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

2008－MUS－74 (31)
2008－SLP－70 (31)
2008／2／9

音響的特徴に基づくノンバーバル発話の意図識別

吉川哲史† 牧本慎平† 柏岡秀紀†‡§ ニック キャンベル†‡§
†奈良先端科学技術大学院大学 情報科学研究科
‡情報通信研究機構 知識創成コミュニケーション研究センター
§国際電気通信基礎技術研究所 音声言語コミュニケーション研究所
{satoshi-yo, shimpei-m, kashioka, nick}@is.naist.jp

あらまし 多くの対話システムでは応答を言語情報のみを使って生成しており，語彙としては意味をもたないノンバーバル発話は使われて来なかった．しかし，「うん」，「えー」などのノンバーバル発話は意図や態度，感情などのパラ言語情報を伝達している．本稿では，ノンバーバル発話の意図を肯定，相槌，否定の3つとし，これらの意図を音響的特徴を用いたサポートベクターマシンによる識別手法を提案する．一人話者で構築した識別器に対して他話者への適用を行い，ある話者の音響的特徴によって，他話者の意図が識別できるか実験を行った．

Intention classification in non-verbal utterance based on acoustic features

Satoshi Yoshikawa† Shimpei Makimoto† Hideki Kashioka†‡§ Nick Campbell†‡§
†Nara Institute of Science and Technology
‡National Institute of Information and Communications Technology
§Advanced Telecommunications Research Institute International
{satoshi-yo, shimpei-m, kashioka, nick}@is.naist.jp

Abstract Paralinguistic features of speech communication are now beginning to be included in interactive speech processing devices. Nonverbal speech sounds such as "erm, yeah, ahah, etc.", are frequent in conversational speech and can carry different meanings according to speaker intentions. This paper discusses the acoustic characteristics of non-verbal speech communication and attempts to distinguish "positive, negative, and backchannel utterances" in a speaker-independent manner. We tested 2 statistical learning techniqes to distinguish three speaker intentions from a set of backchannel and non-verbal utterances taken from a natural conversational speech corpus.

1 はじめに

会話の中で頻繁に用いられるものに，「うん」や「ええ」と言ったノンバーバル発話がある．ノンバーバル発話は文字だけでは意味を持たない．しかし，韻律や声質などの発話スタイルを変えることで意図や態度を伝達している．ノンバーバル発話の理解は，人間の円滑なコミュニケーションを考える上で欠かせないものである．ノンバーバル発話を含めたノンバーバルコミュニケーションの研究は，近年盛んに行われている [1, 2]．ノンバーバル発話を扱った研究としては，「ん」が持つ発話印象を疑念/確信，肯定/否定，好印象/悪印象の3次元に定め，これらの発話印象の韻律特徴の分析と合成を行ったもの [9, 10] や，「え」について，肯定，感心，非難などの15個の知覚ラベルを定義し，韻律，声質につ

いて分析したもの [6] がある．また，文献 [7] では，「うん」をパラ言語ラベルや対人態度ラベルなどを用い，韻律特徴との相関を分析している．これらの研究は，個別の「ん」や「え」を対象にしたものである．

ノンバーバル発話は，曖昧性が多く「うん」が実際には「ああ」に近い発声になると言う特徴がある．また，コーパスに現れる発話などを見てもノンバーバル発話は非常に種類，発声のバリエーションが多い．文献 [8] によると，「ん」の印象表現の持つ f0 変化特性が，印象表現を伴う一語発話においても同様に反映されることが示されており，同義のノンバーバル発話が，同様の音響的特徴を持つ可能性を示唆している．また，我々は同義のノンバーバル発話が同様の音響的特徴を持つかを調べるために，[11] において，「うん」を用い，肯定，相槌，否定の 3 意図についてのサポートベクタマシン（以下 SVM）による識別器を構築した．その結果，他発話への適用において発話継続長の近い「ふん」と「はい」では，F 値で 0.7 の精度を得られている．

本稿では，[11] で構築した SVM を元に新たな特徴量を加え，他話者への適用実験を行う．あわせて HMM による意図識別の結果も報告する．

2 音声資料

音声資料には，自然対話コーパスである ESP_C コーパスを用いる．バランスの良い音声データを得るには，各意図について発話してもらう方が良い．しかし，演技をしながら発話をした音声では，普段の会話に出てくる発声を得られにくいと考え，自然対話コーパスを用いている．

