への 震えるような 感動を 綴る人もいる

読み上げ見本： 半分近い 六業者が 密輸品と 見られる 鯨肉 の 売買を 持ちかけられた
実際の発音： 「半分近い六業者が 密いむん密輸品とみ見られる 鯨肉の 売買を持ちかけられた」
ラベル付加文： 半分近い六業者が @密輸品と@ 見られる 鯨肉の 売買を持ちかけられた
ラベル削除文： 半分近い六業者が 密輸品と 見られる 鯨肉の 売買を持ちかけられた

読み上げ見本： 今度は 逆に 黒い 部分に 注目してみる
実際の発音： 「今度は 逆に く くーろい 部分に 注目してみる」
ラベル付加文： 今度は 逆に @ くーろい 部分に 注目してみる
ラベル削除文： 今度は 逆に くーろい 部分に 注目してみる

4. 実験に使用したデータ

実験には日本語話し言葉コーパス（Corpus of Spontaneous Japanese：CSJ），新聞記事読み上げ音声コーパス（Japanese Newspaper Article Speech：JNAS），新聞記事読み上げ高齢者音声コーパス（S-JNAS）といったよく使われる日本語大規模コーパスと，独自のデータセットである超高齢者音声コーパス（The Elderly Adults Read Speech Corpus：EARS）を用いた．ここでは各データセットについて紹介する．

4.1 CSJ[15]

約 700 万語，約 661 時間分の発話音声とその書き起こし，実験用の様々な付加情報が収録されている．収録音声は，理工学，人文学，社会学の学会講演や日常的話題についての模擬講演を中心とした独話音声を中心としており，その他に朗読や対話音声も収録している．話者数は 1,417 人と多いが 80 歳以上の話者は 17 人と少ない．

4.2 JNAS[16]

306 人（男女各 153 人）の話者が 1991 年から 1994 年の毎日新聞新聞記事の抜粋約 100 文と，ATR 音素バランス 503 文約 50 文の計約 150 文を読み上げた音声約 60 時間文とその書き起こしが含まれている．新聞記事は各約 100 文 155 セット，音素バランス文は各約 50 文 10 セットからなる．年代が明確なのは 10 代 1 人，20 代 159 人，30 代 83 人，40 代 27 人，50 代 10 人，60 歳以上 8 人と比較的若い話者が多い．

4.3 S-JNAS [17]

60〜90 代の話者を対象に，133.4 時間分の音声とその書き起こしを収録している．音韻モデル作成用は 301 人 (男性 151 人・女性 150 人) が各話者新聞記事文 1 セットと音素バランス文 2 セットの合計 200 文を読み上げたデータで，音声認識実験用はタスク文 100 文と新聞記事 100 文の 200 文を 101 人 (男性 51 人・女性 50 人) が読み上げたデータである．モデル作成用データの話者の平均年齢は 67.6 歳，実験用では 65.5 歳，全体では 67.0 歳である．60 代の話者が多く 80 歳以上の話者は 8 人である．

4.4 EARS[13]

EARS は高齢者音声認識の精度向上を目的として，高齢者の中でもより高齢な人の音声を録音・収集したコーパスである．現在名古屋・徳島・木更津・鈴鹿で録音した 120 人の ATR 音素バランス 503 文，毎日新聞記事の読み上げ音声とその書き起こしを収録している．訓練用データセットは音素バランス文各約 50 文 119 人の計 5,981 文，評価

用データセットは新聞記事各 10 文 22 人の計 220 文あり，収録時間は約 13.4 時間である．評価用データセットの話者は一人を除き，訓練用データセットの録音にも参加している．このコーパスの一番の特徴は話者の年齢の高さである．同じように高齢者音声を対象としている S-JNAS ですら平均年齢は 67.0 歳だが，EARS の平均年齢は今回実験に使用したバージョンでは，訓練用データセットで 83.2 歳，評価用データセットで 83.5 歳，全体で 83.3 歳となっている．80 歳以上の話者が 55 人と，小規模なコーパスではあるが高齢者の話者数が多い．**図 2** に EARS と S-JNAS（音韻モデル作成用）の年齢分布を示す．

