2008－SLP－72 (2)
2008／7／18

confusion networkと語彙制約なし音声認識を用いた動的発音モデリング

緒方 淳　　後藤 真孝

産業技術総合研究所

{jun.ogata,m.goto}[at]aist.go.jp

あらまし　自由発話音声認識において，発音変動をいかにモデル化するか，すなわち発音モデリングは，認識性能を左右する重要な問題の１つである．発音モデリングにおいては，発音辞書に発音変動を表す音韻系列を追加していくアプローチが最も一般的である．しかし，単純に多くの発音変動を追加することは，単語間の音響的混同を引き起こし，認識性能が劣化する原因となってしまう．これに対処する手段として，本研究では，発話ごとに固有の発音辞書を生成する，動的発音モデリングに着目する．本稿では，confusion networkと語彙制約なし音声認識を利用した，新たな動的発音モデリング手法を提案する．日本語話し言葉コーパス (CSJ) を用いた自由発話連続音声認識実験により，本手法の有効性を確認した．

キーワード　自由発話，発音変動，動的発音モデリング，静的発音モデリング，音響的混同，confusion network

Dynamic Pronunciation Modeling Using Confusion Networks and Unconstrained Speech Recognition

Jun Ogata　　Masataka Goto

National Institute of Advanced Industrial Science and Technology (AIST)

Abstract　To better model pronunciation variations is an important issue in spontaneous speech recognition. The most common approach is to work at lexicon level, by simply adding phonetic sequences to a basic lexicon. However, it is also known that adding too many variants increases acoustic confusability between words and sometimes even decreases the recognition performance. To solve this problem, we propose a novel dynamic pronunciation modeling using confusion networks and unconstrained speech recognition. In our experiments using the corpus of spontaneous Japanese (CSJ), the effectiveness of the proposed method was confirmed.

Keyword　spontaneous speech, pronunciation variations, dynamic pronunciation modeling, static pronunciation modeling, acoustic confusability, confusion networks

1　まえがき

実環境において音声認識システムが幅広く利用されるようになるためには，自由発話，話し言葉を頑健に認識する技術が必要不可欠となる．自由発話は，読み上げ音声と比べ，不明瞭な発声や口語表現，言い淀み，発話速度の変動など，現状の音声認識システムでは取り扱うことが困難な様々な要因を含んでいる．このような自由発話音声に関する研究を目的とし，近年では，日本語話し言葉コーパス (CSJ)[1] やCIAIR車内音声コーパス [2] など，大規模な自由発話音声データベースが構築されており，自由発話音声認識に関する研究の進展が期待されている．

本研究では，自由発話特有の現象に対して頑健な音声認識システムを構築することを目的としている．本稿では，このような現象の１つとして**発音変動 (pronunciation variation)** を取り上げる．発音変動とは，ある単語が発声される際にその単語の「標準的な発音」とは異なった発音で発声される現象であり，自由発話音声認識において認識性能を劣化させる大きな原因となっている．このような発音変動を音声認識システムにおいてモデル化する，いわゆる**発音モデリング (pronunciation modeling)** に関する研究はこれまでにも様々な検討がなされており [3]，主として，音響モデルにおいて個々の音韻内で発生する変動を考慮するアプローチ，単語発音辞書において音韻レベルのシンボリックな系列の並びによって発音変動をモデル

化するアプローチ，に大きく分類される．ここでは，音韻間で発生するドラスティックな変動(音韻の消失，置換，挿入など)に対処するため，後者のアプローチに着目する．本稿では以降，便宜上，「発音モデリング」とは基本的に後者のアプローチのことを指す．

発音変動を考慮して単語発音辞書を生成するアプローチとしては，これまでにも多くの手法が提案されている[4]-[7]．従来の手法では，基本的に，発音変動を考慮した形で単一の(グローバルな)単語発音辞書を認識前にあらかじめ生成しておき，それによってあらゆる入力発話に対して認識を行うことを前提としている(**静的発音モデリング**)．しかしながら，自由発話音声，あるいは実環境での音声における発音変動は，話者の違い(発音上の癖など)，話者の状態，周囲の環境など，様々な要因によって発生するため，そもそも単一の単語発音辞書内で全ての変動をカバーしきれないという問題がある．また，たとえ同一話者であっても，発音変動のパターンは，上記のような要因に応じて，時々刻々と(発話ごとに)変化していく．

