補助情報を同時に認識する日本語音声認識における出力表現形式の比較*

☆小堀聡太, 藤江真也 (千葉工大)

1 はじめに

音声認識における補助情報を検出する際の最適な出力表現形式について提案する．

音声対話システムなどにおいて人の自然な会話音声を対象とする音声認識では，フィラーや言い直しなどの非流暢性現象や,「うん」や「はい」といった感動詞など，読み上げ音声には現れない内容が含まれるため，それらを適切に扱う非通用がある．

対話において，フィラーとは「えー」「まー」といった言葉である．フィラーは発話権の保持 [1] や，話者の状態を表す情報 [2] としての機能を持つ．この機能は，対話における重要な役割を担っている．感動詞は「あっ」「うん」などを表す言葉である．これらは話者の感情や態度を直接的に表出する機能 [3] を担う．対話の流れを調整する役割 [3] も果たしている．言い直しは,「そそうですね」の最初の「そ」のような，繰り返し同じ言葉を発話することや，誤って発声した言葉を修正することを指す．この言語現象は発言の緩和 [4] に寄与している．会話の円滑化 [5] にも役立っている．本研究ではこれらを総称して補助情報と呼び，テキストと同時に認識する音声認識手法を提案する．

本研究では出力トークンにカタカナ文字のみの発音形を使用し，補助情報を検出する際に複合方式を組み合わせた手法を提案する．従来研究として，北岡ら [6] は，非言語現象を同時に検出する音声認識システムの研究を行っている．フィラーや笑い，言い誤り，非言語音声（咳等）といった 9 種類の言語・非言語現象の同時認識を試みている．Joint CTC/Attention Transformer モデルを用いた実験により，言語情報のみを認識する従来モデルと比較して認識精度が向上することを示している．非言語現象タグの付与位置について，対象の発話の前，後ろ，両側（囲む）の 3 条件で比較を行い，それぞれの認識精度への影響を調査している．

Inaguma ら [7] は，BLSTM-CTC モデルを用いて社会的シグナル検出と音声認識の統合を試みている．社会的信号（笑い，フィラー，バックチャネル，言い直し）のラベル付与方法として，サブワード単位の左側のみに付与する方法と両側に付与する方法で比較を行っている．実験結果から，従来の DNN-HMM モデルと比較して，音声認識の性能を劣化させることなく社会的信号検出が可能であることを示している

Tanaka ら [8] は，音声情報とテキスト情報を同時に活用した非流暢性現象の検出手法を提案している．BLSTM-CRFs をベースとした手法に Attention 機構を導入することで，音声とテキストのアライメントを必要とせずに両方の情報を効果的に利用することを可能にしている．言語的な特徴のみを用いた場合と比較して，音響特徴量と韻律特徴量を追加することで検出精度が向上することを示している．特に，location-based attention を用いた場合に，フィラーと言い直しの両方において最も高い F1 値が得られることを報告している．

本研究では，北岡らの手法である，出力トークンに平文（かな漢字混じり文）で構成された書字形と，補助情報を検出する際に提案した独立方式 [6] で比較した際の，補助情報の検出率と音声認識精度について報告する．

2 補助情報付き音声認識

2.1 補助情報

本研究では日本語日常会話コーパス（Corpus of Everyday Japanese Conversation; CEJC）[9] を使用する．このコーパスは自宅，飲食店，職場等の日常場面で自然に生じるリアルな活動を対象としている．多様な場面・多様な話者の音声と転記テキスト等が収録している．

CEJC の転記テキストには，多様なタグが付与されている．本研究ではこのうち，表 1 に示した三種類のタグを取り上げる．それ以外のタグに関係する補助情報は検出対象外とするため，発話内容のみを取り出しタグは削除する．

本研究では明示的にタグが付与されているもの以外

Table 1 補助情報タグの定義．各タグは CEJC の転記テキストに記載されている補助情報の付与方法を表す.

