Vol.2016-NL-226 No.1
Vol.2016-SLP-111 No.1
2016/5/16

話し言葉音声認識における非言語情報を考慮したRNN言語モデル

外山 翔平[1,a)]　齋藤 大輔[1,b)]　峯松 信明[1,c)]

概要：音声認識は一般に音響モデル，言語モデル，及びデコーダの組み合わせで構築される．言語モデルとして標準的に使われている n-gram 言語モデルを Recurrent Neural Network 言語モデルによって補完することにより，認識精度が改善されると報告されている．さらに，RNN 言語モデルに対して，単語の品詞や発話のトピックなどの付加的な情報を与えることで，言語モデルを発話に適応させる研究も数多くなされている．ところで，音声には話者性や発話状況といった非言語情報が含まれており，このような情報は話者が発する単語やその連接に影響を与えると考えられる．そこで本研究では，話し言葉音声認識において，入力音声を音声処理して抽出される非言語情報を RNN 言語モデルに組み込み，発話に適応させるモデルを提案する．また，日本語話し言葉コーパスを用いた実験において，提案手法によってパープレキシティが改善されることを示した．

キーワード：話し言葉音声認識，RNN 言語モデル，リスコアリング，非言語情報

Recurrent Neural Network Language Model using Non-Verbal Features for Automatic Speech Recognition

Shohei Toyama[1,a)]　Daisuke Saito[1,b)]　Nobuaki Minematsu[1,c)]

1. はじめに

音声認識では，入力音声に対し音響モデルと言語モデルをデコーダによって統合し，単語列に書き起こしている．言語モデルにはこれまで長年 n-gram モデルというシンプルかつ強力なものが使われてきたが，近年では Neural Network (NN) を使った言語モデルが提案されており，これらは n-gram モデルよりも優れているとされる [1], [2]．

例えば，これらの，認識システムが出力する認識仮説に対して，新しい言語モデルに基づくリスコアリングを行ない，再評価するという形でこれらの言語モデルを導入する研究が行われてきた [3]．特に NN による新しい言語モデルは，従来の言語モデルでは扱いづらかった付加的な情報を挿入しやすいため，単語を予測する上で効果的な情報を付加することで，より発話や文書に適した言語モデルを作ることができる [4], [5]．言語モデルに付加される情報としては，単語単位の品詞や上位語，文書単位のトピックなどの言語的なものと [6], [7]，単語単位のピッチやポーズの長さなどの音響的なものが挙げられる [3]．

音声認識の対象となる話し言葉には，年齢や性別，話し方に見られる癖などの個人性などの話者性や，話し相手との人間関係や場所などの発話環境など，言語には直接表れてこない，非言語情報と呼ばれる要素が必ず含まれている [8]．例えば日本語の場合，相手が目上か否かによって敬語という言葉かどうかを変える．このように，非言語情報は発話される単語に対して影響を与えると考えられる [9]．

そこで本研究では，音声認識において話し言葉特有の非言語情報を利用して，言語モデルを発話に適応させるモデルを提案する．本モデルによって，非言語情報から話者性や発話環境を推定し，その情報を考慮した言語モデルが得

[1] 東京大学大学院工学系研究科
Graduate School of Engeneering, The University of Tokyo
[a)] toyama@gavo.t.u-tokyo.ac.jp
[b)] dsk_saito@gavo.t.u-tokyo.ac.jp
[c)] mine@gavo.t.u-tokyo.ac.jp

られると期待される．

2. 音声認識

2.1 音声認識システム

音声認識とは，音声波形から発話された単語列を推定する過程を言う．この過程では，音響モデルと言語モデルという 2 つのモデルと照合することで膨大な仮説の中から最も適した仮説を選択している（デコーディング）．すなわち，入力音声 V に対して，書き起こし文 X を次の式に従って推測する．

$$X = \underset{X_i}{\operatorname{argmax}} P(X_i|V) \tag{1}$$

$$= \underset{X_i}{\operatorname{argmax}} \frac{P(V|X_i)P(X_i)}{P(V)} \tag{2}$$

$$= \underset{X_i}{\operatorname{argmax}} P(V|X_i)P(X_i) \tag{3}$$

式 (3) における第一項 $P(V|X)$ は音響モデルと呼ばれ，単語列と音声波形の関係をモデル化したものである．第二項 $P(X)$ は言語モデルと呼ばれ，書き起こし文の単語列 X がどのくらいの確率で生成されるのかをモデル化したものである．この言語モデルは「ある発話として単語列がどの程度自然なものか」を定量化したものと見なすことができる．