したがって，各入力発話ごとの発音変動のモデル化が可能であれば，より頑健な音声認識が実現できると考えられる．これは，各入力発話ごとに発話固有の単語発音辞書を生成することから，本研究では**動的発音モデリング**と呼ぶ．これと同様の考えに基づいた手法は，あまり多くはないが，これまでにもいくつか検討がなされている[8]-[11]．Fosler-Lussierは，音声認識の中間結果(N-bestリスト，あるいはワードグラフ)中の各単語候補に対して，決定木を用いた発音変動パターンの展開を行い，得られた発音ネットワーク上を再認識(リスコア)する手法を提案している[8]．Willettらは，認識結果に対して，日本語の単純な発音変動ルール(5種類)を段階的に適用した，マルチパス音声認識手法を提案している[9]．ただし，これらの手法では，後述する音響的混同の影響もあり，大きな性能改善には至っていない．一方，Leeらは，発話ごとに音韻特徴(phonetic features)をニューラルネットワークによって検出し，その結果に基づいて発音変動パターンを選択し，発話固有の単語発音辞書を生成する手法を提案している[10]．従来の静的発音モデリングと比較して大きな改善が得られているが，この手法は読み上げ音声(TIMIT音声データベース)を対象にしており，中語彙での評価に留まっている．また，自由発話音声，あるいは雑音環境下音声においては，読み上げ音声同様に音韻特徴を頑健に検出することは困難であると考えられる．

本稿では，confusion networkと語彙制約なし音声認識を利用した動的発音モデリング手法を提案する．提案手法では，まず入力発話に対して通常の音声認識を行い，confusion networkを生成する．次に，あらかじめ学習しておいた発音変動パターンのリスト(静的発音辞書)を用いて，confusion network中の各単語候補に対して，発音変動パターンを展開することで，発話固有の動的発音辞書を生成する．このとき，発音変動パターンの抽出には，汎用性も考慮し，語彙制約なし音声認識を用いる．confusion networkを利用することで，動的発音モデリングにおいて従来利用されてきたN-bestリスト，ワードグラフと比較して，発音変動モデル化の対象とすべき単語候補を大幅に絞り込むことができる．

2　発音モデリングにおける問題

ここでは，一般的な発音モデリングにおける問題点，課題について考察し，動的発音モデリングの重要性について述べる．

2.1　音響的混同

一般的に発音モデリングは，自然発話中の様々な発音変動パターン(音韻系列)を，単語発音辞書においてなるべく多くカバーすることを目的としたものである．しかし，多くの発音変動がカバーされる一方で，異なる単語間においてそれらの音韻系列が近くなることで，**音響的混同**(acoustic confusability)が高くなるという問題が発生する[3]．音響的混同は，音声認識の性能に大きく影響することが知られており，変動パターンの音韻系列を単純に追加するだけでは，ベースラインよりも認識率が大きく下がることが報告されている[12]．とくに日本語においては，欧米の言語と比べて，単語(形態素)の長さ(1単語あたりの音韻数)が比較的短いため，発音モデリングを行う際に音響的混同の影響が大きくなると考えられる．

したがって，単一の単語発音辞書を認識前にあらかじめ生成する静的発音モデリングにおいては，辞書に追加すべき発音変動パターンを取捨選択することが必要不可欠となり，そのための選択基準に関する検討がなされてきた[3]．しかしながら，発音変動のカバレージと音響的混同はトレードオフの関係にあり，両者の間で音声認識システムとして最適なモデル化を行うことは難しい．