タグ	概要
(F)	「あの」「その」等がフィラーとして用いられる場合
(I)	「あ」「え」等の感動詞が挿入構造の内部にあり発話単位として分割されていない箇所
(D)	語の言いさし

*Comparison of Output Formats for Japanese Speech Recognition with Auxiliary Information by KOBORI, Souta, FUJIE, Shinya (Chiba Institute of Technology)

にも補助情報としていくつか加えたものがある．フィラーは，フィラー（F）のタグが付与されたものに加えて，形態素情報において品詞が感動詞-フィラーとされているものを用いる．言い直しは，言い淀み（D）のタグが付与されたもののみを用いる．感動詞は，感動詞（I）のタグが付与されたものと，形態素情報において品詞が感動詞-一般とされているものを用いる．

2.2 補助情報の検出方法

2.2.1 書字形と発音形

一般的に日本語音声認識の出力は，平文（漢字かな混じり文）が用いられる．実際にはトークンと呼ばれる記号の列が出力されるが，その単位としては，文字や Sentence Piece といったサブワードが用いられる．本研究ではこのような平文で文字をトークンとした出力形式を**書字形**と呼ぶ．

漢字は表音文字ではないため，発音と一対一で対応せず同音異義語が存在することや，語彙数（漢字の種類）が発音記号と比較して多くなることから，平文を出力表現とした音声認識は発音記号を出力としたものよりも難しいことが考えられる．そこで，本研究では日本語の発音に対応したカナ文字をトークンとした出力形式を**発音形**と呼び，補助情報検出のための音声認識のベースとして用いる．

2.2.2 独立方式と複合方式

音声認識における補助情報は，前節で説明した文字などの通常のトークンに対応した音声が入力に含まれるのではなく，それに対して正に補助的に付与されるものである．例えば，フィラーは実際には「えー」「えーと」などの文字として発音され，それがフィラーかそうでないかが補助情報として付与されるといった具合である．

テキストに対してこのような補助情報と付与する代表的な方法として，タグを別記号として加える方法がある．北岡ら [7] もこの方法を採用しており，例えばフィラーとして発声された部分を

（F え ー と ）そ れ は ...

のように，開始記号「(F」と終了記号「)」で囲むといったものである．このように，言語情報とは異なる特殊な記号を導入して補助情報として付与する方法を独立トークン方式（**独立方式**）と呼ぶ．独立方式には，**前置方式**，**後置方式**，**前後置方式**の三種類がある．前置方式は (F ○○) のように開始記号にのみ情報を付与するもの，後置方式は (○○ F) のように終了記号にのみ情報を付与するもの，前後置方式は (F ○○ F) のように両方に情報を付与するものである．

本研究では，補助情報に特別な記号を割り当てるのではなく，言語情報を表す通常トークンに対して +F などといった記号を付与した別のトークンを追加する方法を提案する．例えば，上記のフィラーを含む区間は

エ+F ー+F ト+F ソ レ ワ ...

と表されることになる．つまり，フィラーなどの補助情報を表す特別な記号を用意するのではなく，通常発声のトークン（「エ」）と補助情報を含んだ発声のトークン（「エ+F」など）を別々のトークンとして持つということである．本研究ではこの方式を複合トークン方式（**複合方式**）と呼ぶ．

本研究における提案手法は，発音形に複合方式を組み合わせた，**発音形複合トークン方式**である．複合方式を用いる場合，元々持っている通常のトークンの種類数と補助情報の種類数の乗算で合計のトークンの種類数が決まる．そのため，いたずらに問題を複雑にしないために通常のトークンの種類を抑えた発音形が適していると考えた．また，トークンに対応する発音は補助情報の有無によって変化することが考えられる．例えば,「えーと」という言葉の発音は，フィラーとして発声される場合と「A と」「エイト」などの言語情報として発声される場合では異なる発音になる．独立方式ではこれらの区別なく共通のトークンとして出力し，補助情報は周辺に音声とは直接的に関係のない記号として出力するため問題が難しくなると考えた．

Table 2　データセットの分割と各補助情報の出現頻度．各数値は補助情報の発生回数を示し，時間は各データ種別の総音声時間を表す．

	補助情報出現回数（回）			音声時間
	フィラー	感動詞	言い直し	(h)
学習	25,473	207,639	33,423	586.7
開発	566	3,725	810	10.3
評価	1,524	10,243	1,728	27.5

3 実験

3.1 実験条件

提案手法である発音形複合トークン方式とその他の方式について，補助情報の検出性能と音声認識結果を実験によって比較する．実験データは **2.1** で説明した CEJC で構成するが，評価データに多様な年代，性別の音声情報を含むように，表 2 に示すように学習・評価データに分割した．また，書字形と発音形，および独立方式と複合方式を組み合わせた各条件におけるトークンの種類は，表 3 に示す通りである．

音声認識モデルとして Transducer ベースのアーキテクチャを採用し，エンコーダには Contextual Block Conformer を使用する．エンコーダは 12 ブロックで

