日本語話し言葉コーパスにおける発声様式の自動分類

著者	森 大毅，藤本 雅子，浅井 拓也，前川 喜久雄
雑誌名	言語資源活用ワークショップ発表論文集
巻	1
ページ	347-354
発行年	2017
URL	http://doi.org/10.15084/00001490

日本語話し言葉コーパスにおける発声様式の自動分類

森 大毅 (宇都宮大学大学院工学研究科) *
藤本 雅子 (国立国語研究所)
浅井 拓也 (北陸先端科学技術大学院大学)
前川 喜久雄 (国立国語研究所)

Automatic classification of phonation type in the Corpus of Spontaneous Japanese

Hiroki Mori (Utsunomiya University)
Masako Fujimoto (National Institute for Japanese Language and Linguistics)
Takuya Asai (JAIST)
Kikuo Maekawa (National Institute for Japanese Language and Linguistics)

要旨

喉頭音源由来の声質の違いは，話者のパラ言語メッセージならびに心的・認知的状態を伝えるシグナルであり，自発音声コーパスに求められる重要な情報であるが，そのアノテーションは音声学の専門家でなければ難しくコストが大きい。本研究は，機械学習による声質の自動アノテーションの可能性を探ることを目的とする。本研究では，非流暢性にも関連する従来よく用いられてきた発見的な音響特徴量に加え，近年音声からの感情認識で広く用いられるようになった大規模な特徴量セットの効果を検証した結果を報告する。

1. はじめに

書き言葉と異なり，話し言葉はさまざまな音響的手がかりによって話者の意図や態度の違い、すなわちパラ言語情報を伝達している。また同時に、心的状態のように、話者の伝達の意志とは関係なく伝達される情報もある (森ほか 2014)。

話者のパラ言語メッセージや心的・認知的状態を反映する音響的手がかりには、ピッチや大きさの抑揚、テンポ、リズムなどの韻律や、母音の質に代表される声道の特徴が含まれる。中でも、喉頭における発声様式の違いに起因する声質 (Laver 1980, Ni Chasaide and Gobl 1997) は、生成面に着目した声質の基本的記述として、音声学的にも工学的にも重要である。また、自発音声に現れる声質の違いは、心的・認知的状態を無意識に伝える社会的シグナルであり、実時間コミュニケーションの動的な構成に関わるターンテイキングや非流暢性などの現象とも密接な関係がある。このため、これらの研究の基礎資料となる自発音声コーパスには、声質のアノテーションを有することが望まれる。

しかしながら、自発音声の場合には発声中の喉頭の観察は難しい。また、『日本語話し言葉コーパス』(CSJ) のように、音声以外に測定された信号を有しないコーパスも多い。この場合、

* hiroki-public@speech-lab.org

声質の記述を声帯振動の観察により行うことは不可能であり、そのかわり、専門家が聴覚的に判断する必要がある。コーパス構築において、このような声質のアノテーションに要するコストは膨大であり、このことが大規模なデータに基づいた研究を難しくしている。

本研究は，機械学習による声質の自動アノテーションの可能性を探ることを目的とする。著者らはこれまで、CSJ に含まれる長母音/aH/ /eH/を対象に、機械学習アルゴリズムによりフィラーの判別を試みた (Maekawa and Mori 2016)。音響的特徴として継続時間、強度、F0、フォルマント周波数、ジッタ、シマ、調波対雑音比、スペクトル傾斜を用い、フィラーと語彙項目のサンプルが同数の条件で交差検証を行った結果、同定精度は $F = 0.89$ であった。フィラーの多くは通常の語彙項目とは異なる声質で発声されていると考えられるため、提案した手法は声質の自動分類においても有効である可能性がある。

近年は、機械学習技術の発達に伴い、音声から話者の感情や年齢・性別などの情報を推定する問題において、これらと強く関連すると予想される少数の音響パラメータを使うのではなく、網羅的に抽出された音響パラメータ列 (LLD; 4.2 参照) から組織的に生成された非常に多数の要約統計量をそのまま使って機械学習を行うことが一般的になってきた。この種の方法は一見たいへん非効率であるが、少数精鋭のパラメータセットを用いる場合に比べ、性能が大きく改善する場合が多い。本研究では、過去の研究で用いられてきた発見的な特徴量に加え、これらの大規模な特徴量セットを用いることの有効性もあわせて検証する。

