音声言語情報処理 29－36
(1999. 12. 21)

ニュース音声認識における直前原稿を利用した認識性能の改善

松井 淳, 加藤 直人, 小林 彰夫, 今井 亨, 田中 英輝, 安藤 彰男

NHK 放送技術研究所

〒 157-8510 東京都世田谷区砧 1-10-11

TEL: 03-5494-2319

E-mail: {a_matsui, katonao, akio, imai, tanakah, ando}@strl.nhk.or.jp

あらまし ニュースのリアルタイム字幕作成を目的とした音声認識には一般のタスクと違う特徴がある．すなわち発声される可能性のあるテキスト，あるいはそれに類似したテキストを放送の前に入手できる特徴である．著者らはこの点に着目して認識性能を向上させる手法をいくつか研究してきた．これらは計算処理量の比較的重いものと軽いものがあり，軽い処理ほど放送直近の原稿を利用できる．本稿では，比較的処理の重い「言語モデルの適応化手法」，処理の軽い「放送直前の原稿を利用した n-gram 確率の動的計算法」，および「未知語自動登録法」を使った実験を報告し，これらを組み合わせて利用することで単語正解精度が 89.92%から 92.36%に改善されたことを示す．

キーワード ニュース音声認識，直前記者原稿

Improvement Methods for Broadcast News Transcription System Using the Latest News Manuscripts

Atsushi Matsui, Naoto Kato , Akio Kobayashi , Toru Imai, Hideki Tanaka, Akio Ando

NHK Science and Technical Research Laboratories

1-10-11, Kinuta, Setagaya-ku, Tokyo, 157-8510, Japan

TEL: 03-5494-2319

E-mail: {a_matsui, katonao, akio, imai, tanakah, ando}@strl.nhk.or.jp

Abstract Although it is almost impossible to guess what will be uttered beforehand in general speech recognition task, we can do this with high possibility in news dictation task thanks to the manuscripts that well hold the words to be uttered by announcers. In this paper, we will describe three ways of utilizing such news manuscripts and will show empirically that they greatly helped improving the news recognition rate.

key words broadcast news transcription, news manuscripts

1 まえがき

近年，難聴者や聴覚障害者への情報サービスを促進するため字幕放送を求める声が急速に高まりつつある．この要求に答えて，ドラマなど一部の番組の字幕放送が実現しており，その数は着実に増加している．しかし現在日本ではニュースのリアルタイム字幕放送は実現していない．英語と違って日本語の仮名漢字交じり文の作成には時間がかかるため，ワープロでリアルタイムに字幕を作成することが困難なためである[1]．

こうした背景をふまえて，NHKでは，リアルタイムでニュース音声を文字に変換するニュース音声認識システムの研究に取り組んでいる [1].

著者らはこれまでニュースの中でスタジオアナウンサーが担当する部分を中心に研究をすすめてきた．この部分にはアナウンサーが読み上げる「読み原稿」がある場合が多い．読み原稿は記者が電子的に入稿した「記者原稿」を元に作成している．このため，記者原稿の情報を音声認識システムに反映することで認識率を向上できる可能性がある．放送内容に一致する直前の記者原稿がもっとも有効であるが，ニュース内容によっては一定期間繰り返し放送されることもあるため[2] ある程度前の原稿も有効である．

放送の直前の記者原稿を音声認識システムに反映するには高速な手法が必須である．一方1日前の原稿の情報を反映するにはそれほど高速な手法でなくともよい．

小林ら [2] は，前日までのニュース原稿を認識システムに反映させる手法として時期依存言語モデルを提案し，約2%の単語正解精度の改善を示した．時期依存言語モデルの作成には1時間程度の処理を必要とするため，放送前1時間以内に入稿された原稿は利用できない．

これに対して加藤ら [5] は，直前の記者原稿を利用してデコード時にn-gram確率を動的に補正する手法（単語出現位置辞書）を提案している．この手法は小林らの手法より処理が軽いが，デコーダーに実装した場合の改善効果に関しては報告されていない．

著者ら [7] は，未知語を言語モデルに追加登録する手法を提案している．[7] では，未知語が事前に与えられたとして認識実験を行って効果を示している．しかし，直前の記者原稿から未知語を検出する部分を含めた評価はない．

本報告では，まず，アナウンサーの発話と記者原稿の類似性を定量的に評価する．次に，加藤らの報告と著者らの報告で不足していた部分を実装して，小林らの方法との併用による効果を認識実験で示す．

