自発モノローグにおける息継ぎ音自動アノテーションの試み

浅井拓也 (早稲田大学), 菊池英明 (早稲田大学), 前川喜久雄 (国立国語研究所)

Automatic Annotation of Breathing Noise in Spontaneous Monologue: A Preliminary Report

Asai Takuya(Waseda university), Hideaki Kikuchi(Waseda university), Kikuo Maekawa(NINJAL)

1. はじめに

本研究はイントネーションや談話研究への応用を念頭におき，『日本語話し言葉コーパス』(Corpus of Spontaneous Japanese，CSJ) 前川 (2004), の音声信号中に含まれる息継ぎ音（話者の息継ぎによって生じる雑音) ラベルを付与することを目標にしている．本稿では第一報として学会講演，模擬講演から合計８名分の音声を選択して分析し，息継ぎ音自動検出器を作成した結果を報告する．

従来，話し言葉を対象にした息継ぎ音の自動検出では，Price et al. (1989) や Wightman and Ostendorf (1994) などがある．Price et al. (1989) はプロのアナウンサーが発話した音声に対して，ケプストラムを特徴量としたガウス混合分布モデル (Gaussian mixture model，GMM) ベースの検出器を使い，93% の識別率を得た．Wightman and Ostendorf (1994) は同じくケプストラムを使用したベイズ検出器を使い，オープンテストで 73% の識別率を得た．また，歌唱における息継ぎ音検出として，中野ほか (2008) は，息継ぎの自動検出に有効な音響的特徴量を解析し，検出器を作成している．この研究では，MFCC(Mel-frequency cepstral coefficients)，ΔMFCC，ΔΔMFCC，Δ パワーを特徴量とし，隠れマルコフモデル (Hidden Markov Model，HMM) を使用して，再現率 97.5%，精度 77.7% を得た．

しかし，これらの結果はいずれも，教示によって統制された音声を対象としており，また，自発音声における発話者の発話特性の変化や個人差の影響を考慮していない．そこで本研究では，自発音声の息継ぎ音に関しても上記のような音響的特徴量が息継ぎ音の検出に有効であるのかを検討し，発話者の個人性に合わせた検出器を作成することで，発話者の性質が息継ぎの自動検出の精度に与える影響を観察する．

2. 自動検出の目標

まずは，本解析における息継ぎ音自動検出器の最終的な出力について定義する．検出器の出力結果のイメージを図 1 で説明する．図 1 は今回解析の対象とした音声ファイルの一部を praat を使用して可視化したものである．最上部から音声波形，スペクトログラム，CSJ にもともと付与されている情報の一つである word 層，そして今回の自動検出結果のイメージを表示している．

息継ぎ音の検出において問題になるのは，例えば母音や摩擦音などの音響的特徴が似ている音素との分離であるが，本解析では，これを CSJ にもともと付与されている情報を利用し，分離を行った．具体的には，非発話区間 (図 1 の word 層中の#で示されている部分) のみを解析の対象としている．そのため，今回の解析では，息継ぎ音の継続時間長や，発話区間における息継ぎ音の検出は行わない．