2.1 ESP_C コーパス

ESP_C コーパス[3] は，JST/CREST の「表現豊かな発話音声のコンピュータ処理システム」プロジェクトによって作成された，JST/ATR Expressive Speech Processing Corpora の一部である．1 セッション 30 分で，それぞれのペアについて約 10 セッション収録されている．話者は，母語が日本語の男性 3 名，女性 3 名，中国語の男性 1 名，女性 1 名，英語の男性 1 名，女性 1 名の計 10 名である．本稿では，日本人女性 2 名と日本人男性 1 名の計 3 名を対象とした．図 1 に ESP_C の書き起こし例を示す．書き起こしには，発話者，発話開始時間，発話時間が含まれている．

```
JFA_JMA_E04 596.440 1.518 いやーんでももうね
JFA_JMA_E04 598.023 0.947 あのー
JFA_JMA_E04 598.994 1.427 一回やってることだから
JMA_JFA_E04 600.335 0.446 うん
JFA_JMA_E04 600.782 0.499 うん
JMA_JFA_E04 601.070 0.440 まー
JMA_JFA_E04 601.543 0.590 ほんで
JFA_JMA_E04 601.598 0.398 また
JMA_JFA_E04 602.189 0.800 そんな
JMA_JFA_E04 603.245 1.133 あれもないし
```

図 1 ESP_C コーパスの例

10073 うん	1084 あっ
8608 はい	981 は あい
3487 うーん	942 あの
2906 ええ	941 ふーん
1702 はーい	910 そう
1573 うーーん	749 えー
1348 ズー	714 あーー
1139 ふん	701 あ
1098 あのー	630 あーーー

図 2 書き起こし行単位での単独発話の順位

図 2 は書き起こし行単位での単独発話の順位である．これを見ると，1 万発話ある「うん」以外にも，「ええ」や，「あーー」と言った，ノンバーバル発話が多く見られることが分かる．本稿では，最も発話数の多い「うん」を代表的な発話と定め，他の発話に，2 番目に多い「はい」，他に「ふん」，「うーん」を用いた．

2.2 利用するデータの抽出

コーパスから切り出した音声の中から利用するデータを抽出するために，主観評価による意図（肯

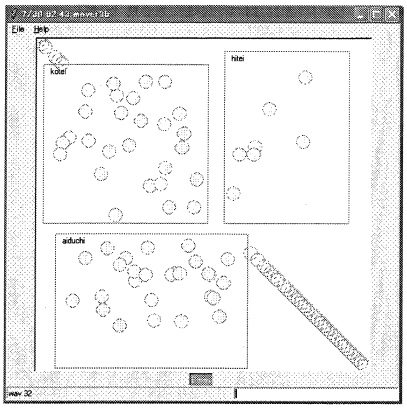

図 3　mover の使用画面

表 1　JFA の 4 人以上一致発話

	肯定	相槌	否定	合計
うん	65	109	48	222
はい	25	9	0	34
ふん	8	51	3	62
うーん	13	25	15	53

表 2　JFC，JMA「うん」の 3 人以上一致発話

	肯定	相槌	否定	合計
JFC	27	134	1	162
JMA	23	96	6	125

定，否定，相槌）の分類を行った

2.2.1　データの分類条件

ESP_C の JFA の「うん」400 発話，「はい，ふん，うーん」各 100 発話では 5 名，JFB，JMA の「うん」各 400 発話では 4 名に分類してもらった．音声は文脈なしで聞いてもらい，各発話について，肯定，否定，相槌のどの意図にあてはまるか，図 3 で示す，mover[3] という音声分析ツールを用いて，選択してもらった．尚，文脈なしであるのは，人が意図の識別に用いている音響的特徴を調べるには，音声の持つ情報のみで評価する必要があると考えたからである．

2.2.2　抽出されたデータ

データの中から分類者 5 名のデータに関しては，一致率 80% を基準に，4 名（一致率 80%），分類者 4 名では，3 名（一致率 75%）以上の一致率が得られたものを正解データとして使用することとする．JFA「うん，はい，ふん，うーん」で，4 人以上が同じ意図と判断した発話数を表 1 に示す．JFC，JMA「うん」で，3 人以上が同じ意図と判断した発話数を表 2 に示す．