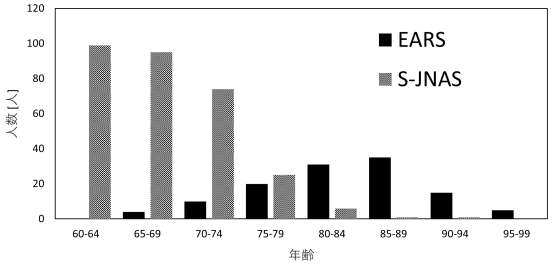

図 2 EARS と S-JNAS の年齢構成
Fig. 2 Age Structure of EARS and S-JNAS.

S-JNAS は 6，70 代の話者が多いが 80 代以上になると EARS の方が人数が多くなっている．同じ高齢者音声コーパスでも，EARS がより高齢者を対象としていることがわかる．また、各話者に対し HDS-R（Hasegawa's Dementia Scale for Revised）による認知症簡易テストを行っており，話者の中には認知症傾向のある者も含まれる．

5. 音声認識実験

5.1 実験条件

NVIDIA 社の GPU，GeForce RTX 3090 を 1 基備えたマシンで，ESPnet2（バージョン：v.0.9.9）を用いて実験を行った．

ベースラインのモデルは ESPnet2 の Joint CTC-Attention Transformer ASR モデルを CSJ, JNAS, S-JNAS で学習したもので，CSJ は ESPnet2 の CSJ 用の ASR レシピで自動的にデータを学習用，評価用，validation 用に分け，JNAS も同様に ESPnet のレシピで自動的に分けた．JNAS の評価用データのうち，語彙サイズの大きい評価用セット 500 も今回学習用データに含む．S-JNAS では卓上型マイクで録音された音韻モデル作成用のデータを学習データとして用い，その一部を手動で validation データとした．ベースラインの学習データ数は 497,680 発話，validation データは 7,321 発話である．学習は 20epoch 繰り返し，その中で validation の acculacy が高い 10epoch の各 epoch 時点でのモデルを平均したモデルを ASR に用いた．

学習・認識時のパラメータは ESPnet2 の初期値のままである．ESPnet2 では言語モデルを用いて shallow fusion[18] を行うことも可能であり，shallow fusion によって精度がよくなることもある [19] が，本実験では言語モデルを使わずに ASR モデルのみで認識を行った．認識・評価に用いたデータは CSJ の評価用データセットと EARS の評価用データセットである．超高齢者音声である EARS だけでなく，データ量が多く言い淀みを含む自発的発話音声である CSJ も評価に用いることで，ラベルの有効性を確認できるソースが増えると共に，高齢者音声といった特殊な音声だけでなく，一般的な自発的発話でのラベルの有効性を評価することが可能となる．CSJ の評価用データは話者 10 人ごとに eval1，eval2，eval3 の 3 つのディレクトリに分け，EARS の評価用データを eval4 とした．データ数は eval1 が 1,272，eval2 が 1,292，eval3 が 1,385，eval4 が 220 発話で全部で 4,169 発話である．CSJ の評価用データは話者オープンである．

各データの正解の書き起こし文は，複数のコーパスを使用していることで，数値の表記法がアラビア数字，漢数字，その中でも位取り記数法とそれ以外のものが混在していたため，すべて CSJ のスタイルに近い表記法に正規化した．

5.2 実験方法

2.3 節と 2.4 節で示されたように，高齢者音声認識を行うには高齢者音声で学習したモデルが有効であるため，最初にベースラインの学習データに EARS の訓練用データを加え，モデルの学習を行った．訓練用データの一部は手動で validation データに分類し，この時点の学習データは 503,207 発話，validation データは 7,774 発話で，EARS の評価用データセットの話者は訓練用データセットの話者と重複するため，EARS の評価データは一人を除き話者クローズである．