この非発話区間に，息継ぎ音が含まれている場合には 1 を，含まれていない場合には 0 を出力すること (図 1 の最下層) が今回の検出器の目標である．

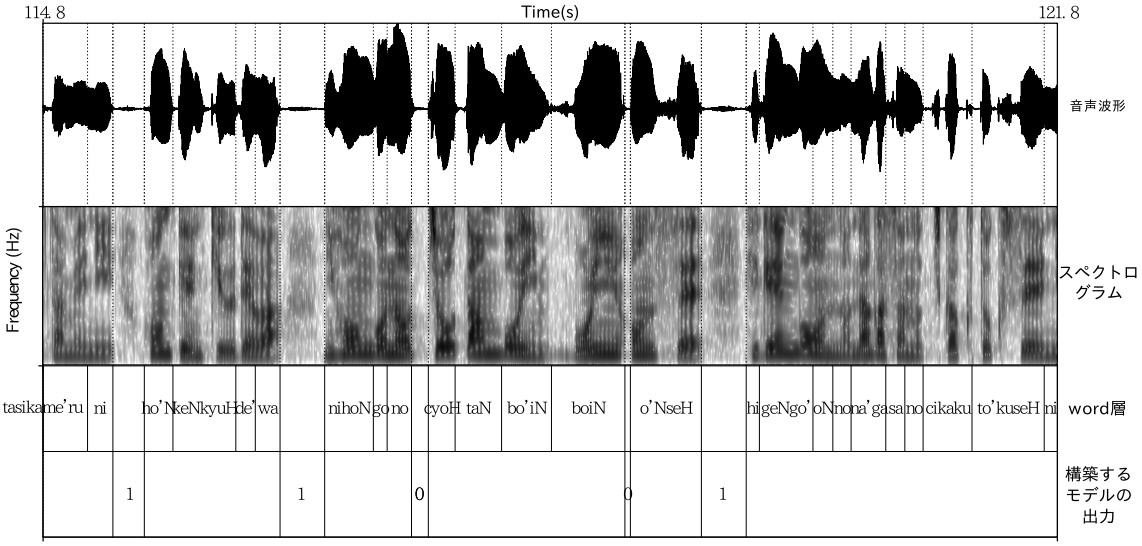

図 1　検出器の出力結果イメージ

3. データ

本章では，今回の解析で使用したデータについて説明する．今回の解析では，CSJ コーパスのコアデータから男女 4 名ずつ，合計 8 データを使用した．

息継ぎ音の特性に話速の影響がある可能性を考慮して，CSJ コーパスのコアデータの中から，特に話速の速いものと遅いものを男女それぞれ 2 名分ずつ用意した．なお，話速に関しては平均発話速度 (Mora/秒) により決定した．

これらの解析ファイルに関して，まず試験的に解析者 2 名 (エキスパート 1 名，学生 1 名) による息継ぎ音の手動アノテーションを行った．これは，そもそも自発音声における息継ぎ音は録音された音声から一貫性を持ってアノテーションを行うことが可能なのかを調査することが目的である．アノテーション範囲は上記 8 ファイルの冒頭 70sec と終端 60sec であり，非発話区間単位に息継ぎ音が含まれているか否かを評価した．

解析者の回答の一致率は κ 係数で算出した，その結果，解析者間の κ 係数は 0.92 となり，人間は息継ぎ音の検出を一貫性を持って行うことが可能であることが示された．

そのため，解析対象のデータはファイル全区間に対し学生 1 名によって手動によるアノテーションが行われた．ここでのアノテーションは，非発話区間に含まれる息継ぎ音の開始，終了時間を切り出している．この際切りだされた音声は息継ぎ音の音響的特徴量の観察と，自動検出器の教師情報及び，評価用の正解ラベルとして使用する．

本稿では中野ほか (2008) の解析を参考に，MFCC12 次元，ΔMFCC12 次元，ΔΔMFCC12 次元とパワー，Δ パワー，ΔΔ パワーを算出した．これら音響的特徴量の算出には HTK に付属している Hcopy を使用した．この際，フレームシフト幅は 10msec，フレーム幅は 25msec に設定した．対象の非発話区間の解析フレーム数は総数 171075 個であり，そのうち息継ぎ音と認定された解析フレームは 56219 個であった．

表 1 に今回の解析で使用したデータの基本情報 (発話者 ID，平均発話速度，講演時間，性別，息継ぎ音フレーム数，非息継ぎ音フレーム数) をまとめる．

表 1　解析対象データの基本情報

ID	平均発話速度	講演時間	性別	息継ぎ音フレーム数	非息継ぎ音フレーム数
A01M0015	11.57	716.24	男性	5855	10537
A03M0059	11.48	890.57	男性	6904	11746
A02M0076	8.27	1474.08	男性	7227	8494
A02M0098	7.10	1551.37	男性	10330	32441
A01F0067	10.43	942.75	女性	5996	18860
A03F0153	10.00	991.77	女性	7319	8488
A01F0122	8.17	828.17	女性	5548	15473
A06F0028	7.85	959.69	女性	7040	8817

3.1　音響的特徴量の観察

上記データに関して，音響的特徴量の観察を行った．この観察では，CSJ コアデータの音響的特徴量を，静的な特徴量 (MFCC)，動的な特徴量 (ΔMFCC，ΔΔMFCC)，パワー (パワー，Δ パワー，ΔΔ パワー) の三軸で観察する．ここでの目的は，これらの音響的特徴量のうち，息継ぎ音の自動検出器に効果的な特徴量を選択することである．

MFCC12 次元のペアプロットを図 2 に示す．この図では，それぞれ 1 次元目から 12 次元目までを上下左右に配置している．ペアプロットの下半分には，それぞれの特徴量の組み合わせの散布図を，上半分にはそれぞれの特徴量の確率密度を示す．なお，確率密度の推定にはカーネル密度推定を行い，それぞれの等高線はこの値を 10 分割して示している．最下にはそれぞれの特徴量のヒストグラムを，最右には，それぞれの次元の箱ひげ図を示した．図中，濃く示しているのが息継ぎ音，薄く示しているのが非息継ぎ音の解析フレームである．散布図や確率密度を確認すると，特に低次成分において，息継ぎ音，非息継ぎ音の分離が行えることが確認できる．これは，非発話区間における息継ぎ音には，ある種の母音的な特徴があることが原因であると思われる (図 1 を参照，息継ぎ音と認定したフレームにおいてはフォルマント成分が確認できる).