[1] 速記ワープロなどを使った手段も考えられるが，コストがかかり特殊な技能を必要とする問題がある．

[2] 通常の事件ではほぼ同一内容のニュースが一日程度放送される．さらに飛行機事故のような大事件は同一話題が続報の形で少しずつ内容を変えながら長期間放送される．

2 読み原稿と記者原稿

アナウンサーが放送で実際に読み上げる原稿を「読み原稿」と呼ぶ．これは記者がワープロ入力した原稿，すなわち「記者原稿」や，現場からFAXで送られてきた原稿に，手書きによる修正が加えられて作成される．

読み原稿は現時点で電子化されておらず認識システムで利用することはできないが記者原稿は利用することが可能である．例えば記者原稿には，人名や地名，専門用語など言語モデルにない語彙が含まれているため，この原稿を利用した認識精度の改善が期待できる．

以下にNHKニュースでの発話文と，それに対応する記者原稿の例を示す．(1998年9月30日放送分より)

（記者原稿） 台風九号の接近 に伴い 九州 沖縄各地で
交通機関に 影響が 出ています
（発 話 文） この 台風九号の接近 で 九州 では
交通機関に 乱れが 出ています

実際のニュースにおいて，放送前に入稿された記者原稿とニュースでの発話がどの程度重なっているかを見るため，1998年9月30日放送分の4つのNHKニュース（6時台，7時台，12時台，19時台）ごとに，発話文と記者原稿の類似の程度を定量的に調べた．

ここで「発話文」とは，ニュース音声から人手で発話内容を書き起こしたものを指す．ひとつのニュースの発話文には，アナウンサーやリポーターによる発話以外に，インタビュー等での一般話者による発話も含まれているが，一般話者による発話は，記者原稿として準備しない．そこで，一般話者の部分を除いた計782文を発話文とした．また記者原稿は全て，各々のニュースの放送前12時間内に入稿されたものを用いた．

記者原稿と発話文の類似の程度（類似度）としては，記者原稿を正解候補，発話文を出力文とみなした場合の単語正解精度を用いた．

各ニュースについて調べた類似度の度数分布を図1に示す．図1のグラフでは，いずれの時間帯のニュースも類似度90%から100%の階級にピークがある．

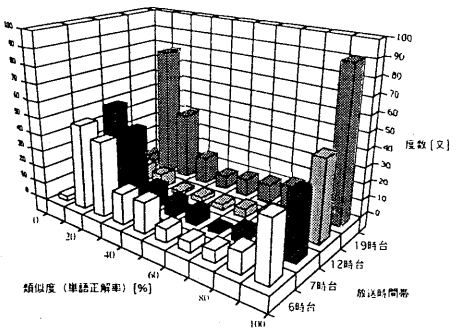

図1 発話内容と記者原稿の類似度

さらに発話と記者原稿が完全に一致する場合（類似度100%）の全発話に占める割合を表1に示す．この表から少なくとも1割ないし2割程度のニュース発話は記者原稿をそのまま読み上げたものであることがわかる[3]．

表1：記者原稿と完全一致する発話文の割合

放送時間帯	完全一致の割合 [%]
6時台	10.9
7時台	11.6
12時台	24.0
19時台	18.6

類似度の低い部分（10%～30%）の発話については，中継先との掛け合いや気象情報など，もともと記者原稿を作らない発話や，助詞や文末表現（「です」「ました」）など機能語の部分が一致しているに過ぎず内容の対応がほとんど認められない発話が大半であった．

3 直前記者原稿の利用法

前章で述べたように，アナウンサーの発話で記者原稿と100%一致するものは2割程度ある．1章で述べた3手法は，このようなニュース発話の性質を利用したものである．これらは以下で説明するように処理時間や対象とする問題に違いがある．以下では，各々の手法の概略と特徴を説明した後，同一の評価データを使用した認識実験について報告しその有効性を議論する．

3.1 時期依存言語モデル

小林ら[2]は，ニュース原稿のための言語モデルについて，学習期間，学習時期による統計的性質の変化を調べ，その結果，学習データは長期間であるほど，また，評価時期に近いほど有効であることを示した．このことから，長期間のニュース原稿と最新のニュース原稿を重み付け混合したデータを言語モデルの学習に用いる方針を示し，次のような方法を提案している．