2. 対象とする声質

2.1 Creaky voice

Creaky voice(きしみ声) とは、creak または vocal fry と呼ばれる、極端に低いピッチやパルス的な音で特徴づけられる発声様式の特徴をある程度有する声であることを意味する。自発音声においては、ピッチの低下と同じように、次のような場所・場面でよく観察される。

- フィラー
- 発話末や句末
- 自信のない心理状態

Ishi et al. (2008) は、周期性と声帯パルスの類似性を利用した vocal fry の検出法を提案し、自然発話データに対して 73% の検出率と 13% の挿入誤り率を達成したと報告している。ただし、対象としたデータは、Sadanobu (2004) が「りきみ」と呼ぶ pressed かつ creaky な発声に限られており、一般の creaky voice に対する提案手法の有効性は明確ではない。

2.2 Breathy voice

Breathy (気息性) な声は、声門閉鎖の不完全により、声帯振動による周期音と同時に生じる乱流雑音によって生成される。乱流雑音の程度および様態により、whisper → whispery → breathy → modal と様々な語によって形容される。すなわち、気息性は程度の問題であり、非気息性の声との間に明確な境界は存在しない。Breathy な声は、ある種の個人性を特徴づけるほか、落胆などの心的状態を反映した低緊張の状態に関連する。音響的には、大きなスペクトル傾斜と低い調波対雑音比により特徴づけられる (Klatt and Klatt 1990)。

表 1　サンプル数

(a) フィラー

母音	全データ数	creaky	breathy
/a/	67	20	11
/e/	110	42	34
/i/	0	0	0
/o/	54	19	9
/u/	0	0	0

(b) 通常語彙項目

母音	全データ数	creaky	breathy
/a/	222	19	31
/e/	126	21	24
/i/	84	8	14
/o/	182	15	22
/u/	53	1	6

3. データ

データとして、CSJ のコア部分中の独話 (学会講演と模擬講演) を使用する。母音のみから構成されるフィラー 231 個、および通常の語彙項目中の母音 667 個をランダムにサンプリングした。

声質のラベリングは、発声様式に関する豊富な研究経験を持つ第 2 著者および第 4 著者が行った。2 名のラベラーは、全てのサンプルを 2 回ずつ聴取し、“creaky”, “breathy”, “modal” のいずれであるかを判定した。1 発声の中で、creaky から modal、または modal から creaky へのように変容が生じていると判断される場合には、その旨を記述した。

表 1 に、ラベリング結果から得た声質の分布を示す。creaky および breathy に分類されているサンプルは、ラベラー 2 名による全 4 回の判定のうち、1 回でも creaky または breathy と判定されたものである。

4. 音響特徴量

4.1 基本 15 次元

今回検討する音響特徴量のうち、この節で説明するものは、著者らが過去にフィラーの分析に使用したものである (Maekawa and Mori 2016)。

■継続時間　継続時間 (duration) は対象母音の始点から終点までの時間である。

■強度　強度 (intensity) は、単位をデシベルとして求めたものを対象母音区間において平均した。

■基本周波数　F0 は、対象母音区間から Praat の Pitch(ac) コマンドにより求め、対数を取った後に、話者ごとに平均と標準偏差を正規化した。また、ピッチの変動の指標として標準偏差 (sdPitch) を求めた。

■フォルマント周波数　フォルマント周波数 (F1, F2, F3) は、線形予測分析によって得られたものを対象母音区間において平均し、対数を取った後に、話者ごとに平均と標準偏差を正規

化した。

■ジッタ、シマ　基本周波数ゆらぎであるジッタ (jitter) としては PPQ5 (Gelzinis et al. 2008) を、振幅ゆらぎであるシマ (shimmer) としては APQ5 を用いた。