2.2 言語モデル

言語モデルでは，単語列 $X = \{x_1, x_2, \cdots, x_n\}$ を先頭から順に 1 単語ずつ生起されてできたものとみなし，この単語列の生起確率を各単語の生起確率の積で表す．すなわち，

$$P(X) = \prod_{i=1}^{n-1} P(x_{i+1}|x_1, \cdots, x_i) \tag{4}$$

として求める．

2.2.1 n-gram モデル

この $P(x_{i+1}|x_1, \cdots, x_i)$ のモデルとして，これまでは n-gram モデルと呼ばれるモデルが使われてきた．このモデルは，計算コストが低い一方で比較的精度が高く，自然言語処理や音声認識において広く利用されてきた．

n-gram モデルでは，ある単語 x_{i+1} の前にある N 個の単語との共起頻度を，学習データにおける出現回数から求める．

$$P(x_{i+1}|x_1, \cdots, x_i) = P(x_{i+1}|x_{i-N}, \cdots, x_i) \tag{5}$$

$$= \frac{Count(x_{i-N}, \cdots, x_{i+1})}{Count(x_{i-N}, \cdots, x_i)} \tag{6}$$

ここで，$Count(x_{i-n}, \cdots, x_i)$ は，単語列 $x_{i-n}, \cdots, x_i$ が学習コーパス中に現れた回数を表すものとする．一般に注目する単語数 N としては，3 から 5 がしばしば用いられている．

音声認識における言語モデルとしては長年 n-gram モデルが使われてきているが，n-gram モデルにはいくつかの欠点が存在する．まず，それぞれの単語を別のものとして扱うため，単語間の関係や意味を考慮することが難しい．また，注目する単語長 N が大きければ大きいほど学習データ中の多くの単語列パターンについて分析することができる一方で，各パターンの出現回数が少なくなってしまうという，データスパースネスと呼ばれる問題が知られている．

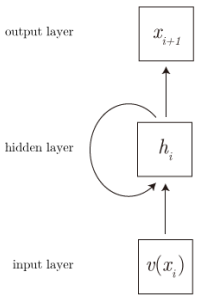

図 1 RNN 言語モデル
Fig. 1 RNN Language Model (RNNLM)

2.2.2 RNN 言語モデル

n-gram モデルに対して，近年では Recurrent Neural Network Language Model (RNN 言語モデル, RNNLM) [1] と呼ばれる言語モデルが提案されている．このモデルは**図 1** のような再帰構造を持った NN による言語モデルで，固定長のベクトルで表現された単語を，同様に固定長で表現される隠れ層に順に合成していき，文脈ベクトルから次に出てくる単語を推定するモデルである．

RNN 言語モデルでは式 (4) の $P(x_{i+1}|x_1, \cdots, x_k)$ を 2 つのステップによって再帰的に求める．まず，現在の単語 x_i の単語ベクトル $v(x_i)$ と，一つ前の隠れ層 h_{i-1} を合成し，現在の隠れ層を得る．

$$h_i = f_h(x_i, h_{i-1}) \tag{7}$$

$$= f(W_{hx}v(x_i) + W_{hh}h_{i-1} + b_h) \tag{8}$$

次に，現在の隠れ層 h_i から次の単語の出現確率を求める．

$$P(x_{i+1}|x_1, \cdots, x_i) = f_x(h_i) \tag{9}$$

$$= \text{softmax}(W_{xh}h_i + b_x) \tag{10}$$

なお，W_{xh}, W_{hx}, W_{hh} は NN の重み行列，b_x, b_h は NN のバイアスベクトルである．また，f は活性化関数と呼ばれる非線形関数で，sigmoid 関数や tanh 関数などが用いられる．

RNN 言語モデルは単語を連続値として表現，学習するため，単語の意味の類似度を捉えることができるとされている [10]．また，n-gram モデルが決められた個数の単語しか注目していないのに対し，RNN 言語モデルでは隠れ層にこれまで現れた単語の情報が合成されており，理論的には過去全ての文脈を保持している．その結果，RNN 言語モデルは n-gram モデルに比べて優れているという結果が数多く報告されている [10], [11], [12].

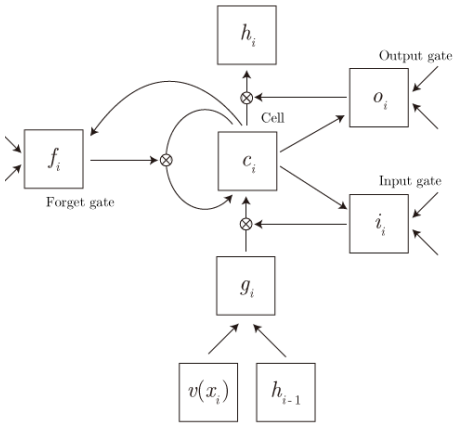

図 2　LSTM の構造
Fig. 2　Long Short-Term Memory

2.2.3　LSTM 言語モデル

さらに近年では，RNN を改良した Long Short-Term Memory (LSTM) [13] と呼ばれる Network を利用した言語モデル (LSTM 言語モデル, LSTMLM) も用いられるようになってきた [2], [14], [15]．LSTM では過去の状態がより残りやすいよう，中間層を再帰的に求める際，Input gate, Output gate, Forget gate, Cell と呼ばれる 4 つの素子を使う (**図 2**)．