長期間原稿（過去のニュース原稿）から作られた言語モデルをP_0，最新の記者原稿から作られた言語モデルをP_1として，それらの線形補間により時期依存言語モデルPを表現する．

$$P(y \mid z) = \lambda P_0(y \mid z) + (1-\lambda) P_1(y \mid z). \quad (1)$$

モデルへの確率重みλはEMアルゴリズムで求める．

時期依存言語モデルの特徴は，大量の長期間原稿と最新の記者原稿を学習時に併用することで，パラメータの統計的な信頼度を保ちつつ，新しく出現した単語に対応した言語モデルを作成できることである．しかし，基本的に学習データから言語モデルを構築し直す作業となるため，言語モデルの作成には，およそ1時間程度の処理時間を必要とする[4]．そのため，時期依存言語モデルで利用可能な記者原稿は，放送1時間以上前に入稿されたものに限られる．

3.2 単語出現位置辞書による n-gram 確率の動的再計算

加藤ら[5]は，記者原稿と一致する単語列のn-gram確率を動的に補正し，認識候補として優遇する手法を提案している．n-gram確率の補正計算については，次のような手法を採用している．

直前の記者原稿に出現した単語列は，発話での出現が十分期待できるので，記者原稿と一致する単語列を履歴に持つ認識候補を，記者原稿との一致の長さNに応じて優遇する．具体的には，Nがある閾値N_0をこ

[3] 発話文（書き起こし）と記者原稿には表記の不統一よる不一致がある．この影響で類似度が100%にならなかったものがあるため表1の数字は多少低めの数字となっている．

[4] Alpha21264 500MHz プロセッサ搭載の計算機上で語彙サイズ約2万の言語モデルを作成した場合．

えた場合について，n-gram 確率を次式の右辺で計算した値と置き換える．

$$p(w_i|w_{i-N-1},...,w_{i-1}) \cong Q_0(1-exp^{-\lambda \times N}) \quad (2)$$

ここで，パラメータ Q_0，λ は，あらかじめ適当な値に設定した定数とする（ただし，$0 \leq Q_0 \leq 1$, $\lambda > 0$）．

本手法の特徴は，bi-gram よりも長距離の情報を第1パスの段階から利用できることである．また，主な処理はデコード時に動的に行われるため，事前に必要とされる処理が軽く，より直近の記者原稿まで利用可能であることも，重要な特徴のひとつである．

加藤らの報告では，本手法の認識実験による評価が行われていなかったので，今回これをデコーダーに実装した[5]．実装するにあたっては，bi-gram 確率を式 (2) の右辺で置き換えた後の，全確率を正規化する計算を省略した．これは，正規化による認識率の改善よりも，それによる認識時間の増大の方が，遥かに大きいと考えたためである．また，パラメータ Q_0，λ の設定によっては，式 (2) による補正値が元の bi-gram 確率より小さくなり，候補として残したい単語の認識を妨げる結果となる危険があるので，補正による値が元の bi-gram 確率よりも大きくなる場合に限定する条件を，今回，新たに追加した．

3.3 未知語の自動登録

著者ら [7] は，未知語が事前に与えられることを前提とした未知語の言語モデルへの登録法を提案している．また認識実験では，評価データに出現する未知語全てを手動で登録した場合について，単語認識率で1%程度の改善効果を確認した．今回，より現実的な設定にするため，直前の記者原稿に出現した未登録語全てを未知語として，自動的に言語モデルに登録した．

未知語の検出を自動的に行うため，ここでは，簡易な方法ではあるが，記者原稿中に出現した言語モデルの未登録語を未知語とした．

未知語には，発音辞書への登録に必要な「読み」の情報と，言語モデルへの自動登録で必要な「品詞」の情報を，あらかじめ付与しておく必要がある．これには形態素解析ツール「茶筌」[6] で自動的に付与できる「読み」と品詞・活用の情報をそのまま利用した．

また未知語に与える確率値などの情報は，品詞を用いたスムージングによって求める [7]．具体的には，以下に示す方法を用いた．

以下，未知語を u，既知語の全体集合を V，u 以外の小文字は任意の単語を示すものとする．また，$G(x)$ は単語 x の属す品詞とし，$|G(x)|$ は $G(x)$ に属す既知語の総数とする．

まず，直前の記者原稿から検出した未知語と，言語モデルの登録語（既単語）を，品詞（単語クラス）に分類する．次に，未知語 u が品詞 $G(u)$ に属するとき，未知語 u を n-gram 言語モデルに登録するのに必要なデータを次のように推定する．