続いて ΔMFCC の観察を行う．ΔMFCC12 次元のペアプロットを図 3 に示す．分布を確認すると，ΔMFCC12 次元においては，息継ぎ音，非息継ぎ音の分布の中心点がほぼ同じ位置にあり，分離が行える次元が存在しないことが分かる．

図 4 に ΔΔMFCC12 次元のペアプロットを示す．ΔΔMFCC の分布も ΔMFCC と同様，息継ぎ音，非息継ぎ音の弁別に有効ではない可能性が強い分布となった．

これら 2 つの図の結果を考慮すると，CSJ の非発話区間から，息継ぎ音の自動検出を行う場合は，動的な特徴量よりも静的な特徴量を使用したほうが有効であることが判明した．ただし，この結果に関しては，解析フレームの幅の問題が存在する．特に動的な特徴量に関しては，解析幅の違いにより結果が異なる可能性がある．そのため，息継ぎ音の分離に適切な解析フレーム幅の検討は今後の課題である．

最後にパワーに関するペアプロットを図 5 に示す．パワーに関する分布を観察すると，特にパワーと ΔΔ パワーの組み合わせにおいて，非息継ぎ音の分離がある程度行えることが判明した．だたし，この結果に関しても，特に Δ パワー，ΔΔ パワーに関しては動的な特徴量の解析と同様，フレーム幅の問題が存在する．