次に，EARS を含む学習データに対し，本研究で提案する言い淀みラベル付けを行った．ラベル付けは CSJ と EARS に対して行い，EARS では独自にラベル付けをし，CSJ は元々コーパス中の SDB ファイルで言い淀みとラベル付けしてある部分を CSJ レシピのデータ整形用スクリプトを改変することで言い淀みラベルに変換した．言い淀みラベルの数を**表 2** に示す．

学習，validation データ中には 1 割程度，評価データ中には eval3 を除き 2 割程度の言い淀みラベル付き文が含まれている．学習されたモデルで認識を行う際，学習データがラベル付きの場合は認識結果もラベル付きになる．評価時の正解文はラベル付きの文からラベル部分を削除した文とし，認識結果がラベル付きの場合ラベル部分を削除してから評価を行った．認識結果と正解文から文字誤り率（Character Error Rate：CER）と文誤り率（Sentence

表 2　データ中の言い淀みラベルの数
Table 2　Number of the hesitation labels in the data.

データ	個数 [個]	文数(LS)[文]	文全体の数(S)[文]	LS/S[%]
train	85,184	65,087	503,207	12.93
validation	1,238	938	7,774	12.07
eval1	321	260	1,272	20.44
eval2	303	234	1,292	18.11
eval3	140	128	1,385	9.24
eval4	52	42	220	19.09

Error Rate：SER）を算出し，ベースライン・ラベルなし・ラベル付きで比較し，ラベル付きデータに関しては実際の認識結果の観察を行った．

5.3　実験結果

実験結果は**表 3** のようになった．CER，SER 共に正解率ではなく誤り率であるので数値は低い方が精度が良い．

表 3　実験結果
Table 3　Experimental result.

	評価データ	CER [%]	SER [%]
ベースライン	eval1（CSJ）	7.3	61.6
	eval2（CSJ）	5.0	57.4
	eval3（CSJ）	6.3	45.3
	eval4（EARS）	12.5	74.5
ラベルなし	eval1（CSJ）	7.5	61.7
	eval2（CSJ）	4.9	56.5
	eval3（CSJ）	6.3	45.6
	eval4（EARS）	**10.9**	71.8
ラベルあり	eval1（CSJ）	**7.1**	**59.7**
	eval2（CSJ）	**4.8**	**55.0**
	eval3（CSJ）	**6.1**	**43.1**
	eval4（EARS）	11.8	**70.0**

ベースラインに対し，学習データに EARS を加えることで，EARS では CER が 1.6 ポイント，SER が 2.7 ポイントと大きく減少し，CSJ においても eval1 を除き，全体に精度の向上が見られた．ベースラインは高齢者に特化したモデルではないため，高齢者音声ではない CSJ の eval1 の CER が良かったと考えられる．

ラベル付けにより，SER が全ての評価データで良くなっている．これはラベルを付けたことで文全体を正しく捉えられるようになったと考えられる．CER は高齢者に限らない自発的発話である CSJ では良くなっているが，高齢者音声である EARS ではラベルなしの方が良い精度である．これは学習データ中のラベル付き文の割合が EARS よりも圧倒的に CSJ のものが多かったからであると考えられる．CSJ を含めたラベルの数は表 2 の通りであるが，EARS のみのラベル数は**表 4** のようになっており，データ全体に対し極めて小さい．データ中の言い淀みラベルの多くは CSJ のものであり，CSJ も EARS も同じ言い淀みではあるものの，高齢者以外の話者にも現れる一般的な言い淀みと高齢者の高齢化による言い淀みで実際の特徴は大きく異なる可能性もあり，今回は高齢者の言い淀みではなく一般的な言い淀みを多く学習したために EARS のラベル付きでの認識が上手くいかなかったと考えられる．

表 4　データ中の EARS の言い淀みラベルの数
Table 4　Number of the hesitation labels of EARS in the data.