LSTM 言語モデルでは，式 (7) における $h_i = f_h(x_i, h_{i-1})$ を次の手順に従って求める．

$$i_i = \sigma(W_{ix}v(x_i) + W_{ih}h_{i-1} + W_{ic}c_{i-1} + b_i) \quad (11)$$
$$f_i = \sigma(W_{fx}v(x_i) + W_{fh}h_{i-1} + W_{fc}c_{i-1} + b_f) \quad (12)$$
$$g_i = f(W_{gx}v(x_i) + W_{gh}h_{i-1} + b_g) \quad (13)$$
$$c_i = f_i \otimes c_{i-1} + i_i \otimes g_i \quad (14)$$
$$o_i = \sigma(W_{ox}v(x_i) + W_{oh}h_{i-1} + W_{oc}c_i + b_o) \quad (15)$$
$$h_i = o_i \otimes \tanh(c_i) \quad (16)$$

$W_{ix}, W_{ih}, W_{ic}, W_{fx}, W_{fh}, W_{fc}, W_{gx}, W_{gh}, W_{ox}, W_{oh}, W_{oc}$ は重み行列であり，b_i, b_f, b_o, b_g はバイアスベクトルである．また，$\otimes$ はベクトルの要素積を表す．

RNN では隠れ層に対する活性化関数による正規化が各ステップにおいて行われるため，長い単語列を処理する場合，勾配消失と呼ばれる学習がうまく行われない問題が発生しやすい．これに対して，LSTM では上記の 4 つの機構を導入することでこれを避けており，長い文に対しても適切に学習を行うことができる．

2.2.4　言語モデルの評価

言語モデルの性能の指標としては，しばしばパープレキシティ (Perplexity, PP) を用いる．これは，文書中の各単語を言語モデルによって予測した時の平均候補数（平均分岐数）を表している．すなわち，同一のテストセットに対して，PP をより小さくする言語モデルの方が，候補数を削減できているため，効果的な絞り込みが可能となる．

単語数 N の文書 D におけるパープレキシティ $PP(D)$ は次の式 (17) で表される．

$$PP(D) = \left(\frac{1}{P(D)}\right)^{\frac{1}{N}} \quad (17)$$
$$= \left(\prod_{i=1}^{N} \frac{1}{p(x_i|x_1, \cdots, x_{i-1})}\right)^{\frac{1}{N}} \quad (18)$$

この他，音声認識に用いる言語モデルの評価においては，音声認識の精度である Word Error Rate (WER) も言語モデルの性能の比較に利用される．

2.3　リスコアリング

通常の音声認識では n-gram モデルに基づいた処理を行なっているが，近年の他の優れた言語モデルを利用する方法の一つとして，リスコアリングが挙げられる．リスコアリングでは，まず，式 (3) に従って通常の音声認識を行うが，この際に認識結果を確率 $P(X_i|V)$ が最大の一つに絞るのではなく，認識仮説を一定数に絞るに留める．得られた認識仮説 X_i それぞれに対し，別の言語モデルによって評価し，言語スコア $Q(X_i)$ を算出する．n-gram モデルによる言語スコア $P(X_i)$ と異なるモデルで得られた言語スコア $Q(X_i)$ を重み w_l で足し合わせ，新たな言語スコア $P'(X_i)$ とする．

$$P'(X_i) = w_l P(X_i) + (1 - w_l)Q(X_i) \quad (19)$$

このようにして得られた新たな言語スコアと音響スコアを用いて認識仮説を再評価し，認識結果 X を求める．

$$X = \underset{X_i}{\operatorname{argmax}}\, P(V|X_i)P'(X_i) \quad (20)$$

これまでドメイン適応した n-gram モデルによるリスコアリングの研究などが行われてきたが，近年では RNN 言語モデルや LSTM 言語モデルによるリスコアリングの研究も注目されており，ツールキットも公開されている [16]．

3.　付加情報を考慮した言語モデル

言語モデルは「ある単語列がどの程度自然か」を確率的にモデル化したものである．一般の音声認識で用いられる言語モデルは，式 (4) に従い，これまで出てきた単語列から，これから出てくる単語を予測するというものである．これを発展させて，これまで出てきた単語列の他に，単語を予測する上で手がかりとなるような付加的な情報を付け加えた言語モデルが提案されている．特に，NN をベースにした言語モデルでは様々な付加情報を挿入することが容易なため，盛んに研究が行われている [3], [4], [5]．ここで加える付加情報としては，大きく分けて言語的な情報と非言語的な情報がある．

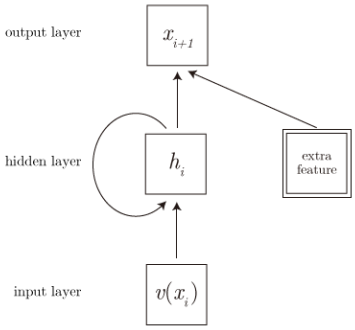

図 3 Document Vector を加えた RNN 言語モデル [17]
Fig. 3 RNNLM with a document vector [17]

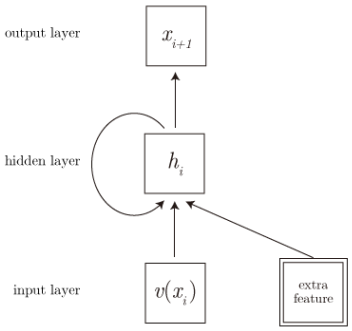

図 4 局所的言語情報を付与した RNN 言語モデル [5]
Fig. 4 RNNLM with additonal features [5]