表 1 の結果から，JFA「うん」においては，全体の 56 % が 4 人以上の一致となった．他の発話においても 4 割強の発話を取得できた．しかし，「はい」については，否定と分類したものが少なく，2 人までしか否定と判断したものはなかったため，このデータからは否定を取得できなかった．「はい」は語彙自体の意味の影響があり，肯定に分類がよっている．

表 2 の結果では，JFC，JMA については，200 発話中 162 発話，125 発話と非常に多くの発話が 3 人以上判断が一致している．しかし，相槌に偏る結果となってしまい，否定も JFC では 1 個，JMA でも 6 個と極端に少ないものとなってしまった．否定をテストデータに用いるにはあまりに少ないため，本稿では，実験に用いるデータとして，JMA の肯定，相槌，JFC の肯定，相槌を用いることとした．

3　機械学習による自動識別実験

JFA「うん」の音響的特徴を用いて，他話者 JFC，JMA の意図識別の可能性について実験する．そのために，機械学習によって，JFA「うん」の肯定，相槌，否定の意図識別器を作成し，他話者 JFC，JMA の「うん」への意図識別実験を行った．機械学習のモデルには，サポートベクターマシン，比較のために韻律を付与した HMM を用いる．

サポートベクターマシン（SVM）[4] は，他クラスとのマージンが最大になるよう境界面を設定するので，汎化能力に優れ，少ないデータでも高い精度で識別できると言う特徴を持っている．一方，HMM は連続的かつ伸縮しうる信号列のパターン抽出に優れるている．時系列データであるスペクトル情報を確率モデルで定義できるため，音声との親和性が高

い．また，韻律はパラ言語情報の理解に重要なパラメータであるので，これを使用できる韻律を付与したHMMを用いた．他にも機械学習手法は提案されているが，対象データの特徴を考慮して，この2つを比較するモデルとした．

SVMには，統計解析環境であるRのe1071パッケージを用いた．HMMのモデル作成には，スペクトル情報に加えて韻律を付与したHMMを作成できる，HMM-based Speech Synthesis System (HTS)[5]のHMMモデルを用いた．

3.1 実験条件

使用する発話は，JFA「うん」220発話 (肯定65，相槌108，否定47)，「ふん」，「はい」，「うーん」．JFC「うん」162発話 (肯定27，相槌134)，JMA「うん」125発話 (肯定23，相槌96) である．JFC，JMA共に否定発話が極端に少ないため，今回は評価対象からはずしている．

SVMは，肯定，相槌，否定の3クラス分類のSVMを構築する．学習データは，JFAの「うん」だけを用いたものと，JFAの発話全てを用いたもの2種類用意した．使用する特徴は3.1.1節で述べる23個である．カーネルには，radialを使用した．

HMMは，肯定，相槌，否定の3単語認識とし，肯定，相槌，否定の3単語のHMMを学習した．学習データは，JFA「うん」である．HMMに使用する特徴としては，MFCC，δMFCC，logf0，δf0を用いた．

3.1.1 使用する特徴量

SVMに使用する音響的特徴は，以下のものである．各特徴は話者性を取り除くために，各特徴でセンタリングを行い，それぞれの値を各特徴の標準偏差で割ることによって標準化した．

ピッチの特徴として，f0の平均 (log fmean)，最大 (log fmax)，最小 (log fmin)．f0のピーク位置の割合 (fpct)，音の有声度を表す (exp fvcd)，f0の標準偏差の (log fsd)．パワーの特徴として，パワーの平均 (log pmean)，最大 (log pmax)，最小 (log pmin)．パワーのピーク時間の割合 (ppct)．発話継続時間 (log duration)．声質の特徴として，第1倍音の振幅 (h1)，第3ホルマントの振幅 (a3)，この二つの差を取った，(h1a3)，第1倍音との第2倍音の差を取った (h1h2)．スペクトル情報として，第1から第8ホルマント平均 (log f1- log f8)．以上，23個を音響的特長として使用する．

