7Q-05

会話内非言語音声情報抽出のための音響特徴量の検討

柴田 健作† 中村 圭佑† 中臺 一博†
（株）ホンダ・リサーチ・インスティチュート・ジャパン†

1. はじめに

音声処理分野では，コンピュータに人間の会話を理解させるため音声認識や自然言語処理の研究が行われている．これらの研究では入力が言語音声のみであることを仮定しているため，笑い声や咳などの非言語音声を含む自然な会話の扱いが困難である．そこで，会話音声からの非言語音声情報抽出について検討する．ここで非言語音声情報抽出とは，言語音声，非言語音声，音声ではない環境音をクラス分類することを示す．本稿では，非言語音声は笑い声と咳を扱い，それぞれを独立したクラスとして扱う．

非言語音声情報の学習に関する先行研究として Björn[1]らが INTERSPEECH2013 Social Signals Sub-Challenge で発表した，笑い声と相槌の分類が挙げられる．Björn らは Harmonic-to-Noise Ratio(HNR)や基本周波数(F0)など複数の音響特徴量を組み合わせた 141 次元の特徴量セットを用いて Support Vector Machine(SVM)による笑い声と相槌の識別を行い，各クラスに対して Area Under the receiver operating Curve(AUC)評価で 82.9%と 83.6%と高い識別性能を報告している．この結果は非言語音声のクラス識別には複数の音響特徴を組み合わせて用いることが有効であることを示唆している．これは非言語音声にはクラスを決定づける顕著な特徴が少ないためだと考えられる．我々は，音圧の高い部分には，クラスを決定づける顕著な特徴が多く含まれると考え，Björn らの手法を音圧を考慮するよう拡張する手法を提案する．

2. 検討手法

2.1 音響特徴量の検討

INTERSPEECH2013 で用いられた特徴量を基にした 4 種類の音響特徴量セットおよび，一般的な音声特徴として MFCC ベースの特徴量セットの計 5 種類の特徴量セットに対して，提案法を適応した場合としない場合について比較検討を行う．用いた特徴量セットの詳細を以下に示す．

(1) INTERSPEECH2013 COMPARE 特徴量セット［2］6373 次元
(2) Björn らの用いた特徴量セット 141 次元
(3) (1)の特徴量セットのうちフレームベースで有効な特徴量のセット 426 次元
(4) 特徴量セット(3)をデータ分析した結果，クラス分類に有用だと思われた特徴量を抽出した特徴量セット 186 次元
(5) MFCC，MFCC の一階微分と音圧の一階微分の特徴量セット 25 次元

(1)は音響イベントベースの特徴量セット，(2)～(5)はフレームベースの特徴量セットとなっている．(4)のデータ分析にはデータ分析ツール Weka[4]を用い，各特徴量に対する各クラスのデータ分布を分析した．

2.2 音圧によるフレーム選択の検討

本稿では各音響イベントの平均音圧以上のフレームのみを非言語音声情報抽出対象として選択する．これにより，信号対雑音比が向上し，良好な識別性能が得られることが期待できる．この手法は 2.1 節で示した特徴量セットのうち(2)～(5)のフレームベースの特徴量セットに対して適用した．提案するフレーム選択処理について説明する．まず，fフレーム目の入力音響信号を短時間フーリエ変換して得られる$\mathrm{X}(\omega, f)$から，k番目の音響イベントを検出する．ここで，ωを周波数，音響イベントが検出された区間を$\underline{F_k} \le f \le \overline{F_k}$とする．フレーム単位の音圧$P_k(f)$を式(1)のように計算する．

$$P_k(f) = \frac{1}{\overline{\omega} - \underline{\omega} + 1} \sum_{\omega=\underline{\omega}}^{\overline{\omega}} \mathrm{X}(\omega, f) X^*(\omega, f)$$

ここで，$()^*$は複素共役演算子を表す．また，$\overline{\omega} \le \omega \le \underline{\omega}$は，処理に用いた周波数帯域を表し，本稿では音声帯域である$500\mathrm{Hz} \le \omega \le 2800\mathrm{Hz}$とした．平均音圧$\overline{P_k}$を以下のように計算する．

$$\overline{P_k} = \frac{1}{\overline{F_k} - \underline{F_k} + 1} \sum_{\underline{F_k}}^{\overline{F_k}} P_k(f)$$

フレーム選択は，$\underline{F_k} \le f \le \overline{F_k}$に対し，$P_k(f) < \overline{P_k}$となるフレームの特徴量を棄却することにより行った．

3. 評価実験

実験データには「音声チャットを利用したオンラインゲーム感情音声コーパス(OGVC)」［5］を利用し，

A study of acoustic features for extra-linguistic information extraction in natural conversations
† Kensaku Shibata, Keisuke Nakamura, and Kazuma Nakadai (Honda Research Institute Japan)

発話の転記テキストをもとに音データから言語音声，非言語音声(笑い声，咳)，環境音を抽出した．個人差や録音環境，録音機器に対するロバスト性も評価するため，学習データとテストデータには異なる話者のデータを用い完全なオープンテストを実施した．表 1 に学習とテストに用いたデータ数を示す．