3.1 言語的情報を付加した言語モデル

言語モデルに付加される言語的情報は，単語や形態素単位の局所的なものと，文や文章全体にまたがる大局的なものに大別できる．これらは，音声認識だけではなく自然言語処理における言語モデルにおいて利用されている．

3.1.1 大局的言語情報

大局的言語情報とは，自然言語処理によって得られる，発話や文書を通して一貫した特徴を指す．代表的なものとして，文書や発話が何について述べているのかを表すテーマやトピックを扱った研究がある．

Aaron らは，Latent Dirichlet Allocation によってトピックを抽出し，トピックごとに n-gram モデルを学習するモデルを提案している [7]．また，Rei らは，文書を固定長のベクトルで表現した Paragraph Vector [4] を逐次的に更新するようにした Document Vector を RNN 言語モデルに組み込んだ結果，言語モデルの精度を改善している [17]．このモデルでは，文書 X の大局的言語情報 $S(X)$ を図 **3** のように挿入している．この時，出力層を求める式 (9) は次の様になる．

$$P(x_{i+1}|x_1,\cdots,x_i) = f_x(h_i, S(X)) \quad (21)$$
$$= \mathrm{softmax}(W_{xh}h_i + W_{xs}S(X) + b_x) \quad (22)$$

3.1.2 局所的言語情報

局所的言語情報とは，単語や形態素単位に与えられる言語的な情報を指す．例えば，多くの文書や発話における単語列は文法という規則に則って並べられている．そのため，各単語や形態素における品詞や活用形によって次に続く単語を絞ることができると考えられる．

Shi らは，各単語の品詞と語幹という局所的な情報と，14 種類に区分した会話の社会的な状況という大局的な情報を組み込んだ RNN 言語モデルを構築した [5]．単語 x_i に関する特徴量を $Local(x_i)$ とし，文書全体に関する大局的な特徴量を $Global(X)$ を連結して特徴量 $S(x_i) = [Local(x_i), Global(X)]$ とし，隠れ層 h_i を次の式によって求めている (図 **4**)．

$$h_i = f_h(x_i, h_{i-1}) \quad (23)$$
$$= \sigma(W_{hx}v(x_i) + W_{hh}h_{i-1} + W_{hs}S(x_i) + b_h) \quad (24)$$

なお，$\sigma(\cdot)$ はシグモイド関数を表す．また，W_{hs} は W_{hx}, W_{hh} 同様，NN の重み行列である．

Arisoy らは，単語単位の RNN 言語モデルと形態素単位の RNN 言語モデルの出力層のみを共有することで，並列して学習が行われるような RNN 言語モデルを提案している [14]．英語やアラビア語などの形態素がそれぞれ意味を持つ言語において，形態素単位の言語モデルや形態素のクラスの言語モデルを単語単位のものと別に構築することで，言語モデル全体の性能が上がるという研究もなされている [18], [19]．

3.2 非言語情報を付加した言語モデル

言語モデルに付加される非言語情報としては，単語やモーラ単位の局所的な音響情報がある．

3.2.1 局所的音響情報

局所的音響情報とは，各単語やモーラなどに対応する，音声波形から得られる音響的な特徴量を指す．

Gangireddy らの研究では，各単語や形態素に対して「単語の発話長」「前単語との間隔」「基本周波数分布」などの局所的な音響特徴量を抽出し，これを RNN 言語モデルの中間層の計算に組み込んでいる [3]．RNN 言語モデルの構造としては，局所的言語情報を RNN 言語モデルに組み込んだモデル図 4 と類似している．単語 x_i に対応する音響特徴量を $S(x_i)$ とすると，RNN 言語モデルの中間層 h_i を次の式 (25) に従って求める．

$$h_i = \sigma(W_{hx}v(x_i) + W_{hh}h_{i-1} + W_{hs}S(x_i)) \quad (25)$$

Fu らは各単語に対応するフレームのピッチ，エネルギー，発話長を平均した特徴量を RNN 言語モデルに加えることで，英語読み上げコーパスである LibriSpeech におけるパープレキシティや WER が改善したと報告している [20]．

4. 非言語情報を付加した RNN 言語モデル

話し言葉音声認識の対象となるような自由発話を考えてみると，そこで使われる単語は話者や発話環境に依存すると想像できる [9]．例えば，性別・年齢・出身地・口癖・感情など，話者によって発話に使う単語が変わってくると考えられる．また，会議・インタビュー・ニュース・自然対話や話し相手との関係など，発話環境によっても使う単語は変わる．そこで，話し言葉音声認識に用いる言語モデルについて，これらの要素を付加情報として与えたい．

書き言葉においては，単語列は文法規則というものに則って配列されているため，構文解析やトピック分析などの自然言語処理による言語的情報の抽出を比較的高い精度で行うことができる．しかし，音声認識においては扱う対象が話し言葉であり，繰り返しの発生やフィラーの挿入など，単語列は必ずしも文法規則に従わないため，書き言葉に比べて言語的情報の抽出が難しい．また，言語的な情報はある程度の単語が集まらないと安定した解析が難しいといった欠点がある．

