社団法人　情報処理学会　研究報告
IPSJ SIG Technical Reports

2005－ＳＬＰ－59（６）
2005／12／21

実況・対談における発声変形を考慮した音響モデルの検討

佐藤　庄衛†　尾上　和穂†　本間　真一†　小林　彰夫†　今井　亨†

† NHK 放送技術研究所　〒157-8510 東京都世田谷区砧 1-10-11
E-mail: †{satou.s-gu,onoe.k-ec,homma.s-fc,kobayashi.a-fs,imai.t-mq}@nhk.or.jp

あらまし　本稿では，放送音声の認識率の低下の要因の一つである発話スタイルの影響を改善するため，学習音声中の発声変形頻度を基準とする発声変形用音響モデルの導入法を検討した．発声変形を取り扱う単位として，トライフォン単位と単語単位，およびこれらの単位間で HMM を共有する場合と独立した HMM を用いる場合を取り上げ，次の二つのタスクにおいて，誤認識単語の削減効果を比較した．第一のタスクは，メタデータ制作を目的としたＪリーグ中継の実況音声の認識であり，絶叫発話と話速の速い部分を対象とし，発声変形の単位を単語として HMM を共有した場合に，キーワード誤りの 15%が削減された．第二のタスクはニュース番組中の対談部分の認識であり，話速の速い部分に起こる発声変形を対象とした．認識実験の結果，発声変形の単位を単語とし，HMM を共有しない場合が，認識率と探索空間の面から有利であることが確認された．一方，発声変形とみなす単位を単語とした場合の実験結果から，発声変形に単語依存性があることが確認され，HMM を共有した場合に認識誤りの 10%が削減された．

キーワード　音声認識，音響モデル，対談，発声変形，話速

Acoustic models for utterance variation in broadcast commentary and conversation

Shoei SATO†, Kazuo ONOE†, Shinich HOMMA†, Akio KOBAYASHI†, and Toru IMAI†

† NHK Science and Techinical Research Laboratories　1-10-11 Kinuta Setagaya, Tokyo, 157-8510 Japan
E-mail: †{satou.s-gu,onoe.k-ec,homma.s-fc,kobayashi.a-fs,imai.t-mq}@nhk.or.jp

Abstract　This paper investigates acoustic models for utterance variation and their units to improve recognition performance. Using a phoneme or a word as a unit of the variation, the models for utterance variation were trained for frequently observable variation. The variation models sharing HMMs among these models were also examined. The models trained for excitedly uttered words reduced 15% of key word erorrs in sports commentary. The models trained for rapidly pronounced reduced words 10% of word errors in conversational news. The results also showed dependence of the variation upon words.

Key words　Speech recognition, Acoustic model, Conversation, Utterance variation, Speech rate

1. まえがき

NHK では，ニュース番組への字幕付与とスポーツ中継番組でのメタデータ（番組関連情報）の効率的な制作を目的として，音声認識精度の改善を図ってきた．これまでの自動字幕付与を目的とした音声認識は，スタジオのアナウンサーによる原稿読み上げ部分において十分な認識精度が得られているが，中継現場からのリポート部分や対談部分では，背景雑音や発話スタイルによる認識率の低下が問題となっている [1]．また，効率的な制作が課題であるスポーツ中継などの生放送番組において，種々の自動認識結果 [2] [3] を統合してメタデータを制作する方法 [4] を検討している．しかし，ハイライトシーンや実況部分でのアナウンサーの興奮や，発話速度などの発話スタイルによる音声認識率の低下などにより，イベント抽出精度の低下が問題になっている [3]．

本稿では，発話スタイルによる発声変形に着目し，これらの変形部分での認識精度の改善を図る．発声変形に着目した先行研究には，本稿で取り上げたメタデータ制作と同様の目的で，野球の実況中継の認識に音声興奮音響モデルを導入した状態推定音声認識 [5] があり，発話を単位として興奮音響モデルを学習して，通常または興奮状態を推定しながら認識を行うものがある．また，発話速度を考慮したものには，品詞をもとに決定した話速種別を音素環境として音響モデルを学習した例や [6]，発話速度を隠れ変数としてベイジアンネットを学習した例 [7]