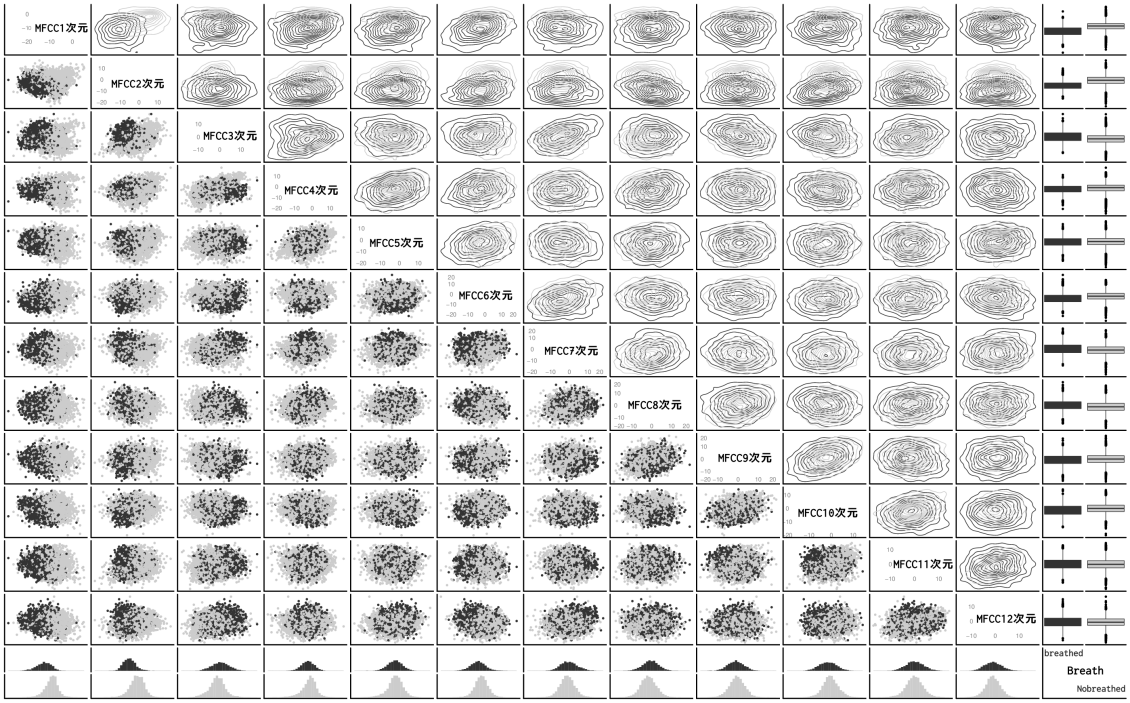

図 2　MFCC12 次元のペアプロット

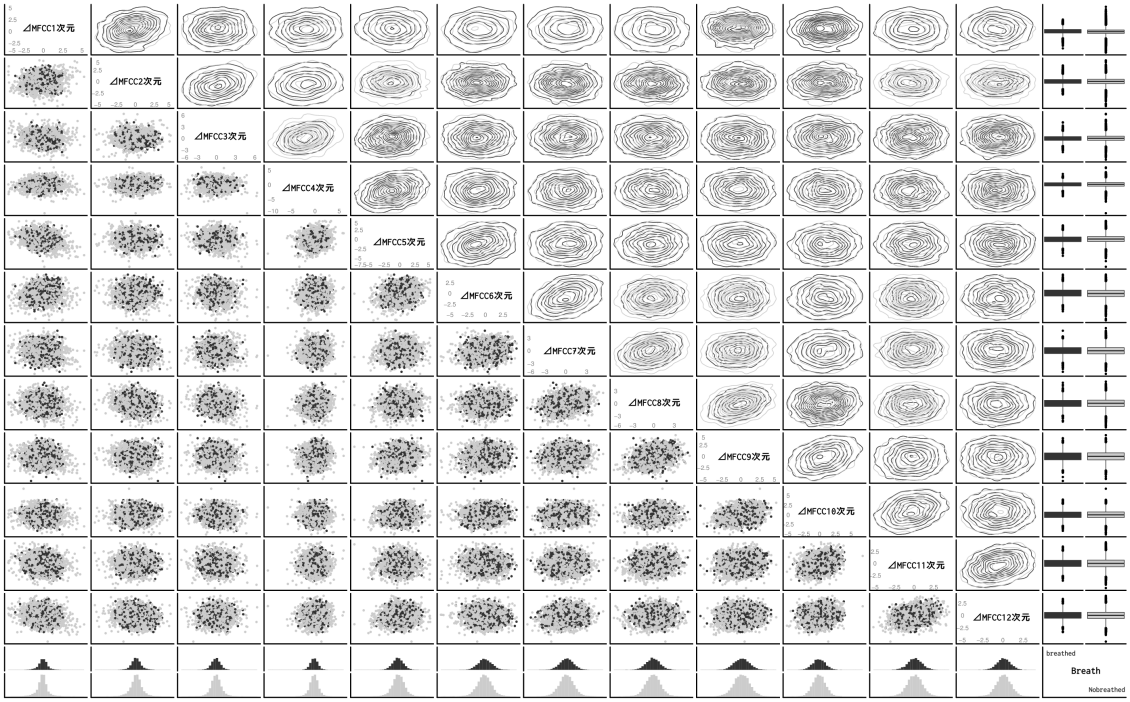

図 3　ΔMFCC12 次元のペアプロット

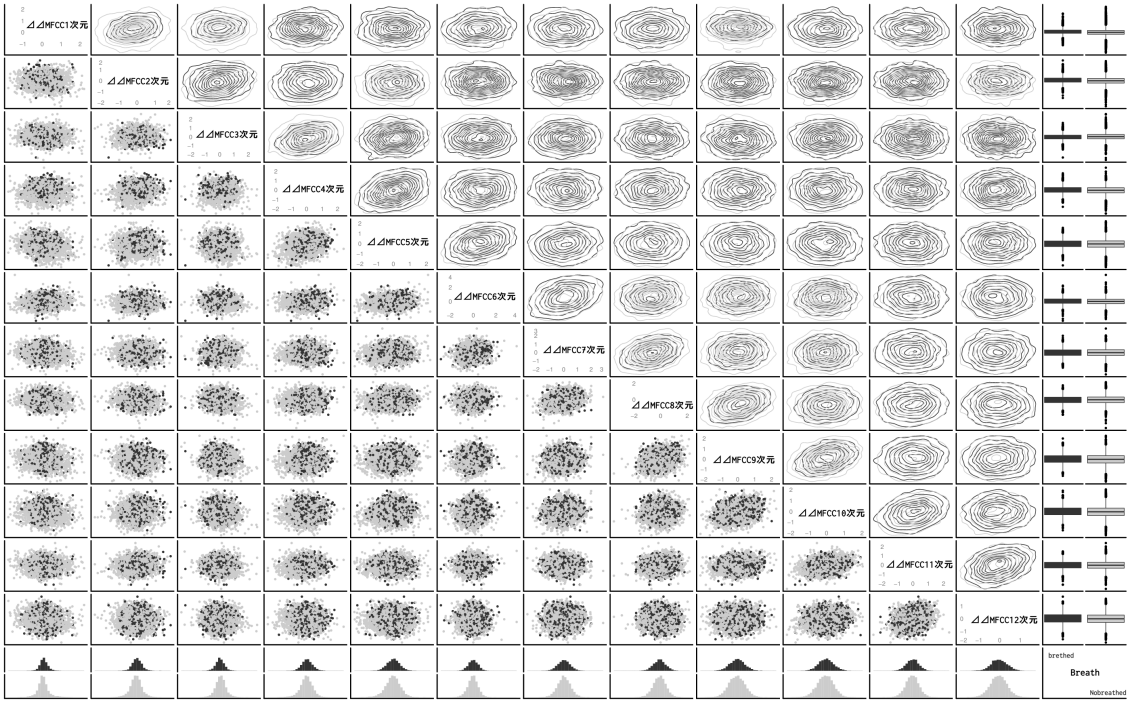

図 4　ΔΔMFCC12 次元のペアプロット

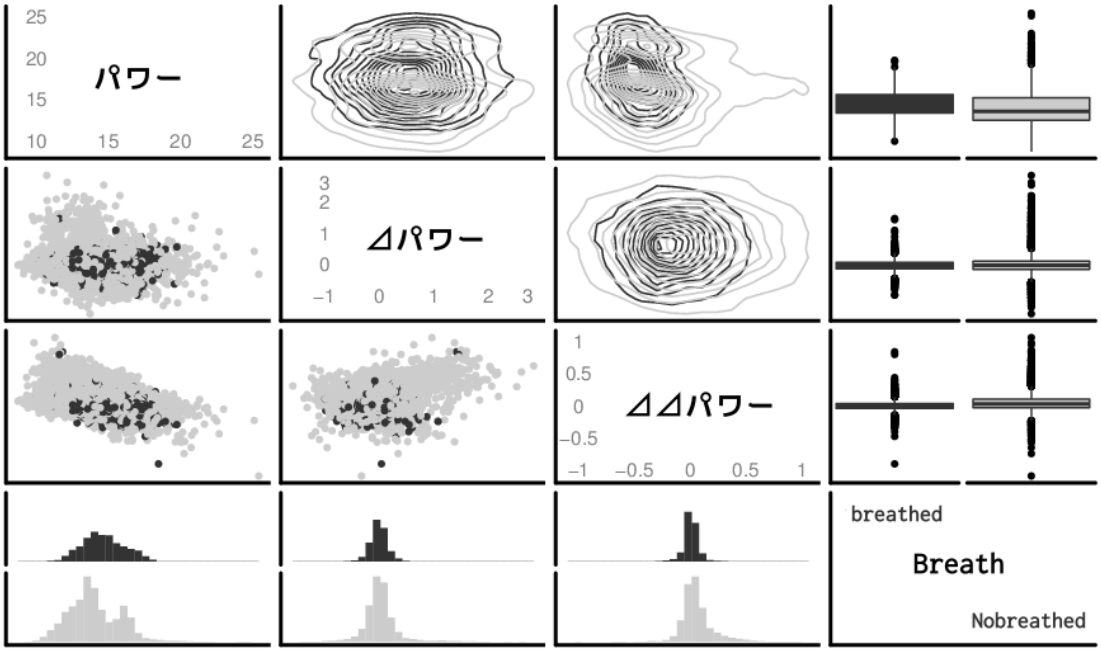

図 5　パワーのペアプロット

4. モデルの作成と評価

ここでは，上記の特徴量を利用して息継ぎ音の自動検出器の作成を行う．それぞれの特徴量を観察すると，どこか特定の次元のみで，息継ぎ音の検出を行えるような特徴量は存在せず，多次元の特徴量を使用する必要がある．また，今回の問題は，ある非発話区間に息継ぎが存在するか否かの二値の値を出力することである．このような場合において，最も効果的であるとされている分類器としてサポートベクターマシン (Support vector machine，SVM) がある．

今回の解析では，上記特徴量の観察の際に使用した解析フレームを 1 つの単位とし，それぞれのフレームが息継ぎ音であるか否かを SVM に分類させている．なお，モデルの学習時には計算コストを削減するために，すべてのデータの中から 6000 フレーム分のデータをランダムサンプリングして使用している．また，SVM のアルゴリズムは統計解析用プログラム言語 R の e1071 library を利用し，ラジアル基底関数 (RBF, radial basis function) カーネルを使用し，パラメタ C と γ はグリッド検索で最も精度の高くなる値に決定した．

以下の解析では，それぞれの条件別に，MFCC，ΔMFCC，ΔΔMFCC，それぞれのパワーを特徴量に，SVM による教師あり学習を行い，その精度の比較を行う．モデルの評価には recall，precision，F 値，accuracy，κ 係数を使用した．モデル評価時には，解析に使用したデータのうちの $\frac{2}{3}$ のデータを学習用に，$\frac{1}{3}$ をテスト用にし，3-fold cross validation 法を行った．