Table 3　各方式における書字形と発音形のトークンの種類の比較．独立方式と複合方式の各手法についての，書字形と発音形でのトークンの種類を表す．

	トークンの種類	
形式	独立方式	複合方式
発音形	138	426
書字形	2,792	3,444

構成され，出力次元数 256 次元，Attention ヘッド数 4 個，Feed-forward 層のユニット数 2,048 とした．

3.2　実験結果

独立方式と複合方式で同一条件での実験を実施する際，補助情報の検出数に差異が生じる．そのため，複合方式において複数に分かれてトークンごとにラベルが出現したものを，連続して同一のラベルが出力した区間を一つとしてカウントし，独立方式と同等の数が出力されるようにした．音声認識精度の評価は，補助情報のタグが付与された状態と，タグを削除した状態の両方について実施した．

表 4 に提案手法と従来手法における補助情報の適合率，再現率，F1 値を示す．書字形では，独立トークン方式が全体的に良い精度を示した．補助情報検出の F1 値については，フィラー では後置方式，感動詞では前置方式，言い直しでは前後置方式がそれぞれ最も高かった．補助情報別の形式による F1 値の分析から，フィラーについては後置方式を除く全ての方式で発音形が高い性能を示した．複合方式では 79.33%から 82.04%へと F1 値が向上し，2.7 ポイントの改善が確認された．感動詞については複合方式以外で書字形が優位であった．言い直しについては，発音形では前置方式と複合方式，書字形では後置方式と前後置方式が高い F1 値を示した．いずれの補助情報においても，提案手法である発音形複合トークン方式がわずかながら最もよい F1 値を示した．

音声認識精度を表 5 と表 6 に示す．表 5 は補助情報を含む認識器の認識結果から補助情報を除去して求めたもので，表 6 は補助情報を除去したテキストを出力として学習した認識器の結果から求めたものである．この結果から，書字形，発音形にかかわらず，補助情報ありの認識結果における認識精度の差はわずかであるが，補助情報なしで学習された認識器に比べると精度が向上が向上していることがわかる．

4　考察

発音形が書字形と比較して良好な結果を示した要因として，表 3 に示したトークンの種類数の違いが挙げられる．発音形は書字形よりも種類が少ないため，補助情報の検出率と音声認識精度の向上したと考えられる．

補助情報の検出率における発音形と書字形の間の差は表 4 に示した通り，比較的小さい結果となった．この理由としては，書字形では漢字による意味的な情報が文脈理解を助ける一方，発音形ではカタカナ文字のみによる制限された情報しか得られない．この特性が，トークン数の差がもたらす影響を部分的に相殺していると推察される．

補助情報の検出に関するエラー分析から，従来手法（独立方式）の問題点が明らかになった．独立方式は，音声情報と直接関係のない補助情報を示すタグの出力が必要となる点である．具体的には，(F，(D，(I や) が単独で出力される事例が多く観察された．この現象は独立方式全般で確認された．以下が発音形と前置方式を適用したときの具体例である．

(F ア ノ ナ ン カ キ カ ン ガ ア ッ …

理想としては，(F ア ノ) と出力されるべきだが，) が出力されていない．) は複数の開始タグを示す (F，(D，(I に対する，共通の終端記号として機能する．このため，適切な対応関係の判断が困難になっている．これが検出精度の低下につながると考えられる．

音声認識精度に関するエラー分析から，両形式の特徴が明らかになった．Sub が最も高い割合を示したことから，モデルは不確実な場合でも認識を省略せず，別の文字として認識する傾向が強い．Ins が最も低い割合を示したことから，モデルは存在しない要素を追加することに対して保守的な判断を行う特性を持つ．

補助情報タグの有無による比較から，システムの安定性が確認された．表 6 に示した通り，書字形，発音形のいずれにおいても，補助情報タグを付与した場合の方が CER が低くなることが確認された．これは，補助情報タグを認識対象に含めることで，通常のテキスト部分の認識性能も向上することを示している．具体的には，書字形では約 1.46%，発音形では約 1.01%の改善が見られた．書字形でより大きな改善が見られた理由として，補助情報タグが文脈の区切りを明示的に示すことで，周辺テキストの認識がより正確になった可能性が考えられる．このことから，補助情報タグの付与は認識精度を損なうどころか，むしろ全体的な認識性能の向上に寄与することが示された．

5　まとめ

本研究では，音声からの補助情報と言語情報の同時検出を行った．実験を通じて，提案手法が最も高い精度を達成することが確認された．補助情報の検出精度についてはさらなる向上の余地が残されている．