一方，音声には非言語情報と呼ばれる，その名の通り文字面には直接現れない情報が含まれている [8], [21]．音声から聞き手が推測できる話者の性別・年齢・感情などが非言語情報に該当し，これらは音声を音響的に解析することで得られる．非言語情報は，感情認識や音楽分類などのタスクにおいて有効であることが示されている [22]．

そこで本研究では，話し言葉音声認識における RNN 言語モデルに，発話を音響的に解析すること出られる非言語情報を利用することで，言語モデルの精度を改善することを目的とする．前述の通り，非言語情報には発話の話者や環境を読み取ることができ，これらを考慮することで言語モデルの性能が改善されると期待される．また，非言語情報は比較的短時間の音声からでも話者や発話状況を推定することができる．

4.1 非言語情報の抽出

本研究では，非言語情報を表すものとして，openSMILE [23] というツールキットから得られる特徴量と i-vector の 2 種類を扱う．これらの非言語情報を Feed Forward 型の NN によって次元圧縮したものを RNN 言語モデルに付加する．なお，この圧縮部についても RNN 言語モデルと同時に学習を行う．

openSMILE は，感情認識や音楽分類などのタスクにおいてしばしば特徴量抽出に用いられているツールキットである．単一音声データをフレーム分割してそれぞれについて音響特徴量とその動的特徴量を計算し，それらに対して様々な統計処理を行うことで，発話を単位とした大域的特徴量を生成する．特徴量と統計量の種類は少なくとも十数パターンに上るため，その組み合わせで生じる特徴量の数は数百から数千となる．openSMILE で抽出される音響特徴量には，ピッチ・エネルギー・MFCC・線形予測係数などがあり，それらの統計量としては平均・分散・レンジ・四分位範囲などがある．

i-vector は，少量の発話からでも安定して得られる話者性の情報を持った特徴量ベクトルであり，話者認識や言語識別において広く用いられている [24], [25]．一発話から推定された Gaussian Mixture Model (GMM) 中の平均ベクトルを連結し，発話を GMM スーパーベクトルとして表す．この GMM スーパーベクトルを主成分分析して i-vector を得る．

4.2 RNN 言語モデルへの付加情報の組み込み

第 3 章で触れた RNN 言語モデルへの付加情報の組み込み方に倣い，非言語情報を RNN 言語モデルに組み込む．図 4 のように中間層に情報を挿入する hidden model と図 3 のように出力層に情報を挿入する output model に加え，これらを合わせたような dual model (**図 5**) において性能を比較する．hidden model は，式 (24) において，S_i に非言語情報を対応させたものである．output model は，式 (22) において，$S(X)$ に非言語情報を対応させたものである．発話 V の非言語情報を表すベクトルを $S(V)$ とするとこの複合モデルは次の式で表される．

$$P(x_{i+1}|x_1,\cdots,x_i) = \mathrm{softmax}(W_{xh}h_i + W_{xs}S(V) + b_x) \tag{26}$$

$$\begin{aligned}
i_i &= \sigma(W_{ix}x_i + W_{iS}S(V) \\
&\quad + W_{ih}h_{i-1} + W_{ic}c_{i-1} + b_i) && (27)\\
f_i &= \sigma(W_{fx}x_i + W_{fS}S(V) \\
&\quad + W_{fh}h_{i-1} + W_{fc}c_{i-1} + b_f) && (28)\\
g_i &= f(W_{gx}x_i + W_{gS}S(V) + W_{gh}h_{i-1} + b_g) && (29)\\
c_i &= f_i \otimes c_{i-1} + i_i \otimes g_i && (30)\\
o_i &= \sigma(W_{ox}x_i + W_{oS}S(V) \\
&\quad + W_{oh}h_{i-1} + W_{oc}c_i + b_o) && (31)\\
h_i &= o_i \otimes tanh(c_i) && (32)
\end{aligned}$$

なお，$W_{iS}, W_{fS}, W_{gS}, W_{oS}$ はいずれも NN の重み行列である．

5. 実験

RNN 言語モデルにおいて，非言語情報を考慮することで言語モデルの性能が改善されるかどうかを実験によって確かめる．

5.1 単一発話スタイルからなる中規模データにおける実験

始めに，発話スタイルが比較的均一である中規模のデータにおいて非言語情報が有効にはたらくことを確かめる．

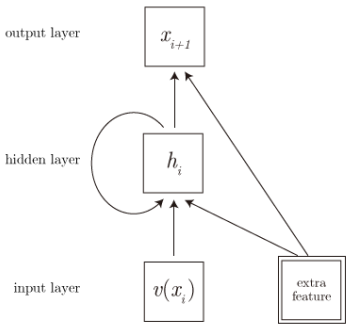

図 5　中間層と出力層に非言語情報を挿入する dual RNN 言語モデル
Fig. 5　dual RNNLM with non-verbal information