表 1　絶叫候補単語とその頻度

	絶叫頻度	総出現頻度	単語
1	115	227	だ
2	80	221	シュート
3	64	1,644	て
4	56	2,934	の
5	50	707	と
6	35	102	アアッ
7	34	1,807	に
8	32	136	チャンス
⋮			

表 2　絶叫候補単語に割り当てた音素

単語	通常発声	絶叫発声
だ	d a	Ed Ea
シュート	sh u: t o	Esh Eu: Et Eo
て	t e	Et Ec
の	n o	En Eo
と	t o	Et Eo
アアッ	a: Q	Ea: EQ
に	n i	En Ei
チャンス	ch a N s (u)	Ech Ea EN Es (Eu)
⋮		

などがあり，これらはそれぞれ，品詞環境または音素を単位として発声変形用の音響モデルを導入したものである．本稿では，このような発声変形の単語依存性を調べ，どのような単位で発声変形を導入すべきかを検討するために，スポーツ実況音声とニュース音声において，興奮発話用と話速の速い部分用の音響モデルを導入して比較した．

2. スポーツ実況用発声変形モデル

筆者らは，Jリーグ中継のアナウンサーの実況音声の認識に際して，発声変形音素を導入してイベント検出精度の改善を検討している[3]．以下，発声変形音素の導入法について説明する．

スポーツ実況において，アナウンサーの実況音声には，シュートや得点などのハイライトシーンの抽出に有用な情報が多く含まれている．そのため，実況音声中のイベントに関連したキーワードの認識精度の向上により，ハイライトシーンの検出精度の向上が期待される．しかし，ハイライトシーンとして重要な得点シーンなどでは，実況アナウンサーが絶叫（興奮した発話）する傾向がある．さらに，試合進行の抽出の手がかりとなる実況文（試合の状況を説明してる発話）では，解説文（試合をわかりやすく再度説明している発話）に比べて話速が速くなる傾向がある．

これらの発声変形に対応すべく，本章では次の二種の発声変形モデルを導入する．

2.1 絶叫音響モデル

Jリーグ中継の実況音声の絶叫部分用の音響モデルを構築するにあたり，音響モデルの学習用音声から，人手により主観的な絶叫部分の抽出を行った．ここでは，次の3つの理由から絶叫用音響モデルの導入単位には単語を選び，絶叫部分を抽出した．

（1） 一文すべてが絶叫されるのはまれである．

（2） 絶叫されている音素を人手により特定するのは困難である．(子音での絶叫による変形の有無の判定が困難)．

（3） “シュート” などイベント抽出に有用なキーワードの絶叫が多い．

表1は主観的に抽出された絶叫単語の例であり，絶叫された頻度順に絶叫頻度と総出現頻度を示したものである．この絶叫頻度を基準として絶叫候補単語を決定し，通常発声の音素に加え，表2に例示するように，絶叫発声用の音素記号を追加し，対応するトライフォンHMMを絶叫発声区間から学習する．ここで，表1の “シュート” の例では，学習音声中の80サンプルから絶叫発声HMMを適応学習し，残りの141サンプルから通常発声HMMを適応学習することになる．

ここで学習された絶叫発声HMMは，主観的に選択された絶叫発声部分から学習されたものであり，絶叫時の音響特徴量の違いを基準にしたものではない．そこで，絶叫時特有の音響特徴量をモデルに反映させるため，通常発声HMMと絶叫発声HMMを用いて，学習音声中の絶叫候補単語を再度クラスタリングした結果から，自動分類に基づく絶叫発声HMMを学習する．

2.2 速い話速用音響モデル

話速が速くなる部分に起こる発声変形をモデル化するため，学習音声の正解音素列に対する強制アライメントの結果から，自己ループを一度も通らないHMMパス（最短通過）の頻度を調べ，最短通過頻度の高いトライフォンまたは単語について，発声変形HMMを学習する．ここでは，発声変形の導入単位を調べるため，図1に示すとおり，3通りの発声変形対策を比較した．図中の左に学習音声の最短通過音素を斜字体で示し，中央に発声変形トライフォンHMM，右に認識時に用いる発音辞書を示している．