2.2　動的発音モデリング

動的発音モデリングは，静的発音モデリングにおいて発生する音響的混同を抑えるための方法と位置づけられる．すなわち，静的発音モデリングのように，事

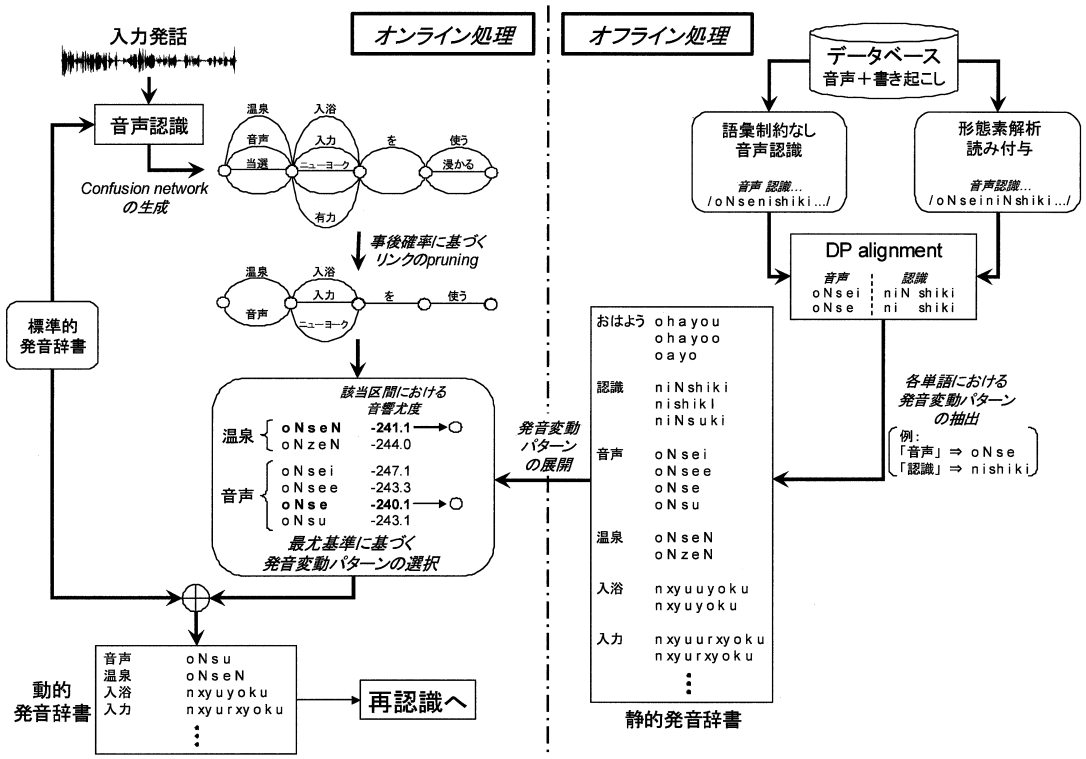

図 1: 提案する動的発音モデリングの概要 (音素表記のうち「xy」は拗音を,「N」は撥音をそれぞれ示している)

前に登録されている全ての単語を対象として，それらの発音変動パターンを単語発音辞書で考慮するのではなく，認識対象の入力発話に対して，まず何らかの形で発話中に出現する単語を推定することで，発音変動モデル化の対象となる単語を制限する．このとき，入力発話中の出現単語の推定には，標準的な単語発音辞書を用いた音声認識が一般的に利用される．その発話において対象外となった単語に関しては，標準的な発音のみが認識時に展開されるので，異なる単語間での音響的混同は起こりにくくなる．ただし，入力発話中の単語を推定する際に，本来正解であるはずの単語が候補から漏れた場合には，その単語に関しては発音変動モデル化の対象にはならない．従来の動的発音モデリングにおいては，このような入力発話中の出現単語の推定には，N-best リストや単語グラフ [8][11] が用いられている．しかしながら，大語彙連続音声認識においては，これらの中間結果に本来の正解単語をより多く含めるためには，それ自体のサイズをかなり大規模なものにする必要があり，動的発音モデリングにおいて音響的混同を効果的に抑えることは難しいと考えられる．

3 confusion network と語彙制約なし音声認識に基づく動的発音モデリング

本研究では，confusion network[13] を用いた新たな動的発音モデリング手法を提案する．confusion network は，単語グラフを圧縮することで，単語レベルのリニアなネットワークで表現した効率的な中間結果である．単語グラフと比較して，単語候補数が大幅に削減されるにも関わらず，本来の正解単語を落とすことなく，精度を保持することが可能である [13]．したがって，動的発音モデリングにおいても，従来の N-best リストや単語グラフを利用した手法に比べて，前節で述べた音響的混同を効果的に抑えることが可能であると考えられる．

本研究で提案する動的発音モデリングの概要を図 1 に示す．提案手法は大きく,「オフライン処理」と「オンライン処理」に分かれる．以下，本手法の各ポイントを順に説明していく．

3.1 静的発音辞書の生成

まず，オフライン処理 (図 1 右側) として，静的発音辞書を生成する．ここでの目的は，学習用データベース中に出現する各単語について，その発音変動パターンをモデル化することである．対象とするデータベースとしては，音声データとその書き起こしが存在することを前提としている (ただし，発音レベルまで忠実に再現した書き起こしは必ずしも必要でない)．まず，データベース中の音声データに対して語彙制約なし音声認識 (ここでは連続音素認識) を行い，得られた音素列と，形態素辞書の読み情報により付与された，標準的な発音を表す音素列との間で，DP によるアラインメントを行う．これにより，各単語における発音変動パターンが抽出でき，図 1 に示すような静的発音辞書が生成される．

以上の手法は，「語彙に依存した発音モデリング」であり，基本的にデータベース中に出現しない単語には適用できないなど，汎用性の面で問題がある．本研究では，第 1 ステップとしてこのようなシンプルな手法を静的発音辞書生成に利用したが，原理的には，文献 [4],[5] で提案されているような「語彙に依存しない発音モデリング」も適用可能である．