4.1 フルモデル vs MFCC モデル

CSJ 非発話区間における息継ぎ音の音響的特徴量を観察すると，MFCC12 次元が最も自動検出の結果に影響を与えやすいデータである．そこで，まずは，この値のみを使用しモデルの学習を行った (MFCC モデル)．モデルの学習結果を表 2 左に示す．この学習の κ 係数は 3 回の試行平均 0.63 となった．

この検討と同様に，ΔMFCC，ΔΔMFCC に関しても，それ単独に学習を行ったが，モデルの学習は収束しなかった．そのため，MFCC モデルにその他の特徴量を全て加えたモデル (フルモデル) を作成し，MFCC モデルと比較した．フルモデルの評価値を表 2 右に示す．フルモデルの 3 回試行の平均 κ 係数は 0.67 となった．わずかではあるが，MFCC モデルより分離精度の向上が確認された．本研究の目的は出来る限り精度のよい自動検出器の作成であるため，以降の検討においてはフルモデルを採用し，発話者の特性によるモデルの精度の変化を観察していく．

表 2　MFCC モデルとフルモデルの評価値比較

	MFCC モデル					フルモデル				
	k=1	k=2	k=3	mean	sd	k=1	k=2	k=3	mean	sd
precision	0.67	0.70	0.70	0.69	0.02	0.74	0.73	0.77	0.75	0.02
recall	0.81	0.78	0.79	0.79	0.02	0.82	0.77	0.81	0.80	0.03
F	0.74	0.74	0.74	0.74	0.00	0.78	0.75	0.79	0.77	0.02
accuracy	0.83	0.84	0.85	0.84	0.01	0.85	0.85	0.87	0.86	0.01
kappa	0.61	0.63	0.64	0.63	0.01	0.67	0.64	0.70	0.67	0.03

4.2 男女差の比較

呼吸法に関しては，一般に男女で異なると言われている．そのため，息継ぎ音の自動検出においても，男性モデルと女性モデルを作成したほうが，検出力が高まる可能性がある．そこで，上記のフルモデルを利用し，学習データを男性のみにした場合と，女性のみにした場合でのモデルの精度を比較した．

男性モデルの評価結果を表 3 左に，女性モデルの学習結果を表 3 右に示す．男性モデルの 3 回の試行平均 κ 係数は 0.60，女性モデルの平均 κ 係数は 0.76 となった．

この試行の結果，男性モデルに関してはフルモデルよりモデルの精度が低くなり，女性モデルはフルモデルよりもモデルの精度が向上することが分かった．この結果は，女性の息継ぎ音は比較的安定した音響的特徴量が存在するのに対し，男性の息継ぎ音は音響的特徴量のばらつきが大きいことを示唆する結果である．

表 3　男女別モデルの評価値比較

	男性モデル					女性モデル				
	k=1	k=2	k=3	mean	sd	k=1	k=2	k=3	mean	sd
precision	0.68	0.65	0.65	0.66	0.02	0.80	0.80	0.85	0.82	0.03
recall	0.81	0.76	0.80	0.79	0.03	0.90	0.84	0.84	0.86	0.04
F	0.74	0.70	0.72	0.72	0.02	0.85	0.82	0.84	0.84	0.02
accuracy	0.84	0.82	0.84	0.83	0.01	0.89	0.89	0.90	0.89	0.00
kappa	0.62	0.57	0.61	0.60	0.03	0.77	0.74	0.77	0.76	0.01

4.3　個人差の比較

男女差の検討に関しては，これが本当に男女差によるものなのか，それとも個人差によるものなのかを検討する必要がある．そのため，解析対象のデータに関して，各発話者一人ひとりを評価データとし，学習データはその発話者以外のデータを使用した検討を行った．この検討で，ある発話者をテストデータとした場合の評価値が低い例が見つかれば，その発話者の特殊性が判断できる．

この検討の結果を表 4 に示す．この表においては各条件の κ 係数のみを示している．Unweighted は通常の κ 係数，Weighted は重み付け κ の値である．

表 4 を確認すると男性ファイル (ID 名に M が含まれるもの) の値は女性ファイル (F が含まれるもの) に比べ評価値の値が低いことが分かる．そのため，男女差の検討での結果は個人差というよりも性別の影響を受けた結果であることが判明した．ただし，男性ファイルでは個人差間のモデルの評価値のばらつきも大きい．特に，最も精度の低かった M0098 は，発話者が高齢であり，息継ぎ音の音響的特徴量が他の発話者と性質が異なるものである可能性が強い．

表 4　各発話者別 κ 係数

	F0067	F0122	M0015	M0076	M0098	F0153	M0059	F0028
Unweighted	0.67	0.75	0.51	0.46	0.36	0.80	0.63	0.53
Weighted	0.67	0.75	0.51	0.46	0.36	0.80	0.63	0.53