■スペクトル傾斜　スペクトル傾斜として、4 種類の特徴量を用いた。C1TL は、対象母音区間の線形予測分析から得られるスペクトル包絡の情報である LPC ケプストラムの第 1 係数から求めたスペクトル傾斜である (前川・森 2005)。H1−H2, H1−A1, H1−A2, H1−A3 は、それぞれ第 2 高調波成分、第 1 フォルマント、第 2 フォルマント、第 3 フォルマントの基本波成分に対する振幅の比である (Gordon and Ladefoged 2001)。

■調波対雑音比　調波対雑音比 (HNR) は Praat により求めた。

4.2　openSMILE

近年の音声からの感情認識研究では、数種類のパラメータを厳選するのではなく，非常に多くのパラメータ，およびそれらから二次的に導出される発話単位の最大・最小・レンジ・平均・四分位数・百分位数・標準偏差などの要約統計量を無制限に利用する方法が主流になっている。音声に関する重要な国際会議の 1 つである Interspeech では 2009 年より感情をはじめとしたパラ言語情報の認識精度を競うコンテストが開催されており， 2013 年に開催されたコンテスト Computational Paralinguistics Challenge における標準特徴は，エネルギー関連の 4 種類，スペクトル関連の 54 種類，有声音源特性に関する 6 種類からなる計 64 種類の低次特徴量 (Low Level Descriptor: LLD) を基に導出された，計 6373 の特徴から構成されている。

このような目的に特化した音響特徴量抽出ソフトウェアに openSMILE (Eyben et al. 2010) がある。今回は、openSMILE を用い、対象母音区間から Interspeech 2009 Emotion Challenge (Schuller et al. 2009) におけるベースライン特徴量を抽出した。これらは、対象母音区間の各分析フレームにおけるゼロ交差率、実効値、F0、HNR、MFCC(メル周波数ケプストラム係数) 12 次元、およびこれらの 1 次回帰係数からなる 32 種類の低次特徴量の系列に対し、区間単位での平均、標準偏差、歪度、尖度、最大値／最小値およびその位置、レンジ、線形回帰係数およびその平均 2 乗誤差からなる 12 種類の汎関数を適用して得られる 384 次元のベクトルである。

5.　声質の自動分類実験

機械学習アルゴリズムとしてサポートベクターマシン (SVM) およびランダムフォレスト (RF) を使い、声質の自動分類を行った。実験は、creaky voice を発見するタスク、および breathy voice を発見するタスクそれぞれについて行った。表 1 に示したように、声質ごとのデータ数は大きく異なる。このため、学習時には、データ数に反比例したコストを定義し、非 creaky または非 breathy に偏って学習されるのを防いだ。機械学習アルゴリズムのパラメータは F 値を基準に調整し、評価は leave-one-out 交差検証法により行った。評価尺度は F 値および ROC 曲線下面積であり、ともに 1 に近いほど検出性能が高い。

667 個の母音全てを対象にした実験の結果を表 2 に示す。表 2 から、SVM に比べランダム

表 2　自動分類結果

(a) F 値

	creaky	breathy
基本 15 次元 (SVM)	0.552	0.581
基本 15 次元 (RF)	0.562	0.624
15 + 384 次元 (SVM)	0.523	0.619
15 + 384 次元 (RF)	0.590	0.619

(b) ROC 曲線下面積

	creaky	breathy
基本 15 次元 (SVM)	0.786	0.797
基本 15 次元 (RF)	0.875	0.876
15 + 384 次元 (SVM)	0.741	0.783
15 + 384 次元 (RF)	0.872	0.903