図1の上段に示す第一の導入単位はHMMつまり，共有されているHMMを含んだトライフォン単位である．学習音声の各HMMの最短通過頻度を求め，頻度の高いHMMに対して発声変形HMMを追加した．この方法では，比較的少数のHMMを追加するだけで，辞書中のより多くの単語に対して発声変形HMMを導入できる利点があるが，認識時には比較的大きな探索空間が必要となる．本稿では，単語内の複数の発声変形トライフォンのコンビネーションによる探索空間の増大を避けるため，全ての発声変形トライフォンを同時に用いた一種類のエントリのみを追加した．この発声変形HMMは，入力音素が最短通過フレーム長以下である場合にも，最短通過フレーム数分の特徴ベクトルを取り込んでしまうため，周囲の音素の特徴もモデルに学習されていると考えられる．このような不整合は，発声変形部分の前後のHMMで問題になる可能性がある．一方，特定の単語末尾に発声変形が起こりやすいなど，単語や単語内の位置に依存した発声変形が起こっている場合，これらの依存性を考慮せずにトライフォンのバリエーションを考える（より

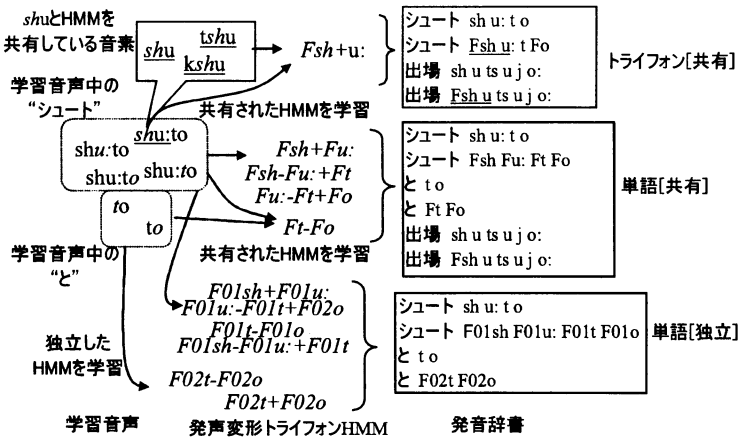

図 1　モデルの概要

表 3　発声変形 HMM を用いた認識結果

	WER[%]	KER[%]
baseline	18.4	12.9
絶叫 (主観分類)	18.0	13.5
絶叫 (自動分類)	17.9	10.9
話速 (トライフォン [共有])	18.9	12.7
話速 (単語 [共有])	17.6	12.0
話速 (単語 [独立])	17.9	11.7
絶叫+話速	17.5	11.1

制約の少ないモデルを導入する）ことになるため，新たな誤認識を引き起こす可能性もある．

図 1 の下段に示す第二の導入単位は単語であり，単語間の発声変形 HMM の共有を許さず，独立した HMM を用いたものである．ここでは，最短通過を含む頻度が高い単語に対して，その単語専用のトライフォン HMM を学習し，単語 HMM としたものである．これにより，単語に依存した発声変形をモデル化できるようになり，発声変形部分の周囲のモデルとの整合性も向上する．しかし，学習音声の確保が難しくなるため，発声変形を考慮できる単語は限られると同時に，多量の HMM を追加しなければならないという問題がある．

図 1 の中段に示した第三の導入単位は前述と同様に単語であるが，単語間の発声変形 HMM の共有を許したものである．導入する HMM の単語依存性をなくすことで，音響モデル規模の増大を軽減するとともに，学習データのスパースネスを軽減して，より多くの単語に対する発声変形を考慮できるようにしたものである．