4.4　話速の差の比較

最後に，話速の差に関する検討を行う．本稿では解析データの選別に関して，ファイル単位での話速の速いものと遅いものを選択的に選んでいる (表 1 参照)．そこで得られた話速別に，話速の速い発話者，遅い発話者別のモデルを作成した．話速の速い発話者モデルを表 5 左に，遅い発話者モデルを表 5 右に示す．これらのモデルにおいての κ 係数はそれぞれ，0.75，0.59 となった．そのため，この結果のみを比較すると話速の影響によりモデルの検出結果が変化するように見える．

しかし，4.2 節での検討の結果もあり，この結果が話速によるものなのか，男女差によるものなのかの検討が必要である．そこで，話速の遅い群に関しては更に，男女別にモデルを作成し，評価を行った．

表 5　話速別モデルの評価値

	話速の速い発話者モデル					話速の遅い発話者モデル				
	k=1	k=2	k=3	mean	sd	k=1	k=2	k=3	mean	sd
precision	0.79	0.81	0.85	0.82	0.03	0.72	0.67	0.68	0.69	0.03
recall	0.89	0.82	0.84	0.85	0.03	0.77	0.72	0.73	0.74	0.03
F	0.84	0.82	0.85	0.83	0.02	0.74	0.70	0.70	0.71	0.03
accuracy	0.88	0.88	0.90	0.89	0.01	0.84	0.83	0.82	0.83	0.01
kappa	0.75	0.73	0.78	0.75	0.03	0.62	0.57	0.58	0.59	0.03

低話速度男性モデルの評価値を表 6 左に，低話速度女性モデルの評価値を表 6 右に示す．それぞれの三回の試行平均 κ 係数は 0.53，0.74 となった．つまり，同じ低話速度条件においても女性モデルの評価値は高いままである．その結果，ここでの検討結果は話速による影響ではなく，男女差の影響を反映したものであることが判明した．

表 6　低話速度男女別モデルの評価値比較

	男性モデル					女性モデル				
	k=1	k=2	k=3	mean	sd	k=1	k=2	k=3	mean	sd
precision	0.68	0.61	0.59	0.63	0.05	0.79	0.83	0.84	0.82	0.03
recall	0.72	0.69	0.69	0.70	0.02	0.92	0.81	0.79	0.84	0.07
F	0.70	0.65	0.64	0.66	0.03	0.85	0.82	0.82	0.83	0.02
accuracy	0.83	0.80	0.80	0.81	0.02	0.89	0.89	0.87	0.88	0.01
kappa	0.58	0.51	0.50	0.53	0.05	0.76	0.74	0.72	0.74	0.02

5. まとめ

本稿では CSJ 講演音声の息継ぎ音自動検出器作成に向けて，息継ぎ音の音響的特徴量の観察を行った．その結果，CSJ コアデータの息継ぎ音自動検出には静的な特徴量である MFCC を使用するのか最も効率がよいことが分かった．続いて，SVM を使用した自動検出器を作成し，3-fold cross validation 法で評価を行った．その結果，今回提案する検出器の精度は κ 係数で 0．6 以上という大変よい精度を得た．また，発話者の性質別にモデルの学習を行い，発話者の性質とモデルの精度の関係を観察した．その結果，作成した息継ぎ音の自動検出器の精度には男女差が存在することが分かった．話速に注目した解析も行ったが，女性の発話においては検出器の精度に大きな差はなく，男性の発話においてのみ話速の差が確認された．

これらの解析は息継ぎの音響的特徴量による自動検出という課題においては，男女差を考慮する必要があること，特に男性の発話においては，話速の影響を考慮する必要のあることを示す結果である．

6. 課題

最後に今後の課題について述べる．

まず，本稿での報告の段階では冒頭で上げたモデルの出力 (各非発話区間毎の二値データ) の形式ではなく，音響解析フレーム毎に息継ぎ音か非息継ぎ音かを検出している．これを当初の目的に合わせ，非発話区間毎に変換をする必要がある．これに関しては今回の提案モデルは音響フレーム毎の評価では非

常に良い成績を出しているため，例えば，ある非発話区間に含まれる音響解析フレームの値の多い方をその区間の正解とすることで, より高い精度での検出が可能になると考えている．

また，今回の解析の結果，特徴量に MFCC12 次元を使用し，SVM による検出器を作成することで，マイクに収録されている息継ぎ音の自動検出はある程度よい精度を得られることが判明したが，最終的にこの結果をイントネーションや談話分析に応用していくことを考えると，マイクに収録されることのない，小さな息継ぎや発話区間中に発生する息継ぎ音に関しても抽出しきれることが望ましい．