データ	個数 [個]	文数(LS)[文]	文全体の数(S)[文]	LS/S[%]
train	1,563	897	503,207	0.18
validation	154	101	7,774	1.30
eval4	52	42	220	19.09

この問題は EARS の学習データの単純に数倍与えたりデータ拡張をしたりして学習データ中の EARS の割合を大きくすることで改善する可能性がある．また，高齢者と高齢者以外の言い淀みに違いがある場合，違いを考慮し，高齢者かの年齢情報のマルチタスク学習等を行うことで精度向上が期待できる．

実際のラベル付きの認識結果とラベル付きの正解文を見比べると，CER の改善が見られなかった EARS でも，

認識結果：　バンの@当局者は九人の生存者を確認したという
正解：　バンの@当局者は九人の生存者を確認したという

認識結果：　だから一度は一度は@おさわ一郎とも手応進んだ
正解：　だから一度は一度は@小沢一郎とも手を結んだ

認識結果：　これをんメディーベディア風にしたのがサボ人@になるＣＶだんだろう
正解：　これをん@メディア風にしたのがサブリ@ミナルＣＤだんだろう

というように正解文と同じ位置にラベルが出た文が多く含まれた．ラベル以外の部分の誤認識や，ラベルも完全に全て正しい位置に出ているわけではないため，CER の改善には繋がらなかったが正しい位置にラベルが出力されるということはモデルがラベルの意味を正しく学習できていると考えられ，ラベル付けは有効であるといえる．

6.　まとめ

本研究では End-to-End 音声認識において，言い淀みをラベル付けすることによって，モデルがその意味を学習で

きると仮定し，高齢者音声や自発的発話といった非流暢発話の精度がどう変化するか実験を行った．その結果，SER はすべての評価データで，61.7%から 59.7%，56.5%から 55.0%，45.6%から 43.1%，71.8%から 70.0%と 1.5 から 2 ポイントほど改善した．CER は CSJ の評価データでは最大 0.4 ポイント，すべて減少が見られたが EARS の評価データではラベルなしの方が良い精度となってしまった．これは学習データの量の調整やマルチタスク学習で改善する可能性がある．実際のラベル付きの認識結果と正解文を比較すると，モデルがラベルの意味を学習できていることが期待できる有効な結果を得られた．

謝辞 本研究は JSPS 科研費 JP19H01125 および 2020 年度国立情報学研究所公募型共同研究（20S0403）の助成を受けたものです．

参考文献

[1] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L. u. and Polosukhin, I.: Attention is All you Need, *Advances in Neural Information Processing Systems* (Guyon, I., Luxburg, U. V., Bengio, S., Wallach, H., Fergus, R., Vishwanathan, S. and Garnett, R., eds.), Curran Associates, Inc. (2017).

[2] Karita, S., Chen, N., Hayashi, T., Hori, T., Inaguma, H., Jiang, Z., Someki, M., Soplin, N. E. Y., Yamamoto, R., Wang, X., Watanabe, S., Yoshimura, T. and Zhang, W.: A Comparative Study on Transformer vs RNN in Speech Applications, *2019 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, pp. 449–456 (online), DOI: 10.1109/ASRU46091.2019.9003750 (2019).

[3] Watanabe, S., Hori, T., Karita, S., Hayashi, T., Nishitoba, J., Unno, Y., Soplin, N., Heymann, J., Wiesner, M., Chen, N., Renduchintala, A. and Ochiai, T.: ESPNet: End-to-end speech processing toolkit, *Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH*, Vol. 2018-September, pp. 2207–2211 (online), DOI: 10.21437/Interspeech.2018-1456 (2018).

[4] Watanabe, S., Boyer, F., Chang, X., Guo, P., Hayashi, T., Higuchi, Y., Hori, T., Huang, W.-C., Inaguma, H., Kamo, N., Karita, S., Li, C., Shi, J., Subramanian, A. S. and Zhang, W.: The 2020 ESPnet update: new features, broadened applications, performance improvements, and future plans (2020).

[5] 馬場 朗, 芳澤伸一, 山田実一, 李 晃伸, 鹿野清宏: 高齢者音響モデルによる大語彙連続音声認識, 電子情報通信学会論文誌. D-II, 情報・システム, II-パターン処理 = The transactions of the Institute of Electronics, Information and Communication Engineers. D-II, Vol. 85, No. 3, pp. 390–397 (2002).