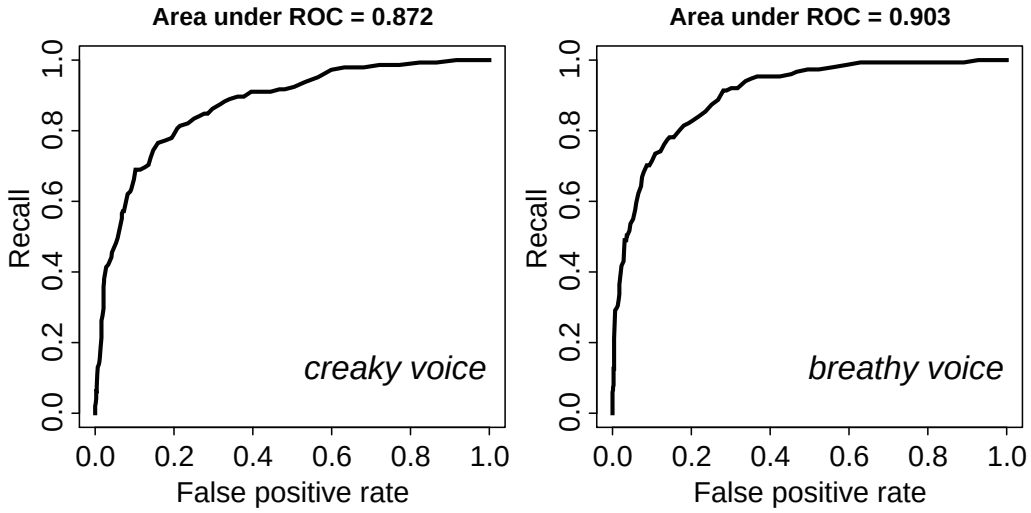

図 1　ROC 曲線

フォレストの方が性能が高いことがわかる。また、openSMILE で抽出した 384 次元を追加することで精度が向上したケースは少なく、特徴の追加の効果は限定的であったと言える。図 1 に、特徴量として基本 15 次元 + openSMILE 384 次元を用い、ランダムフォレストにより分類を行った場合の偽陽性率 (false positive rate) と再現率 (recall) の関係、いわゆる ROC 曲線を示す。この図は、例えばラベラーが creaky voice または breathy voice と判定した母音のうち 90% 以上を検出しようとすれば、ラベラーが creaky voice または breathy voice と判定しなかった母音のうちそれぞれ 39.6% および 28.1% 以上は誤って creaky voice または breathy voice と判定されてしまうことを意味する。

次に、用いた特徴量の重要度を、ランダムフォレストにおける平均不純度減少量を基準に調べた。Creaky voice 検出および breathy voice 検出における重要特徴量の上位を、表 3 (基本 15 次元) および表 4 (基本 15 次元 + openSMILE 384 次元) に示す。 基本 15 次元だけを特徴量として用いた場合、C1TL は、creaky voice 検出の重要度が高い特徴としてコンスタントに上位を占めていた。また、周期性のゆらぎである PPQ5, APQ5 も上位を占めていた。この結果は、スペクトル傾斜が小さく周期性が低い creaky voice の特徴と符合している。C1TL, PPQ5, APQ5 はまた、breathy voice 検出の重要度が高い特徴でもあった。この結果

表 3　重要特徴量 (基本 15 次元: 上位 10)

creaky voice		breathy voice	
0.44	C1TL	0.42	C1TL
0.43	PPQ5	0.41	PPQ5
0.41	APQ5	0.4	HNR
0.36	sdPitch	0.39	APQ5
0.35	H1−H2	0.38	sdPitch
0.35	HNR	0.32	H1−H2
0.33	F1	0.32	H1−A1
0.33	F2	0.31	F1
0.32	H1−A1	0.31	F3
0.31	H1−A3	0.3	H1−A3

表 4　重要特徴量 (基本 15 次元 + openSMILE 384 次元: 上位 10)

creaky voice		breathy voice	
0.58	RMSenergy range	0.67	ΔMFCC$_7$ maxPos
0.57	MFCC$_3$ min	0.66	MFCC$_2$ range
0.56	MFCC$_4$ stddev	0.62	MFCC$_1$ stddev
0.56	MFCC$_2$ minPos	0.62	MFCC$_3$ maxpos
0.54	RMSenergy stddev	0.59	ΔMFCC$_4$ minpos
0.54	MFCC$_2$ amean	0.57	MFCC$_2$ linregc1
0.54	MFCC$_1$ stddev	0.57	MFCC$_3$ minpos
0.53	F2	0.56	MFCC$_2$ minpos
0.53	MFCC$_3$ max	0.55	F1
0.53	C1TL	0.55	sdPitch