3.2 confusion network の生成と pruning

次に，オンライン処理の説明に移る．まず，入力発話に対して，標準的発音辞書 (形態素辞書の読み情報により付与される発音) を利用して音声認識を行い，confusion network を生成する．confusion network は，単語グラフをもとに，2 段階の音響的クラスタリングを行うことで求められる [13]．confusion network の各リンクには，クラスタリングした各クラス (単語の区間) ごとに事後確率が算出され，それらの値は，各区間での単語の存在確率を示す．また，次節での処理のため，network 中の各リンクにおける時間情報 (入力発話に対する各単語の開始時間，終了時間) を保持しておく．ここでは，単語グラフ中に記録された各単語の時間情報を利用する．

本研究では，発音モデリングにおける音響的混同の削減を目的とし，confusion network の更なる圧縮を試みる．ここでは，上記事後確率値に基づき，network 上のリンク (単語) の pruning を行う．具体的には，確率値の閾値を 2 種類設定し (th_{max}, th_{min})，あるリンク l_i の事後確率 $post(l_i)$ が，以下の条件を満たすリンクのみ残す (以降で発音モデリングの対象とする)．

$$th_{min} < post(l_i) < th_{max}$$

3.3 最尤基準に基づく発音変動パターンの選択

pruning 後の confusion network 中の各単語に対して，オフライン処理で求めた静的発音辞書中の発音変動パターンを展開する．ここで，各単語に対する，静的発音辞書中の複数存在する発音変動パターンの中から，その発話固有の発音変動パターンを最尤基準により選択する．具体的には，confusion network 中の各単語の区間 (前節で求めた時間情報を利用) に対応する入力発話中の音響信号に対して，各発音変動パターンの音素系列ごとの音響尤度を計算し，最尤のパターンを選択する．

3.4 動的発音辞書の生成

最後に，各単語ごとに選択された発音変動パターンと，標準的発音辞書中の発音パターンを混合することによって，発話固有の動的発音辞書を生成する．この発音辞書を用いて，同じ入力発話を再度認識し，認識結果を出力する．

4 実験と考察

提案手法を，日本語話し言葉コーパス (CSJ) を用いた実験により評価を行った．

4.1 実験条件

評価用音声データとしては，CSJ 中の男性話者の 6 講演を利用した．また，3.2 節で述べた pruning の閾値や，音声認識器のデコーディングパラメータなどの最適値を実験的に求めるために，development セットとして上記データとは別の 2 講演を用意した．

ベースラインとなる音声認識システムについても，CSJ を用いて学習している．音響モデルには，単語間の調音結合も考慮した状態共有型 triphone モデルを用いた．音素体系は，無音，ショートポーズも含めた 30 音素である．言語モデルは，CSJ の 2670 講演の書き起こしから学習した 2-gram モデルを用いた．書き起こしはすべて，茶筅 Ver.2.4.2 と IPADIC 2.7.0 を利用して形態素解析した．作成した言語モデルの語彙サイズは 28488 である．

4.2 静的発音辞書

3.1 節で述べた静的発音辞書は，CSJ の講演音声 871122 発話から学習した．上記音声データすべてに対して，語彙制約なし音声認識を行い，標準的発音による音素列との間でアラインメントをとることにより，

表 1: 各手法における単語誤り率

	単語誤り率 (%)
ベースライン	30.95
提案手法 (ALL)	31.18
提案手法 (ML)	**29.19**

各単語の発音変動パターンを得た．この際，アラインメントの誤りや極端に低頻度の発音変動の影響を除去するために，発音変動パターンの頻度 (本実験では 2 に設定) によりカットオフを行った．その結果，75228 単語に対して合計 131028 個の発音変動パターンを得た．

4.3 実験結果

提案手法を用いたときの認識性能 (単語誤り率) を表 1 に示す．表中，提案手法のうち「ALL」は各単語の発音変動パターンを全て展開した場合，「ML」は 3.3 節で述べたように，最尤基準で発音変動パターンを選択した場合をそれぞれ示す．また，本実験では，confusion network の pruning 時の閾値 th_{max}, th_{min} は，それぞれ 0.9，0.1 に設定した．

結果より，提案手法において，最尤基準で発音変動パターンを選択した場合，ベースラインより認識性能は向上し，単語誤り率は絶対値で 1.76%削減された．これにより，提案した最尤基準の選択方法により，発話固有の発音変動パターンが適切にモデル化できたといえる．一方，発音変動パターンを全て展開した場合は，動的発音モデリングによって，ベースラインよりもわずかに認識性能が劣化した．これは，提案手法の動的発音モデリングでは confusion network を利用することで，発音モデリング対象の単語の絞り込みが可能であるものの，同一単語で複数登録された発音変動パターンにより音響的混同の影響を回避できなかったことが原因と考えられる．