2.3 認識実験

上述の発声変形音響モデルの有効性を調べるために行った認識実験を以下に説明する．評価音声には，NHK が 2004 年 12 月 11 日に放送した J リーグ中継（浦和レッズ vs 横浜マリノス）の後半の実況アナウンサーの音声を用いた．この音声は実況アナウンサーのマイク出力を直接収録したもので，放送音声に比べると背景雑音の影響は少ない．本実験では認識結果を単語誤認識率（WER）とキーワードの誤認識率（KER）を用いて評価する．ここで，キーワードはメタデータの抽出に重要になると思われる “シュート” や “オフサイド” などの J リーグ・サッカー用語 169 単語と評価試合の出場選手名 47 単語とした．評価音声中の総単語数は 5,307 単語，キーワード総数は 551 単語であった．認識実験に用いた言語モデルは，スポーツニュース，J リーグ中継，ワールドカップ中継の書き起こしを用いて，チーム名，選手名，地名などを評価試合の名称に置換して学習した適応化言語モデルであり，トライグラムのテストセットパープレキシティーは 96，未知語率は 3.2%であった．音響モデルの学習音声には，評価音声と同じ条件で収録した約 4 試合分の J リーグ中継の実況音声（4.5 時間）を用いた．本実験では，ニュース番組音声 473 時間から学習した 5 状態 3 ループ，総計 2,000 状態 8 混合分布トライフォン HMM を MLLR，MAP で適応化して HMM を学習した．以下，発声変形用の HMM を追加せずに学習した音響モデルを baseline として比較する．

絶叫用 HMM は，学習音声中の絶叫頻度が 10 以上の 27 単語を対象として作成した 86 個のトライフォン HMM である．この絶叫用 HMM の学習に用いた延べ音素数は，主観的な分類結果を用いた場合には 2,020 であったが，上述の 27 単語に対して，絶叫または通常発声の自動クラスタリングと HMM の学習を 5 回繰り返したことにより，最終的には，13,963 となった．表 3 上段は，主観分類と自動分類の結果から絶叫用 HMM を学習した場合の認識結果である．主観分類と自動分類を比較すると，WER と KER ともに自動分類に大きな改善が見られた．自動分類の結果を見ると，認識単語全体では本モデルにより認識誤りの 3%が削減されるのに対し，キーワードでは 15%の誤り削減率になっている．これは，キーワードが絶叫される頻度が高いためであり，イベント抽出を目的とした本実験条件においては，単語を単位とした絶叫用 HMM の導入が有効であると考えられる．

速い話速用 HMM は，上述の baseline 音響モデルを用いた学習音声のアライメント結果から，トライフォンと単語を単位とした場合の両者において，学習サンプル数が 30 以上になるよ

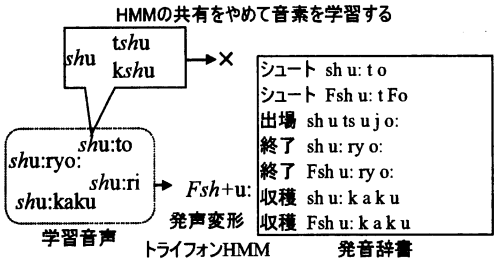

図 2 トライフォンを単位とし，独立した発声変形 HMM を導入した場合

うに発声変形単位を選択し，発声変形 HMM を学習した．表 3 中段に，トライフォン [共有]，単語 [共有]，単語 [独立] を単位とした 3 種類の速い話速用 HMM を導入した場合の認識結果を示す．トライフォンを単位とした場合には，227 の発声変形 HMM が追加され，KER に若干の改善が見られたが，WER は増加した．単語を単位とし，HMM を共有した場合，46 単語に対して 125 個の発声変形 HMM が学習された．この発声変形 HMM により，認識誤りの 4%とキーワード誤りの 7%が削減された．単語を単位とし，独立した HMM を用いた場合には，上記の 46 単語に対して，160 の発声変形 HMM が学習された．単語 [独立] では単語 [共有] に比べて，KER の改善が見られるが，WER の改善は単語 [共有] に比べて小さかった．以上のことから，一部のキーワードにおいては単語に依存した発声変形が起こっているものと考えられる．

表中の “絶叫+話速” は，自動分類の結果から学習した絶叫用 HMM を導入した後，単語 [共有] を単位として速い話速用の HMM を導入したもので，絶叫用 27 単語 86HMM に加え，43 単語 138HMM が追加されたものである．ここで，絶叫かつ速い話速用の HMM は 8 単語の 20HMM であった．両者の併用により 5%の WER 削減率が得られたが，KER の削減は “絶叫（自動分類）” を単独で用いた場合には及ばなかった．