は、スペクトル傾斜が大きく周期性が低い breathy voice の特徴と符合している。基本 15 次元と openSMILE の 384 次元を併用した場合、creaky voice, breathy voice ともに、スペクトルの情報である MFCC に関連した特徴が上位を占めている。また、creaky voice についてはRMSenergy(実効値) すなわち音の強さに関連した特徴が含まれている。

6. おわりに

本研究では、『日本語話し言葉コーパス』(CSJ) に含まれる母音の発声様式 (creaky, breathy) のラベリングを行い、機械学習による声質の自動アノテーションの可能性を探った。

サポートベクターマシン (SVM) およびランダムフォレスト (RF) を用いた声質の分類実験の結果、RF の性能が SVM の性能を上回った。しかしながら、検出性能の指標である F 値は creaky voice で最大 0.59, breathy voice で最大 0.62 程度であり、人手によるアノテーション

を代替するほどの精度は得られないことがわかった。

本研究ではまた、過去の研究で用いられてきた発見的な特徴量に加え、openSMILE と呼ばれる音響特徴量抽出ソフトウェアを用いた大規模な特徴量セットを併用することの有効性も検証したが、その効果は限定的であった。

本稿で述べた手法を声質アノテーションに応用しようとする場合、専門家によるアノテーション作業の補助に使うことが考えられる。しかし、本稿で述べたように、例えば再現率 90% を目標とすると、偽陽性率は 30% から 40% 程度となり、人手による確認の負荷はかなり大きくなることが予想される。今後は、ラベラー間一致度などを基に検出した非 modal 声質の信頼度を推定する方法を検討し、自動アノテーションの実用化につなげたい。

謝辞

本研究は国立国語研究所コーパス開発センターの共同研究「コーパスアノテーションの拡張・統合・自動化に関する基礎研究」および科研費（課題番号 26284062, 研究代表者 前川喜久雄）の成果である。

文　献

森大毅・前川喜久雄・粕谷英樹 (2014).『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—』　コロナ社.

John Laver (1980). *The Phonetic Description of Voice Quality. Cambridge Studies in Linguistics.*: Cambridge University Press.

Ailbhe Ni Chasaide, and Christer Gobl (1997). “Voice source variation.” William J. Hardcastle, and John Laver (Eds.), *Handbook of Phonetic Sciences*. Oxford: Blackwell. pp. 1–11.

Kikuo Maekawa, and Hiroki Mori (2016). “Voice-Quality Difference Between the Vowels in Filled Pauses and Ordinary Lexical Items.” *Proc. Interspeech 2016*, pp. 3171–3175.

Carlos Toshinori Ishi, Ken-ichi Sakakibara, Hiroshi Ishiguro, and Norihiro Hagita (2008). “A Method for Automatic Detection of Vocal Fry.” *IEEE Transactions on Audio, Speech, and Language Processing*, 16, pp. 47–56.

Toshiyuki Sadanobu (2004). “A natural history of Japanese pressed voice.” 音声研究, 8:1, pp. 29–44.

Dennis H. Klatt, and Laura C. Klatt (1990). “Analysis, synthesis, and perception of voice quality variations among female and male talkers.” *Journal of Acoustical Society of America*, 87:2, pp. 820–857.

Adas Gelzinis, Antanas Verikas, and Marija Bacauskiene (2008). “Automated speech analysis applied to laryngeal disease categorization.” *Comput. Methods Prog. Biomed.*, 91:1, pp. 36–47.

前川喜久雄・森大毅 (2005).「フィラーの声質上の特徴に関する予備的分析」　日本音響学会講演論文集, pp. 293–296.

Matthew Gordon, and Peter Ladefoged (2001). “Phonation types: a cross-linguistic overview.” *Journal of Phonetics*, pp. 383–406.

Florian Eyben, Martin Wöllmer, and Björn Schuller (2010). “openSMILE: The Munich versatile and fast open-source audio feature extractor.” *Proc. ACM Multimedia*, pp. 1459–1462.

Björn Schuller, Stefan Steidl, and Anton Batliner (2009). “The Interspeech 2009 Emotion Challenge.” *Proc. Interspeech 2009*, pp. 312–315.