表 1　CSJ 講演データ
Table 1　CSJ academic lecture data

データ	発話数	単語数	OOVR
training	943	3,603,778	0.29%
validation	41	100,035	0.47%
eval1	10	27,651	0.93%
eval2	10	28,424	0.90%
eval3	10	18,283	0.65%

表 2　NN のハイパーパラメータ
Table 2　hyper parameter of Neural Network

パラメータ	値
次元数	200
学習率	SGD+adam
活性化関数	tanh
誤差関数	クロスエントロピー関数
バッチサイズ	32
dropout 率	0.5
イテレーション	20

表 3　講演データにおけるパープレキシティ
Table 3　perplexity of LM trained with CSJ academic lecture data

モデル	特徴量	eval1	eval2	eval3
n-gram	-	64.75	68.18	140.63
LSTMLM	-	53.17	54.53	107.32
hidden	i-vector	52.09	54.11	108.26
output	i-vector	55.50	58.90	111.01
dual	i-vector	53.97	55.47	110.53
hidden	openSMILE	**51.53**	53.49	**102.15**
output	openSMILE	53.78	55.55	108.60
dual	openSMILE	51.82	**53.39**	105.21

また，openSMILE から抽出された特徴量が非言語情報として適切であるかを確かめる．

5.1.1　実験データ

話者性が出やすいと思われる話し言葉コーパスのうち，今回の実験では日本語話し言葉コーパス (CSJ) [26] を利用した．CSJ には学会講演や模擬講演などの自由発話音声とその書き起こし文が含まれている．

この実験では，CSJ 全データのうち，学会講演のデータ (274 時間) を用いた．その詳細は**表 1** の通りである．言語モデルに登録する語彙数は 37,681 単語とし，語彙にない単語は未知語として “UNKNOWN_WORD” という単語に置換した．Out Of Vocabulary Rate (OOVR) とは，全単語のうちの未知語の占める割合を示している．

トレーニングデータのうち約 2.5%を開発データとした．テストデータは音声認識システム kaldi [27] の標準評価セットに従って 3 種類作成し，eval1 は男性による 10 講演，eval2 は男女 5 名ずつの 10 講演，eval3 は模擬講演が 10 講演が含まれている．

5.1.2　実験設定

言語モデルとしては，RNN 言語モデルの発展形である LSTM 言語モデルを用いた．NN の重みは $[-0.1, 0.1]$ の一様乱数で初期化し，誤差逆伝播法によって学習を行う．非言語情報として加える特徴量は 3 層からなる Feed Forward Neural Network によって 100 次元に圧縮され，3 種類の LSTM 言語モデルに付加される．誤差は 35 単語前まで伝搬したら打ち切りとし，逆誤差伝播法によって求めた勾配の最大値は 5 とした．なお，各モデルの実装には Chainer [*1] を利用した．また，n-gram 言語モデルの構築には srilm [*2] を利用し，Good-Tuning-discounting によるスムージングを施して 3-gram モデルを作成し比較を行った．

非言語情報としては，i-vector と openSMILE による特徴量をそれぞれの音声全体から抽出した．i-vector の抽出に用いる Universal Background Model は，トレーニングデータ中の 200 文を用いて構成した．i-vector の次元は 100 とし，抽出には kaldi を用いた．openSMILE の特徴量セットには，感情認識タスクのための特徴量セットである emobase を使い，991 次元の特徴量を得た．

音声認識では，フレーム特徴量として 13 次元の MFCC とその $\Delta, \Delta\Delta$ 特徴量を用いた．この特徴量から，kaldi のレシピに則り，LDA+MLLT+SAT 処理を施した GMM-HMM 音響モデルを構築した．また，リスコアリングの際の認識仮説数は 100 とし，式 (19) における n-gram の重み w_l を 0.25，新しい言語モデルの重み $1 - w_l$ を 0.75 とした．

5.1.3　パープレキシティによる比較

それぞれの言語モデルにおけるテストデータのパープレキシティは**表 3** のようになった．まず，n-gram モデルに対して LSTM 言語モデルはパープレキシティが小さく，言語モデルとしての性能が優れていることが確認できる．非言語情報として i-vector を付加した場合，その効果は見られなかった．一方で，openSMILE で抽出した特徴量を付

[*1] http://chainer.org/
[*2] http://www.speech.sri.com/projects/srilm/

表 4　講演データにおける WER
Table 4　WER of model trained with CSJ academic lecture data

モデル	特徴量	eval1	eval2	eval3
n-gram	-	14.55	12.58	18.09
LSTMLM	-	12.64	**11.21**	17.55
hidden	openSMILE	**12.51**	11.29	17.42
dual	openSMILE	12.69	11.35	**17.32**

加した場合は，特に深い層に入力すると言語モデルの性能が向上している．このことから，言語モデルに付加する非言語情報としては i-vector よりも openSMILE で抽出した特徴量のほうが適していると言える．

また，LSTM 言語モデルの傾向としては，output model, LSTMLM, dual model, hidden model の順にパープレキシティが小さい．output model のように出力層に非言語情報を挿入した場合，隠れ層に関係なく，常に一定のバイアスで単語の確率を考慮することになる．一方，hidden model のように隠れ層に挿入した場合は，各単語における隠れ層を変化させるため，時間に依存して話者性を考慮していると見なすことができる．今回の実験結果から，非言語情報と言語の間には時間に依存した関係があることを示していると言える．