3.1.2 評価尺度

性能評価の尺度として，以下の式で示される，F値 (F-measure) を用いる．

$$\text{精度}\ (P) = \frac{\text{識別に成功した発話数}}{\text{識別された発話数}}$$

$$\text{再現率}\ (R) = \frac{\text{識別に成功した発話数}}{\text{正解意図の全体数}}$$

$$F\ \text{値} = \frac{2P \times R}{P + R}$$

3.2 実験結果

表3に「うん」のSVMによる3分割の交差検定による推定精度，HMMによる10分割交差検定による推定精度を示す．HMMの精度の高さが目立つ結果となっている．

表3　交差検定による推定精度

	精度		
	肯定	相槌	否定
SVM	0.71	0.93	0.82
HMM	0.92	0.9	0.91

次に，表4に他話者JFC，表5に他話者JMA「うん」にSVMを適用した結果を示す．表4の右の列にあるjfc(un+)とは，「うん」発話に加えて「ふん，はい，うーん」もSVMの学習に用いた場合の結果である．「うん」だけの学習では，肯定で5割程度の精度であったものが，他の発話も追加すると6割になり，精度が向上することが分かる．しかし，再現率は落ちており，F値では，微増にとどまっている．しかし，「うん」だけ，他発話追加ともにF値で0.6程度を示しており，前回の[11]の他発話「ふん」などに0.1ポイント程度下回る結果で識別できている．JMAについては，精度は「うん」だけの場合で，肯定で3割であった．他発話を追加したJMA(un+)でも4割と識別制度は良くない．

表4　SVMによる他話者JFCへの適用結果

	JFC(un)		JFC(un+)	
	肯定	相槌	肯定	相槌
精度	0.49	0.95	0.61	0.92
再現率	0.74	0.73	0.63	0.81
F値	0.59	0.82	0.62	0.86

表5　SVMによる他話者JMAへの適用結果

	JMA(un)		JMA(un+)	
	肯定	相槌	肯定	相槌
精度	0.32	0.85	0.41	0.82
再現率	0.52	0.63	0.39	0.76
F値	0.4	0.72	0.4	0.79

最後に，HMMでの結果を表6に示す．HMMでは，ほとんどのデータが相槌に振られてしまい認識が上手く行っていないことがわかった．

表6　HMMの他話者への適用結果，

	JFC		JMA	
	肯定	相槌	肯定	相槌
精度	0.4	0.79	0.29	0.70
再現率	0.15	0.88	0.78	0.94
F値	0.22	0.83	0.42	0.80

4　考察

SVMとHMMの2つの手法による精度の比較を行った．他話者においてはSVMの方が優れた結果となった．HMMは，SVMと比較して時系列データを扱っている分使える特徴は多く，SVMに比べて有利である．しかし，時系列情報全てを用いてしまったことで，HMMが学習話者に適合しすぎてしまった可能性が考えられる．その点，SVMは時系列データのように詳細な変化を観測していないので，発話全体の特徴を捉えており，他話者へもある程度適用できたと考える．

SVM，HMM共に，他話者の女性が他話者の男性より精度がよくなっている．これは，話者性が影響したのではないかと考えられる．SVMについては，話者性の除去に関して，標準化により男性，女性のf0の違いは考慮し，logを掛けるなどにより，ヒストグラムを正規分布に近づけると言ったことを行った．HMMに関しては，log f0とδ f0が用いられており，ある程度の話者性は吸収されている．しかし，実際には精度に差異が生まれており，現時点での結果からは，話者性の中でも，男性・女性の違いなのか，発話スタイルの違いなのかは判断できない．

また，SVMでは，「うん」だけではなく，「ふん，はい，うーん」も学習に加えた場合，肯定では，精度が上がる結果となった．（本稿では取り上げていないが，HMMでは精度の向上は見られなかった．）JFC，JMA共に同様の傾向を示しており，JFAの発話のバリエーションが増えたことにより肯定表現が増え，精度が上がったことが考えられる．しかし，JFC，JMAともに肯定表現においては再現率が低下している．これは，相槌表現のバリエーション増加による相槌へのご認識の増加が原因と考えられる．