実際にマイクに収録されることのなかった息継ぎ音の存在を示唆するために図 6 を示す．この図は今回解析を行った各発話者ごとの，ある息継ぎ音の発生から，次の息継ぎ音の発生までの時間をヒストグラムにしたものである．図中横軸はある息継ぎ音から次の息継ぎ音までの持続時間，縦軸はその出現回数である．

これらの分布を確認すると特に A02M0098 と A06F0028 において分布の二峰性が確認される．その他の分布に関しても分布の混合が確認される．

この図で示しているものは，それぞれの発話者の息を吐き続けている持続時間である．そして，直感に従うのなら，人間が息を吐き続けることができる時間には限界があり，おそらくは正規分布を仮定してもよいはずである．しかし，図 6 はひとつの正規分布ではなく，時間的に数カテゴリある混合分布である．つまり，この混合数分だけ，マイクに収録されることのない息継ぎが存在している可能性がある．

今後の試行として，実際に行われた息継ぎと，マイクに録音された息継ぎ音の差がどの程度のものであるのかについて厳密な実験を行う予定である．

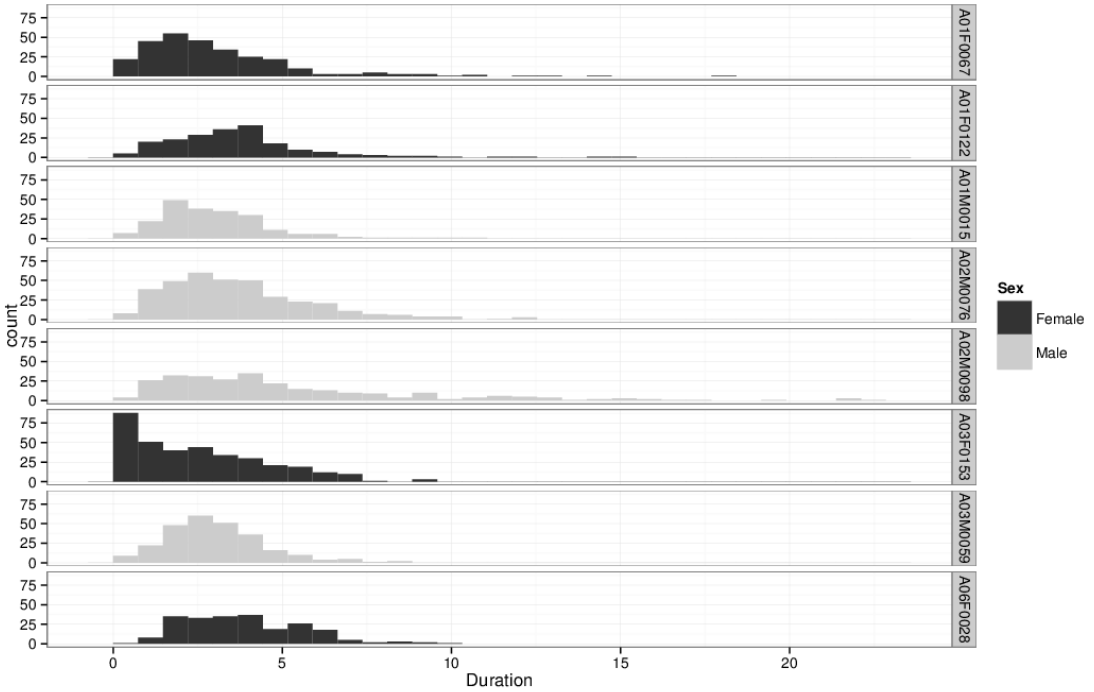

図 6　各発話者の息継ぎ音までの持続時間

参考文献

Price, P. J., M. Ostendorf, and C. W. Wightman (1989). “Prosody and parsing.” *Proc. Work. Speech Nat. Lang. - HLT ’89*, p. 5 Morristown, NJ, USA: Association for Computational Linguistics.

Wightman, C.W., and M. Ostendorf (1994). “Automatic labeling of prosodic patterns.” *IEEE Trans. Speech Audio Process.*, 2:4, pp. 469–481.

中野倫靖・後藤 真孝・緒方 淳・平賀 譲 (2008).「無伴奏歌唱におけるブレスの音響特性とそれに基づく自動ブレス検出 (音響分析一般 (1))」　情報処理学会研究報告. [音楽情報科学], 2008:78, pp. 83–88.

前川喜久雄 (2004).「『日本語話し言葉コーパス』の概要」　日本語科学, 15, pp. 111–133.