5.1.4　WER による比較

次に，ベースラインと最もパープレキシティの良かった 2 つのモデルについて，WER についても比較を行なった．その結果を**表 4** に示す．なお，eval3 のテストデータは自由発話からなるものであり，学会講演のみからなる学習データとドメインが異なることに注意されたい．

まず，LSTMLM によるリスコアリングを行うことで WER が約 1%以上が改善している．LSTM 言語モデルに非言語情報を付加した効果については，今回の結果ではあまり一貫した傾向は見られなかった．パープレキシティの結果から言語モデルの大きな改善には結びついていないと推測でき，WER までその効果が明確に現れなかったのだと考えられる．

5.2　複数の発話スタイルを含む大規模データでの実験

5.1 節において openSMILE によって抽出された非言語情報が言語モデルに対して有効であると確認できた．そこで，学会講演だけではなく，摸擬講演や対話など，より自由発話を交えた大規模なデータについて，このモデルが有効であるかを確かめる．

5.2.1　実験データ

データセットとしては，CSJ の全データ (661 時間) を用いた (**表 5**)．この結果，後者の語彙数は 64,765 単語となった．未知語などの処理については 5.1.1 節と同様である．

5.2.2　実験設定

こちらも 5.1.2 と同様である．なお，この実験では openSMILE から得られる非言語情報のみを扱う対象とした．

表 5　CSJ 全データ
Table 5　all data of CSJ

データ	発話数	単語数	OOVR
training	3,232	7,692,758	0.20%
validation	88	234,145	0.43%
eval1	10	27,651	0.93%
eval2	10	28,424	0.90%
eval3	10	18,283	0.65%

表 6　CSJ 全データで学習した言語モデルのパープレキシティ
Table 6　perplexity of LM trained with csj all data

モデル	特徴量	eval1	eval2	eval3
n-gram	-	69.94	76.92	74.98
LSTMLM	-	53.53	57.52	58.21
hidden	opensmile	**53.24**	**56.46**	**57.90**
output	opensmile	54.43	57.13	59.28
dual	opensmile	54.10	57.53	58.21

5.2.3　パープレキシティによる比較

CSJ 全データで学習した言語モデルの性能は，講演音声における結果と同じ傾向が見られた．隠れ層に特徴量を加える hidden model では，いずれのテストデータに対してもパープレキシティが小さくなっている．一方で，出力層に特徴量を加える output model, dual model については，何もしない LSTM 言語モデルよりも劣る結果となった．これは，出力される単語と非言語情報の間に時間依存性があり，hidden model でその関係性を捉えることができていると考えられる．

6.　おわりに

本研究では，非言語情報を用いて話者性や発話環境を推定し，その情報を RNN 言語モデルに組み込むことで言語モデルをドメイン適応することを提案した．日本語話し言葉コーパスによる実験において，提案手法によってパープレキシティが改善されることを確認した．

しかし，中規模データにおける音声認識実験において，WER の改善に明確な効果が見られなかった．また，学会講演のみを学習データとした場合と，自由発話も交えて学習データとした場合の 2 種類の実験を行なったが，パープレキシティの改善率を見てみると，両者に大きな違いは無かった．そこで，今後はより話者性や発話環境などの非言語情報をクラスの形で抽出し，それぞれのクラスに対してモデルを適応させるモデルを考えたい．

その他，DNN 音響モデルを使い，リスコアリング時の重みを最適化することで，CSJ 全データにおける WER においても提案手法の有効性を示したいと考えている．また，CSJ のデータの約半分は学会講演という話者性のあまり出にくい性質の発話であるため，今後は TED など，より自

由発話に近いデータセットについても同様の評価実験を行いたい.

参考文献

[1] Tomas Mikolov, Martin Karafiát, Lukas Burget, Jan Cernockỳ, and Sanjeev Khudanpur. Recurrent neural network based language model. In *INTERSPEECH 2010, 11th Annual Conference of the International Speech Communication Association, Makuhari, Chiba, Japan, September 26-30, 2010*, pages 1045–1048, 2010.

[2] Martin Sundermeyer, Ralf Schlüter, and Hermann Ney. Lstm neural networks for language modeling. In *INTERSPEECH*, pages 194–197, 2012.

[3] Siva Reddy Gangireddy, Steve Renals, Yoshihiko Nankaku, and Akinobu Lee. Prosodically-enhanced recurrent neural network language models. In *Sixteenth Annual Conference of the International Speech Communication Association*, 2015.

[4] Quoc V Le and Tomas Mikolov. Distributed representations of sentences and documents. *arXiv preprint arXiv:1405.4053*, 2014.

[5] Yangyang Shi, Pascal Wiggers, and Catholijn M Jonker. Towards recurrent neural networks language models with linguistic and contextual features. In *INTERSPEECH*, 2012.