(a) 未知語 u の uni-gram 確率:

$$p(u) = \frac{1}{|G(u)|} \sum_{x \in G(u)} p(x). \quad (3)$$

(b) 未知語 u のバックオフ係数:

$$\alpha(u) = \frac{1}{|G(u)|} \sum_{x \in G(u)} \alpha(x). \quad (4)$$

(c) 未知語 u に連接する既単語の集合:

$$W_{pre} = \{x \mid x \in V, \ ^{\exists}s \in G(u) : p(s \mid x) \neq 0\}, \quad (5)$$

$$W_{post} = \{x \mid x \in V, \ ^{\exists}s \in G(u) : p(x \mid s) \neq 0\}. \quad (6)$$

(注) 既知語 x が複数の品詞を取りうる場合は
W_{pre}, W_{post} の要素から単語 x を除外する．

(d) 未知語 u と既知語 x の bi-gram 確率:

$$p(u \mid x) = \begin{cases} \frac{1}{|G(u)|} \sum_{s \in G(u)} p(s \mid x), & if \ \ x \in W_{pre} \\ \alpha(x)p(u). & otherwise \end{cases} \quad (7)$$

$$p(x \mid u) = \begin{cases} \frac{1}{|G(u)|} \sum_{s \in G(u)} p(x \mid s), & if \ \ x \in W_{post} \\ \alpha(u)p(x). & otherwise \end{cases} \quad (8)$$

(e) 未知語 $u \in G(u)$ と未知語 $v \in G(v)$ の
bi-gram 確率:

$$p(u \mid v) = \frac{1}{|G(u)||G(v)|} \sum_{s \in G(u), t \in G(v)} p(s \mid t). \quad (9)$$

なお，n-gram 追加後の確率値の正規化については，計算が膨大となるため，放送直前での高速な処理を想定している都合上実施しなかった．

[5] 2 パスデコーダー [1] の第 1 パスで，bi-gram により言語スコアを計算する部分に実装した

4 実験

前章で提案した，単語出現位置辞書のデコーダーへの実装，および，直前記者原稿を用いた未知語の自動検出・自動登録による，認識率の改善効果を検証するため，以下の条件で認識実験をおこなった．

評価データは，1998年9月30日放送分のニュース音声データベース[3]の，一般話者を除く発話の中から502文（性別既知，総単語数13,458）を選んだ．

音響パラメータ，音響モデル，デコーダーは，以下のものを使用した．

○音響パラメータ
・ 39次元（12次元MFCC+log Power, $\Delta, \Delta\Delta$）
○ 音響モデル
・男女別，不特定話者
・混合ガウス分布型HMM(triphone)
・アナウンサー適応化[4]
○ デコーダー[1]
・ 2パスデコーダー
- 第1パス：bi-gram, Viterbi beam search, N-best
- 第2パス：tri-gram rescoreing

言語モデルとして，下記の4種類を用意して比較した．この中で直前の記者原稿を利用するもの（時期依存言語モデル，単語出現位置辞書，未知語自動登録）については，評価データの放送時間（6時，7時，12時，19時）ごとに記者原稿を集め，それぞれの原稿について言語モデルを別に作成した．

(1) 標準モデル (Baseline)
NHKニュース原稿 '91. 4/1 - '98. 9/29で学習した言語モデル

(2) 時期依存言語モデル (TDLM)
NHKニュース原稿を長期間原稿，放送13時間前～1時間前のNHK記者原稿を直前原稿とした，時期依存言語モデル．

(3) 単語出現位置辞書の併用（WPD+）
時期依存言語モデルと単語出現位置辞書を併用したモデル（$\lambda = 1.0, Q_0 = 0.1, N_0 = 3$）

(4) 未知語自動登録（OOV+）
(3)の言語モデルに，放送13時間前～10分前のNHK記者原稿から自動検出した未登録語を未知語として追加登録した言語モデル．

なお，語彙サイズは全て約2万単語に統一した．言語モデル学習時のcut-off値は，1 (bi-gram) - 2(tri-gram)とし，Good-Turingのスムージングを行った．

評価データには，記者原稿と一致する発話と，一致しない発話が混在してる．そこで，認識結果を評価するにあたっては，(3), (4)のモデルに利用した記者原稿との類似度（第2章参照）により，評価データを次のように分類した．