この “絶叫（自動分類）” での認識結果を用いて，自然言語処理を用いたイベント抽出 [8] を行ったところ，シュートシーンの検出精度が適合率で 18%(6/11→8/11)，再現率で 17%(6/12→8/12) の改善が得られることを確認した．

3. ニュース音声認識用発声変形モデル

本章では，大規模な学習音声を用いた場合の発声変形モデルの導入単位を検討するために，自動字幕付与を目的としたニュース音声認識への発声変形モデルの導入実験を行った．

3.1 実験条件

評価音声は NHK が 2004 年 7 月 1 日から 15 日に放送したニュース番組のうち，現場リポートや対談などによる認識率の低下が見られた 16 項目，643 発話，8448 単語（All）および，そのサブセットである対話調の 231 発話，3247 単語（Spon）である．この評価セットには多くの不要語や言いよどみが含まれている．この言いよどみの中には，発話内容が不明な部分があるため，認識結果の正確な評価は難しい．また，字幕付与の観点からは，言いよどみ部分の認識結果は人手で削除されるべきものであり，たとえ正しく認識できたとしても挿入誤りとなる．さらに，挿入誤りの削除は置換・脱落誤りの修正に比べて修正コストが小さいことを考慮して，本章での評価は，不要語や言いよどみを除いたリファレンスに対する置換・脱落誤りから求めた単語誤認識率（WER）を用いる．発話変形を考慮しない場合のベースラインの WER は All で 6.1%，Spon で 9.7%であり，誤認識総数は All で 512 単語，Spon で 315 単語である．正解音素列で評価音声の強制アライメントをとった結果を用いて，上記の誤認識単語中の HMM の最短通過の有無を調べたところ，All で 47 単語，Spon で 31 単語に最短通過が含まれていた．さらに，話速が速く HMM 長より短い入力音素によって，誤りが前後の単語に及ぶと考えた場合，All で 78 単語，Spon で 49 単語の誤りが最短通過 HMM 周辺にある．本章では，これらの認識誤りの改善を期待して，2.2 章と同様の手法で速い話速用の HMM を導入する．さらに本章では，図 2 に示すように，トライフォン [共有] 単位と単語 [共有] 単位との発声変形単位の中間的な導入単位として，トライフォンを発声変形単位として，トライフォン間の HMM の共有を許さないトライフォン [独立] を考え，比較検討した．

音響モデルの学習音声は 2002 年 1 月から 2004 年 6 月に放送されたニュース番組から，約 331 時間分の男性発話を収集したものである．この学習音声から，ベースラインの音響モデル (5 状態 3 ループ，4000 状態 16 混合トライフォン HMM) を学習し，正解音素列でのアライメント結果から，HMM の最短通過部分を検出し，トライフォンと単語を発声変形の単位とし，それぞれ発声変形単位間の HMM を共有する発声変形 HMM と独立した発声変形 HMM を，ML 推定を 5 回繰り返して学習した．実験は発声変形単位の導入に必要な最短通過の頻度の下限を変化させて，発声変形 HMM の総数を変えながら WER を測定した．

認識に用いた言語モデルは 400 万文，1 億単語のニュース原稿をもとに，該当期間中に入稿された記者原稿 28,000 文で適応化した適応化言語モデル [9] である．この言語モデルでは，連接する頻度の高い単語から作成した複合語が一部に用いられている．評価音声に対するトライグラムパープレキシティーは 21.9，未知語率は 0.57%であった．

3.2 認識結果

3.2.1 評価音声全体 (All)

図 3 はトライフォン (tri.) を発話変形の単位とした場合の認識結果である．横軸に学習音声中の最短通過頻度の下限を示し，左軸に単語誤認識率（WER），右軸に追加された発声変形用の HMM の総数（Δhmm）を示す．最短通過頻度の下限が大きい場合には，発声変形用 HMM に十分な量の学習音声を確保できるが，下限が小さい場合には過学習の傾向になる．