本研究では，発音変動パターンの生成手法としては，主として語彙制約なし音声認識を利用した．しかし，本実験で評価に用いている CSJ においては，人手によって作成された，発音もできるだけ忠実に再現した書き起こしが存在する (これを「音韻トランスクリプション」と呼ぶ)．ここでは，提案手法の枠組みにおいて，語彙制約なし音声認識と音韻トランスクリプションの，発音変動パターン生成手法としての性能比較を行った．実験結果を表 2 に示す．ここで，各単語の発音変動パターンの選択は最尤基準で行っている．結果より，語彙制約なし音声認識を用いた場合の方が，良い認識性能を示していることがわかる．この理由としては，語彙制約なし音声認識を用いた場合，発音変動だけでなく，音響モデルにおける音響的誤り傾向が，単語の発音パターンとして考慮されるからと考えられる．

表 2: 発音変動パターン生成手法の比較

	単語誤り率 (%)
音韻トランスクリプション	29.50
語彙制約なし音声認識	**29.19**

5 おわりに

本稿では，自由発話音声認識おける発音変動に対処するために，confusion network と語彙制約なし音声認識に基づく動的発音モデリング手法を提案した．confusion network による効率的な中間結果と，最尤基準に基づく発音変動パターンの選択によって，効果的な動的発音モデリングが実現できた．また，動的発音モデリングにおける，発音変動パターン生成手法として，語彙制約なし音声認識を用いることの有効性を示した．

今後の課題としては，静的発音辞書生成部の改良，特に語彙に依存しない手法の適用を検討していく．また，より大規模な実験，他のタスクでの評価などが挙げられる．

謝辞

本研究の一部は，科研費 (19300065) の助成を受けた．

参考文献

[1] 河原達也: “『日本語話し言葉コーパス』を用いた音声認識の進展”, 第 3 回話し言葉の科学と工学ワークショップ講演予稿集, pp.61-66, 2004.

[2] K. Takeda, H. Fujimura, K. Itou, N. Kawaguchi, S. Matsubara and F. Itakura: “Construction and Evaluation of a Large in-car Speech Corpus, IEICE Transactions on Information and Systems, Vol. E88-D, No. 3, pp.553-561, 2005.

[3] H. Strik and C. Cucchiarini: “Modeling pronunciation variation for ASR: a survey of the literature”, Speech Communication, Vol.29, No.24, pp.225-246, 1999.

[4] 秋田裕哉, 河原達也: “話し言葉音声認識のための汎用的な統計的発音変動モデル”, 信学論 (D-II), Vol.J88-D-II, No.9, pp.1780-1789, 2005.

[5] 深田俊明, 匂坂芳典: “発音ネットワークに基づく発音辞書の自動生成”, 信学論 (D-II), Vol.J80-D-II, No.10, pp.2626-2635, 1997.

[6] T. Sloboda and A. Waibel: “Dictionary learning for spontaneous speech recognition”, In Proc. of ICSLP96, pp.2328-2331, 1996.

[7] L. Lamel and G. Adda: “On designing pronunciation lexicons for large vocabulary, continuous speech recognition”, In Proc. of ICSLP96, pp.6-9, 1996.

[8] J.E. Fosler-Lussier: “Dynamic pronunciation models for automatic speech recognition”, PhD thesis, University of California, Berkeley, 1999.

[9] D. Willett, E. McDermott, S. Katagiri: “Unsupervised pronunciation adaptation for off-line transcription of Japanese lecture speeches”, In Proc. of PMLA2002, pp.36-41, 2002.

[10] K-T. Lee, C.J. Wellekens: “Dynamic lexicon using phonetic features”, In Proc. of EuroSpeech 2001, 2001.

[11] M. Finke and A. Waibel: “Speaking mode dependent pronunciation modeling in large vocabulary conversational speech recognition”, In Proc. of EuroSpeech 97, 97.

[12] 南條浩輝, 河原達也, 山田篤, 内元清貴: “講演音声認識のための言語モデルの教師なし適応”, 情処研報, Vol.2002. No.121, pp.189-194, 2002.

[13] L.Mangu, E.Brill and A.Stolcke: “Finding Consensus in Speech Recognition: Word Error Minimization and Other Applications of Confusion Network” Computer Speech and Language,Vol.14, No.4, pp.373-400, 2000.