5　おわりに

ノンバーバル発話の肯定，相槌，否定の3意図識別器を，女性話者JFAの「うん」または，「うん，はい，ふん，うーん」の音響的特徴を用いたSVMによって構築した．「うん，はい，ふん，うーん」で構築したSVMでは，他話者の「うん」発話において，肯定であれば4-6割程度，相槌であれば，8-9割の識別精度を得ることが出来た．否定表現による評定を得られていないが，肯定-相槌の間であれば，他話者においても，一人話者で構築したモデルで，同姓であれば，肯定意図が6割程度の精度で識別できた．また，「うん」だけを学習に用いたモデルよりも「うん，はい，ふん，うーん」全てを用いたモデルの方が他話者の「うん」であっても，精度が良くなることから，ノンバーバル発話では語彙に関係なく意図に共通の表現が存在すると言える．これは，「うん」や「はい」のような種類や，発声の曖昧性を考慮したモデル作成が可能であることを示している．

今後の課題としては，データ抽出では，他話者

JFC，JMA について，否定発話をほとんど抽出できなかったことから，コーパス中から効率よくデータを抽出する方法を検討しなくてはならない．また，男性話者への対応や，さらなる精度の向上のために，学習データを増やすことや，実際の会話には，肯定，相槌，否定以外にも多くの意図があることから，例えば，[9, 10] のように多次元なものや，ESP[3] の発話行為ラベリングなども，実対話システムを考える上では導入を検討しなくてはならない．また，実対話システムでのノンバーバル発話抽出を考えるとき，本稿では書き起こし文内の単独発話からノンバーバル発話と思われる発話を抽出し実験に使用したが，文中の発話も積極的に抽出する必要がある．これらの抽出方法としては，文中からの非語彙的表現の抽出法 [12] が考えられる．

参考文献

[1] Anna Esposito, Maja Bratanic, Eric Keller, and Maria Marinaro, editors. *Fundamentals of Verbal and Nonverbal Communication and the Biometric Issue.* Nato Security Through Science Series E: Human and Societal Dynamics. Ios Pr Inc, 1 edition, 5 2007.

[2] Anna Esposito, Marcos Faundez-Zanuy, Eric Keller, and Maria Marinaro, editors. *Verbal and Nonverbal Communication Behaviours*, Vol. 45 of *Lecture Notes in Computer Science.* Springer, 2007.

[3] Feature extraction and analysis for speech technology. http://feast.atr.jp/.

[4] V.N. Vapnik. The Nature of Statistical Learning. *Theory, Springer*, 1995.

[5] H. Zen, T. Nose, J. Yamagishi, S. Sako, T. Masuko, A.W. Black, and K. Tokuda. The HMM-based speech synthesis system version 2.0. In *ISCA SSW6*, Bonn, Germany, august 2007.

[6] 石井カルロス寿憲, 石黒浩, 萩田紀博. 韻律および声質を表現した音響特徴と対話音声におけるパラ言語情報の知覚との関連. 情報処理学会論文誌, Vol. 47, No. 6, pp. 1782–1793, 6 2006.

[7] 梅野淳史. 発話「うん」の音響的特長に基づくパラ言語・非言語情報の分析. Master's thesis, 奈良先端科学技術大学院大学, 2003.

[8] グリーンバーグ陽子, 津崎実, 加藤宏明, 匂坂芳典. 入力語彙情報に基づく対話韻律制御. 日本音響学会 2005 年秋季研究発表会講演論文集, pp. 273–274, 9 2005.

[9] 渋谷渚, グリーンバーグ陽子, 匂坂芳典. 基本周波数特性に基づく一語発話「ん」の分類について. 日本音響学会 2005 年秋季研究発表会講演論文集, pp. 271–272, 9 2005.

[10] 李克, グリーンバーグ陽子, 渋谷渚, 匂坂芳典. 印象表現によるパラ言語情報を用いた韻律制御. 日本音響学会 2006 年秋季講演論文集, pp. 233–234, 9 2006.

[11] 吉川哲史, 柏岡秀紀, ニックキャンベル. 発話「うん」のの音響的特徴に基づく意図の自動識別. 日本音響学会誌, pp. 243–246, 2007.

[12] 牧本慎平, 吉川哲史, 柏岡秀樹, ニックキャンベル. 統計学習を用いた対話からの非語彙的表現の抽出. 本誌, 2008.