	非言語音声		言語音声	環境音
	笑い声	咳		
学習	150	15	150	150
テスト	80	2	80	80

表 1　実験に用いた各分類クラスのデータ数

OGVC の音データから抽出できる咳クラスの音声データが少ないため，他クラスに比べて学習・テストのデータ数が少なくなっている．学習はアルゴリズムに Sequential Minimal Optimization(SMO)を用いた SVM で行い，評価は適合率(P)と再現率(R)で行った．特徴量抽出に音圧を考慮しなかった場合の非言語音声情報抽出結果を表 2 に，音圧を考慮した場合の結果を表 3 に示す．ただし，表の縦軸ラベルは特徴量セットの番号を表し，評価結果の値は小数点第三位を切り捨てて表す．また，フレームベース特徴量セットを用いた評価実験でのそれぞれの分類クラスの評価結果において，音圧を考慮しない場合音圧を考慮する場合の結果を比較し，良かったほうのスコアを太字で表す．

	非言語音声				言語音声		環境音	
	笑い声		咳					
	P	R	P	R	P	R	P	R
(1)	0.94	0.74	0.00	0.00	0.80	1.00	1.00	1.00
(2)	0.87	**0.65**	0.00	0.00	0.74	0.84	0.98	0.99
(3)	**0.95**	0.42	**0.01**	**0.05**	0.65	0.89	0.96	0.99
(4)	**0.95**	**0.45**	0.00	**0.03**	0.68	0.88	0.95	0.99
(5)	0.00	0.00	0.00	0.00	0.00	0.00	0.56	**1.00**

表 2　音圧を考慮しない場合の評価結果

	非言語音声				言語音声		環境音	
	笑い声		咳					
	P	R	P	R	P	R	P	R
(2)	0.87	0.64	0.00	0.00	**0.76**	**0.94**	**0.99**	0.99
(3)	0.90	**0.47**	0.00	0.00	**0.69**	**0.96**	**0.97**	0.99
(4)	0.92	0.44	0.00	0.00	**0.70**	**0.97**	**0.96**	0.99
(5)	**0.58**	**0.17**	0.00	0.00	**0.66**	**0.60**	**0.65**	0.94

表 3　音圧を考慮する場合の評価結果

咳クラスの適合率と再現率が他クラスに比べて低いことがわかるが，これは学習データ・テストデータが共に他クラスと比べて極端に少ないためだと考えられる．特徴量セット別に比較すると，特徴量セット(5)は音圧を考慮しない場合の環境音の再現率を除く全ての場合において，適合率と再現率が他の特徴量セットよりも低く，特に非言語音声の分類クラスにおいてその傾向が強い．これは，非言語音声が音声信号に特定の調波構造を持つとは限らず，音声特徴量のみで表現することが難しいためだと考えられる．また，咳クラスの結果を除くと特徴量セット(1)が最も良い結果を出している．フレームベースの特徴量セットの場合，学習・テストのデータ数はフレームを抽出する音データの長さに依存するため，クラス間のデータのばらつきが生じる．特徴量セット(1)が最も良い結果を出した理由は他の特徴量よりも高次元の音響特徴量を用いたためだと考えられるが，笑い声，言語音声，環境音のクラス間のデータのばらつきがなかったことも影響している可能性がある．特徴量抽出に音圧を考慮しない場合と音圧レベルを考慮する場合を比較すると，音圧を考慮しない場合のほうが高いスコアをとる実験もあるが，多くの実験において音圧を考慮する場合のほうが高いスコアとなった．また，音圧を考慮しない場合での性能向上は最大でも 0.06 と微少であるのに対し，音圧を考慮する場合では平均で 0.16 と大きく性能向上している．これらのことから，特徴量抽出に音圧を考慮することが非言語音声情報抽出に有用であるといえる．

4. おわりに

本稿では音圧を考慮した特徴量の抽出方法に着目して，会話音声からの非言語音声情報抽出を行った．評価実験の結果，特徴量抽出に音圧を考慮することが非言語音声情報抽出に有用であることを示した．一方，今回検討した特徴量セットでは咳クラスの結果を除くと(1)が最も良い非言語音声情報抽出性能を示したことから，特定の音声特徴量のみを用いて非言語音声情報抽出を行うことは難しいという知見が得られた．

本稿では非言語音声のサブクラスとして笑い声と咳の 2 クラスのみを対象とした．また咳クラスに関しては十分な量のデータを用いることができなかった．そのため，サブクラス数を増やすとともに，咳クラスを含む各クラスに関して十分な量のデータを用いて，提案がより一般化された非言語音声情報抽出に有用であるかを今後確認したい．また，SMO ベースの SVM を用いたが，GMM や DNN などのフレームベースの識別器の性能比較や，HMM などの音響信号の時間的変化を考慮した識別器についても今後検討したい．

参考文献

[1] Björn Schuller et al.: The interspeech 2013 computational paralinguistics challenge: social signals, conflict, emotion, autism (2013)

[2] Florian Eyben et al.: Affect recognition in real-life acoustic conditions - A new perspective on feature selection (2013)

[3] Florian Eyben et al.: *open-Source Media Interpretation by Large feature-space Extraction*, 2.1 edition (2014)

[4] Mark Hall et al.: The weka data miningsoftware: An update. *SIGKDD Explorations*,Vol. 11 (2009)

[5]有本泰子ら: 音声チャットを利用したオンラインゲーム感情音声コーパス. 日本音響学会2013年秋季研究発表会講演論文集, No.1-p-46a, pp.385-388 (2013)