ここで，HMM 共有 [tied] と HMM 独立 [untied] を比較すると，HMM 共有の場合には下限が 1500 付近で WER が増加したり，下限が小さい場合（500 以下）での WER の増加が見られるなど，安定した改善が見られず，HMM 独立とする方が改善率と安定性の面から有利であることがわかる．最短通過頻度

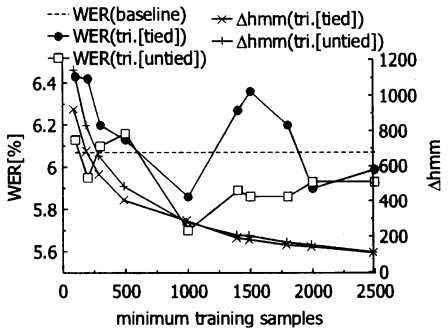

図 3　頻度下限による単語誤認識率と追加された発声変形用 HMM 数 (All) トライフォン:tri. HMM 共有:[tied] HMM 独立 [untied]

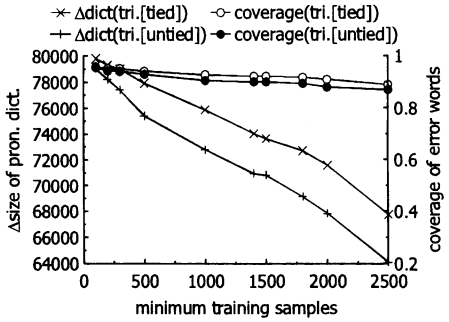

図 4　頻度下限による発音辞書のエントリの増分と誤認識単語のカバー率 (All) トライフォン:tri HMM 共有:[tied] HMM 独立 [untied]

の下限が小さい場合（1000 以下）では，両者の Δhmm の差が大きくなり，モデルの規模の観点からも HMM 独立の場合が有利であることがわかる．図 4 は図 3 に対応して，発声変形モデルの導入による発音辞書の変化を示したものである．左軸に発音辞書のエントリーの増分，つまり発声変形モデルが適用される単語数を示し，右軸に，この発声変形 HMM が適用された単語のベースラインでの認識誤りのカバー率を示す．この図から，HMM 共有の場合に比べ，HMM 独立とすることで，辞書の増分を 1 割程度小さくできるため，探索空間の観点からも，HMM 独立とした場合が有利であることが示される．両者の誤り単語のカバー率は，どの条件でも 9 割程度であった．

一方，図 5 は単語 (wrd.) を発声変形の単位とした場合の認識結果を図 3 と同様に表したものである．図中の HMM 共有 [tied] と HMM 独立 [untied] を比較すると，最短通過頻度の下限が大きい場合（300 以上）では，HMM 独立の方が WER の改善が大きく，その差は最短通過頻度の下限を大きくするほど広がっている．このことから，最短通過の観測頻度が上位の単語では発声変形の単語依存性が高くなっている可能性が考えられる．しかし，最短通過頻度の下限が小さい場合（300 以下）では，HMM 共有により，大きな WER の削減が得らた．これは，HMM 共有により，モデルの規模 (Δhmm) の増大を小さく，過学習を抑制できるため，多くの単語の発声変形 HMM を導入できるためであると考えられる．図 6 も単語を発声変形の単位とした場合の発音辞書の増分と誤り単語のカバー率を図 4 と同様に示したものである．この条件では，HMM 共有と HMM 独立で，発声変形の考慮対象となる単語は同一である．本実験条件では，誤り単語のカバー率は 3 割から 7 割であった．

さらに，図 3 と図 5 において最も大きな WER の改善が得られた条件（tri.[untied]@1000, wrd.[tied]@100）を比較すると，WER は同等 (5.7%) であるが，発声変形モデルの規模 (Δhmm) はトライフォンを単位として HMM を独立とした場合が，単語を単位として HMM を共有した場合の 13%と小さくてすんだが，発音辞書の追加エントリ数は 26 倍大きくなった．評価音声 All では，最大で 6.7%の誤り削減率が得られた．

3.2.2　対談部分 (Spon)