[6] 河原達也, 李 晃伸: 連続音声認識ソフトウェア Julius（＜特集＞研究のツールボックス(2)), 人工知能, Vol. 20, No. 1, pp. 41–49（オンライン）, DOI: 10.11517/jjsai.20.1_41 (2005).

[7] 李 晃伸, 河原達也, 武田一哉, 鹿野清宏: Phonetic Tied-Mixture モデルを用いた大語彙連続音声認識, 情報処理学会研究報告. SLP, 音声言語情報処理, Vol. 29, pp. 43–48 (1999).

[8] Pellegrini, T., Trancoso, I., Hämäläinen, A., Calado, A., Dias, M. S. and Braga, D.: Impact of Age in ASR for the Elderly: Preliminary Experiments in European Portuguese, *Advances in Speech and Language Technologies for Iberian Languages - IberSPEECH 2012 Conference, Madrid, Spain, November 21-23, 2012. Proceedings*, Communications in Computer and Information Science, Vol. 328, Springer, pp. 139–147 (online), DOI: 10.1007/978-3-642-35292-8_15 (2012).

[9] Meinedo, H., Caseiro, D., Neto, J. P. and Trancoso, I.: AUDIMUS.MEDIA: A Broadcast News Speech Recognition System for the European Portuguese Language., *PROPOR*, Lecture Notes in Computer Science, Vol. 2721, Springer, pp. 9–17 (2003).

[10] 二宮利治, 清原 裕, 小原知之, 米本孝二: 日本における認知症の高齢者人口の将来推計に関する研究 平成 26 年度 総括・分担研究報告書（厚生労働科学研究費補助金厚生労働科学特別研究事業), 技術報告, 九州大学大学院医学研究院附属総合コホートセンター (2015).

[11] 河野和彦: ぜんぶわかる認知症の事典, 成美堂出版 (2020).

[12] 松田 実: 非流暢性発話の症候学, 高次脳機能研究 （旧失語症研究), Vol. 27, No. 2, pp. 139–147（オンライン), DOI: 10.2496/hbfr.27.139 (2007).

[13] Fukuda, M., Nishizaki, H., Iribe, Y., Nishimura, R. and Kitaoka, N.: Improving Speech Recognition for the Elderly: A New Corpus of Elderly Japanese Speech and Investigation of Acoustic Modeling for Speech Recognition, *Proceedings of the 12th Language Resources and Evaluation Conference*, Marseille, France, European Language Resources Association, pp. 6578–6585 (2020).

[14] 森 浩一: 吃音 (どもり) の評価と対応, 日本耳鼻咽喉科学会会報, Vol. 123, No. 9, pp. 1153–1160（オンライン), DOI: 10.3950/jibiinkoka.123.1153 (2020).

[15] MAEKAWA, K.: Corpus of Spontaneous Japanese : its design and evaluation, *Proceedings of The ISCA & IEEE Workshop on Spontaneous Speech Processing and Recognition (SSPR 2003)*, pp. 7–12 (2003).

[16] Itou, K., Yamamoto, M., Takeda, K., Takezawa, T., Matsuoka, T., Kobayashi, T., Shikano, K. and Itahashi, S.: JNAS: Japanese speech corpus for large vocabulary continuous speech recognition research, *Journal of the Acoustical Society of Japan (E)*, Vol. 20, No. 3, pp. 199–206 (online), DOI: 10.1250/ast.20.199 (1999).

[17] 新エネルギー産業技術総合開発機構（NEDO）: 新聞記事読み上げ高齢者音声コーパスの構築 (2001).

[18] Gulcehre, C., Firat, O., Xu, K., Cho, K., Barrault, L., Lin, H.-C., Bougares, F., Schwenk, H. and Bengio, Y.: On Using Monolingual Corpora in Neural Machine Translation (2015).

[19] Kannan, A., Wu, Y., Nguyen, P., Sainath, T. N., Chen, Z. and Prabhavalkar, R.: An Analysis of Incorporating an External Language Model into a Sequence-to-Sequence Model, *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 1–5828 (online), DOI: 10.1109/ICASSP.2018.8462682 (2018).