[6] Ryo Masumura, Hirokazu Masataki, Tomohiro Oba, Osamu Yoshioka, and Satoshi Takahashi. Use of latent words language models in asr: a sampling-based implementation. In *Acoustics, Speech and Signal Processing (ICASSP), 2013 IEEE International Conference on*, pages 8445–8449. IEEE, 2013.

[7] Aaron Heidel, Hung-an Chang, and Lin-shan Lee. Language model adaptation using latent dirichlet allocation and an efficient topic inference algorithm. In *INTERSPEECH*, pages 2361–2364, 2007.

[8] 広瀬友紀. 文理解の認知メカニズム 話者の意図と聞き手の理解: 語彙アクセントの隠れた作用. **認知科学**, 13(3):428–442, 2006.

[9] 博也 藤崎, 賢司 阿部, 一滋 黒川, 武田和也, 修一 成澤, and 澄雄 大野. 話者の心の状態遷移モデルに基づく対話音声認識. **情報処理学会研究報告音声言語情報処理（*SLP*）**, 2001(11):79–84, feb 2001.

[10] Stefan Kombrink, Tomas Mikolov, Martin Karafiát, and Lukás Burget. Recurrent neural network based language modeling in meeting recognition. In *INTERSPEECH*, pages 2877–2880, 2011.

[11] Tomas Mikolov and Geoffrey Zweig. Context dependent recurrent neural network language model. In *SLT*, pages 234–239, 2012.

[12] Daniel Renshaw and Keith B Hall. Long short-term memory language models with additive morphological features for automatic speech recognition, 2015.

[13] Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural computation*, 9(8):1735–1780, 1997.

[14] Ebru Arisoy and Murat Saraçlar. Multi-stream long short-term memory neural network language model. In *Sixteenth Annual Conference of the International Speech Communication Association*, 2015.

[15] Hori Takaaki, Hori Chiori, Watanabe Shinji, and John R. Hershey. Minimum word error training of long short-term memory recurrent neural network language models for speech recognition. In *INTERSPEECH*, 2016.

[16] X. Chen, X. Liu, Y. Qian, M.J.F. Gales, and P.C. Woodland. Cued-rnnlm an open-source toolkit for efficient training and evaluation of recurrent neural network language models. In *INTERSPEECH*, 2016.

[17] Marek Rei. Online representation learning in recurrent neural language models. *arXiv preprint arXiv:1508.03854*, 2015.

[18] Daniel Renshaw and Keith B Hall. Long short-term memory language models with additive morphological features for automatic speech recognition. In *Acoustics, Speech and Signal Processing (ICASSP)*, pages 5246–5250. IEEE, 2015.

[19] Amr El-Desoky Mousa, Ralf Schlüter, and Hermann Ney. Investigations on the use of morpheme level features in language models for arabic lvcsr. In *Acoustics, Speech and Signal Processing (ICASSP), 2012 IEEE International Conference on*, pages 5021–5024. IEEE, 2012.

[20] Tong Fu, Yang Han, Xiangang Li, Yi Liu, and Xihong Wu. Integrating prosodic information into recurrent neural network language model for speech recognition. In *2015 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA)*, pages 1194–1197. IEEE, 2015.

[21] 大毅 森, 喜久雄 前川, and 英樹 粕谷. **音声は何を伝えているか：感情・パラ言語情報・個人性の音声科学**. Number 12 in 音響サイエンスシリーズ / 日本音響学会編. コロナ社, 2014.

[22] Samira Ebrahimi Kahou, Vincent Michalski, Kishore Konda, Roland Memisevic, and Christopher Pal. Recurrent neural networks for emotion recognition in video. In *Proceedings of the 2015 ACM on International Conference on Multimodal Interaction*, pages 467–474. ACM, 2015.

[23] Florian Eyben, Felix Weninger, Florian Gross, and Björn Schuller. Recent developments in opensmile, the munich open-source multimedia feature extractor. In *Proceedings of the 21st ACM international conference on Multimedia*, pages 835–838. ACM, 2013.

[24] Najim Dehak, Pedro A Torres-Carrasquillo, Douglas A Reynolds, and Reda Dehak. Language recognition via i-vectors and dimensionality reduction. In *INTERSPEECH*, pages 857–860, 2011.

[25] 哲司 小川 and さやか 塩田. i-vector を用いた話者認識. **日本音響学会誌**, 70(6):332–339, jun 2014.

[26] Sadaoki Furui, Kikuo Maekawa, and Hitoshi Isahara. A japanese national project on spontaneous speech corpus and processing technology. In *ISCA Workshop on Automatic Speech Recognition*, pages 244–248, 2000.

[27] Daniel Povey, Arnab Ghoshal, Gilles Boulianne, Lukas Burget, Ondrej Glembek, Nagendra Goel, Mirko Hannemann, Petr Motlicek, Yanmin Qian, Petr Schwarz, Jan Silovsky, Georg Stemmer, and Karel Vesely. The kaldi speech recognition toolkit. In *IEEE 2011 Workshop on Automatic Speech Recognition and Understanding*. IEEE Signal Processing Society, December 2011. IEEE Catalog No.: CFP11SRW-USB.