次に，評価音声の対談調発話のサブセット（Spon）の認識結

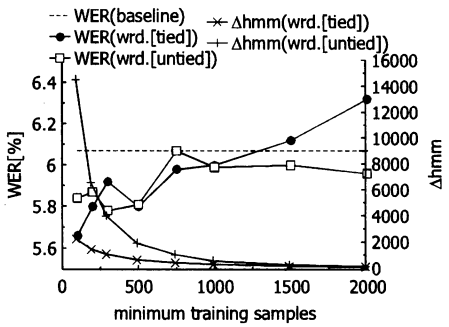

図 5　頻度下限による単語誤認識率と追加された発声変形用 HMM 数 (All) 単語:wrd. HMM 共有:[tied] HMM 独立:[untied]

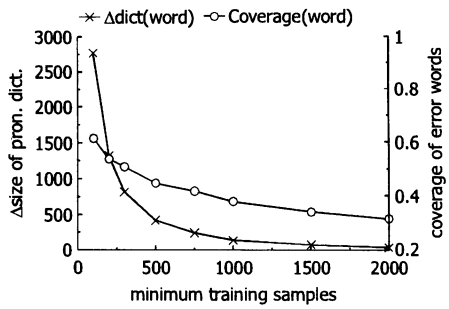

図 6　頻度下限による発音辞書エントリの増分と誤認識単語のカバー率 (All) 単語:word

果を見てみる．対談部分は，自由発話に近い音声が比較的多く含まれ，単調な話速ではない．また，最短通過が含まれる認識誤りの 66%が含まれていることからも，速い話速を考慮した発声変形モデルがより有効であると考えられる．

図 7 は図 3 に対応して，トライフォンを発声変形の単位とし，HMM 共有と HMM 独立の場合の各音響モデルの最短通過頻度の下限ごとの対談部分 (Spon) の WER である．対談部分では，評価音声全体 (All) での認識結果と同様に，HMM を共有するよりも HMM を独立とした方が WER の削減が大きく，両者の差はより顕著に見られた．図 8 は図 5 に対応して，単語を発声変形の単位とし，HMM 共有と HMM 独立の場合の各音

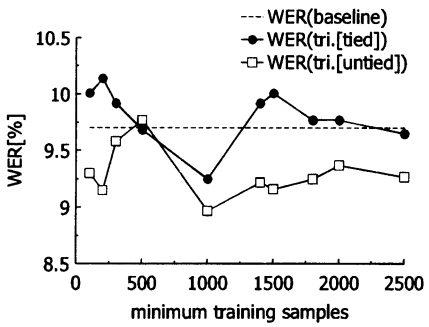

図 7 頻度下限による単語誤認識率（Spon）トライフォン:tri. HMM 共有:[tied] HMM 独立 [untied]

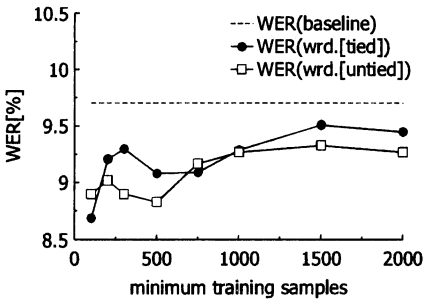

図 8 頻度下限による単語誤認識率（Spon）単語:wrd. HMM 共有:[tied] HMM 独立 [untied]

響モデルでの WER である．対談部分では図 5 に見られるような，頻度下限が大きい部分での発話変形の単語依存性は見られず，HMM 共有と HMM 独立で WER に同様の傾向が見られた．これは，学習された発声変形 HMM と，対談部分での発声変形単語に整合性が高かったためであると考えられる．実験を通して，HMM を独立とした場合の WER の改善が HMM を共有した場合より大きい傾向にあるが，最大の WER 削減率は HMM を共有した場合に得られ，10.4%であった．

3.3 考　察

この実験を通して導き出される結論は次の二点である．一点目は，トライフォンを発声変形の単位とした場合，HMM を独立にしたほうが WER の改善と探索空間の面から有利であったことである．二点目は，話速による発声変形には単語依存性が見られたことである．しかし，単語を発声変形単位として HMM を独立とした場合には，追加単語数の増加とともに，HMM を共有した場合に比べより多くの発声変形 HMM が必要となり，過学習が起こる．結果的に，HMM を共有しない場合に最大の WER 改善が見られた．