他の補助情報と比較して低い検出率を示した言い

Table 4　各方式における書字形と発音形の補助情報の検出率の比較．各方式における補助情報の検出精度を適合率・再現率・F1 値で表す．

補助情報	方式	書字形			発音形		
		適合率	再現率	F1 値	適合率	再現率	F1 値
フィラー	独立トークン前置方式	**82.51**	78.33	80.36	81.76	**81.26**	81.51
	独立トークン後置方式	84.04	78.13	**80.98**	82.54	78.59	80.52
	独立トークン前後置方式	82.02	**79.22**	80.60	82.05	80.03	81.02
	複合トークン方式	81.08	77.65	79.33	**83.19**	80.92	**82.04**
感動詞	独立トークン前置方式	86.48	87.71	**87.09**	86.15	86.12	86.14
	独立トークン後置方式	86.15	87.03	86.59	86.51	86.28	86.40
	独立トークン前後置方式	85.96	**87.83**	86.89	86.82	85.35	86.08
	複合トークン方式	**86.85**	85.92	86.39	**87.67**	**86.66**	**87.16**
言い直し	独立トークン前置方式	41.06	37.01	38.93	41.11	**37.14**	39.02
	独立トークン後置方式	**41.07**	36.33	38.55	39.52	37.07	38.26
	独立トークン前後置方式	40.33	**37.76**	**39.00**	40.92	36.45	38.56
	複合トークン方式	40.49	36.17	38.20	**42.72**	36.97	**39.64**

Table 5　各方式における書字形と発音形の CER の比較．置換・削除・挿入誤りと求めた誤り率で算出した文字誤り率を表す．Sub，Del，Ins はそれぞれ置換，削除，挿入の割合を示す．

方式	書字形				発音形			
	Sub	Del	Ins	CER	Sub	Del	Ins	CER
独立トークン前置方式	13.35	6.05	3.59	22.99	9.58	5.16	2.90	17.64
独立トークン後置方式	13.19	6.08	3.63	**22.90**	9.57	5.26	2.86	17.69
独立トークン前後置方式	13.30	6.11	3.56	22.97	9.50	5.32	2.71	**17.53**
複合トークン方式	13.36	6.08	3.63	23.07	9.61	5.04	3.03	17.68

Table 6　補助情報タグなしの書字形と発音形の音声認識精度．補助情報を含まない場合の置換・削除・挿入誤りと求めた誤り率で算出した文字誤り率を表す．Sub，Del，Ins はそれぞれ置換，削除，挿入の割合を示す．

形式	Sub	Del	Ins	CER
書字形	13.61	5.85	3.78	23.24
発音形	9.69	5.09	2.95	17.74

直しについては，新たな課題が明らかになった．その特有の特徴を考慮したモデルの改良が必要である．データの拡充も検討すべき課題である．

感動詞については，より詳細な区分に基づく検出率の調査が求められる．現状では，応答を表す感動詞「うん」「はい」と，驚きを表す感動詞「えっ」「あっ」など，異なる機能を持つ感動詞が単一のカテゴリとして扱われている．これらの感動詞の種類別の検出難易度を明らかにすることで，より効果的な検出手法の開発が可能になる．

参考文献

[1] 水上悦雄，山下浩二，“対話におけるフィラーの発話権保持機能の検証,” 認知科学，vol.14，no.4，pp. 588–603，2007.

[2] 定延利之，“会話においてフィラーを発するということ,” 音声研究，vol.14，no.3，pp. 27–39，2010.

[3] 森大毅，“対話システムはどのように話すべきか” 音響学会，vol.78，no.5，pp. 283—288，2022.

[4] 池田佳子，“会話に不可欠な「言い淀み」の機能の一考察,” 地域文化研究，vol.7，pp. 1–11，2001.

[5] 吉田悦子，Robin Lickley，“聞き手の行動における言い淀みの役割の日英比較分析,” 言語処理学会，pp. 1094–1097, 2010.

[6] 北岡教英，若林佑幸，塩根凪人，“言語現象と非言語現象も検出する音声認識システムの提案,” 信学技報，vol.123，no.88，pp.109–113, 2023.

[7] H. Inaguma，et al.，“An End-to-End Approach to Joint Social Signal Detection and Automatic Speech Recognition,” Proc. ICASSP，pp.6214–6218，2018.

[8] T. Tanaka, et al., “Disfluency detection based on speech-aware token-by token sequence labaling with BLSTM-CRFs and attention mechanisms,” Proc. APSIPA 2019, pp. 1009–1013, 2019.

[9] H. Koiso, et al., “Design and Evaluation of the Corpus of Everyday Japanese Conversation,” Proc. 13th Language Resources and Evaluation Conference, pp. 5587–5594, 2022.