「完全一致」 類似度100%の発話（76文）
「部分一致」 類似度100%未満，50%以上の発話（148文）
「一致なし」 類似度50%未満の発話（278文）

評価データ全体（Total），および，記者原稿との一致の度合いによる分類（完全一致，部分一致，一致なし）についてまとめた単語正解精度を表2に示す．

表2：単語正解精度（単位は全て%）

分類	Baseline	TDLM	WPD+	OOV+
完全一致	95.39	98.68	99.17	99.28
部分一致	94.04	96.89	97.55	97.92
一致なし	81.75	81.35	81.54	81.60
Total	89.92	91.69	92.15	92.36

また，表2の「完全一致」,「部分一致」,「一致なし」の部分についてのグラフを図2に示す．

表2の「Total」で示すように，記者原稿を利用する3つの手法を全て組み合わせた結果（OOV+），評価データ全体に対する単語正解精度で，baselineの89.92%から92.36%に改善された．これを単語誤り率で表現すると，10.08%から7.64%への改善となり，約24%の誤りが削減されたことがわかる．

また，記者原稿との一致による分類ごとに見てみると，「完全一致」で3.89%，「部分一致」で3.88%，それぞれ単語正解精度が向上し，「一致なし」では，ほとんど変化しなかった．このことから，提案手法が対象とする発話（完全一致，部分一致）については，期待した改善効果が得られ，対象外の発話（一致なし）に

ついては，全体的に悪影響が及ぶことは殆ど無いと言える．

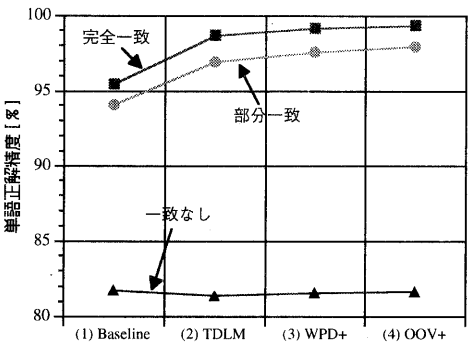

図 2 単語正解精度

提案手法で改善されなかった誤りの内容を見るため，全ての手法を併用した言語モデル（OOV+）で「完全一致」の発話を認識した結果を調べたところ，ほぼ次の 3 パターンに分類できた．

○ 助詞（「は」と「が」など）の置換誤り
○ 不要語（「え」や「あ」など）の部分での挿入・置換誤り
○ 文頭の単語の置換誤り

このうち，文頭の単語の誤りについては，単語履歴が文頭のみであるために，単語出現位置辞書による補正では原理的に改善できない誤りであった．

5 まとめ

放送 1 時間より前の原稿を言語モデルに反映する時期依存言語モデル，処理が軽く直前の原稿の情報を利用できる単語出現位置辞書による n-gram 確率の動的計算法，同じく処理の軽い未知語自動登録法について述べ，音声認識実験によってこれらの効果を議論した．最後に今後の課題を述べる．単語出現位置辞書による確率計算は，今回第 1 パスに実装したが，第 2 パスへの実装を考えたい．未知語自動登録法は電子化された原稿を使って実験した．しかしこの手法は FAX・手書きの文章中の未知語にも使うことができる．今後はこのような応用も考えたい．

参考文献

[1] 今井 亨, 安藤 彰男「ニュース番組自動字幕化のための音声認識システム」情報処理学会研究会, HI80-11(SLP-23-11), pp.59-64(1998).

[2] 小林 彰夫, 今井 亨, 安藤 彰男, 中林 克己「ニュース音声認識のための時期依存言語モデル」情報処理学会論文誌, pp.1421-1429(1999).

[3] 安藤 彰男, 宮坂 栄一「ニュース音声データベースの構築」音響学会平成 9 年春季研究発表会講演論文集, pp.157-158 (1997).

[4] 佐藤 庄衛, 今井 亨, 安藤 彰男,「音響モデル学習サンプル自動選択の検討」音響学会平成 11 年春季研究発表会講演論文集, pp. 9-10(1999).

[5] 加藤 直人, 浦谷 則好, 江原 暉将「記者原稿を利用したニュース音声認識のための言語モデル」言語処理学会第 5 会年次大会発表論文集, pp.269-272 (1999).

[6] 松本 裕治, 竹内 啓, 山下 達雄, 平野 喜隆, 今一 修, 今村友明「日本語形態素解析システム『茶筌』ver.1.5 使用説明書」(1997).

[7] 松井 淳, 小林 彰夫, 今井 亨, 安藤 彰男「ニュース音声認識のための未知語登録法」音響学会平成 11 年秋季研究発表会講演論文集, pp. 67-68 (1999).