次に，トライフォン単位と単語単位の比較であるが，今回作成したトライフォン単位の発声変形 HMM の次の二点を検討する必要があるため，今後の課題としたい．

今回用いた HMM 独立のトライフォン単位の発声変形モデルには，2.2 節で述べたとおり，最短通過長以下の短い音素による周囲の音素モデルとの不整合がある．そこで，発声変形 HMM に，1 → 3，3 → 5 の状態スキップ遷移を許して頻度下限 1000 でのモデルを学習したが，WER の改善は見られなかった（All skip なし 5.7% → skip あり 6.1%）．今後 4 状態，または 3 状態の発声変形 HMM を検討する必要がある．

一方，本実験においてトライフォンを単位とした場合の発音辞書は，一単語内に複数の発声変形単位が含まれている場合でも，全ての発声変形 HMM が同時に用いられた一種類しか追加されていない．このような場合，発声変形単位のコンビネーションを考えたものが望ましいが，辞書への追加エントリ数はさらに多くなり，探索空間がさらに増大することが予想される．これらのコンビネーションによる WER の改善効果とマルチパス音響モデルを利用した探索の効率化は今後の課題としたい．

4. あとがき

放送音声の認識率低下を引き起こす発話スタイルによる発声変形を取り上げ，発声変形用音響モデルの導入法を検討した．本稿では，メタデータ制作を目的とした J リーグ中継実況音声とニュース字幕自動作成を目的としたニュース対談部分を対象として，絶叫発話用の発声変形モデルと話速の速い部分用の発声変形モデルを導入し，発声変形の導入単位による改善率を比較した．認識実験の結果，J リーグ中継実況音声においてはキーワード誤りの 15%が削減され，ニュース対談部分では単語誤りの 10%が削減された．

本稿で取り上げたような発声変形を扱った音響モデルの学習には，発声変形部分の特定が課題となる．今後話速以外の発声変形を扱うためにも，発声変形部分の特定方法の確立が今後の課題である．

文　献

[1] T. Imai, A. Kobayashi, S. Sato, S. Homma, K. Onoe and T. S. Kobayakawa, Speech Recognition for Subtitling Japanese Live Broadcasts, ICA-2004, pp.I165–I168, April, 2004, Kyoto, Japan.

[2] 佐藤庄衛，尾上和穂，小林彰夫，今井亨，スポーツ番組用メタデータ制作のための音声認識，音響学会秋季講演論文集，pp.127–128, 2004.

[3] 佐藤庄衛，尾上和穂，小林彰夫，本間真一，今井亨，発声変形を考慮したスポーツ実況音声の認識，音響学会秋季講演論文集，pp.101–102, 2005.

[4] M. Sano, H. Sumiyoshi, M. Shibata and N.Yagi, Generating Meta-data from Acoustic and Speech Data in Live Broadcasting, Proc. of ICASSP, pp.1145–1148, 2005.

[5] 佐古淳，有木康雄，状態推定音声認識を用いた野球中継の構造化およびイベント検出，音響学会秋季講演論文集，pp.129–130, 2004.

[6] 五十川賢造，西本卓也，篠田浩一，嵯峨山茂樹，品詞情報と単語内位置情報を用いた話し言葉音声認識のための状態クラスタリング，音響学会春季講演論文集，pp.7–8, 2003.

[7] 篠崎隆宏，古井貞熙，発話速度変動を考慮した隠れモード HMM による音声のモデル化，電子情報通信学会技術研究報告，SP2003-41, pp.37–42, 2003.

[8] 山田一郎，佐野雅規，住吉英樹，八木伸行，柴田正啓，アナウンスコメントを利用したサッカー番組メタデータ自動生成，電子情報通信学会技術研究報告，2004-112, pp.37–42, 2005.

[9] A. Kobayashi, K. Onoe, T. Imai and A. Ando, Time Dependent Language Model for Broadcast News Transcription and Its Post-Correction, Proc. of ICSLP, pp. 2435–2438, 1998.