音響特徴量を用いた吹き出しテキストの生成*

☆松宮 翔，Sakriani Sakti，Graham Neubig，戸田 智基，中村 哲 (奈良先端大)

1 はじめに

字幕放送はテレビ番組のナレーションやドラマのセリフなどの音声を文字にして伝える放送であり，テレビの音が聞き取りにくい高齢者や聴覚障害者への重要な情報提供手段である．そして，近年の音声認識技術の発展によって自動字幕付与技術の拡充が進んでいる．しかし，従来の字幕付与では，音声をテキストに書き起こすのみでとどまり，コミュニケーションの分野において重要な役割を果たす感情の表現が欠如している．

字幕において感情を表現する方法としてテキストのサイズや色の変更などが挙げられる．しかしながら，テキスト表記に対して変更を施す方法は，テキストの視認性を劣化させる可能性がある．そこで本稿では，マンガ等で感情をうまく表現するツールとして用いられている吹き出しに着目する．音声に内在する感情に合った形をもつ吹き出しを，字幕と合わせて付与することにより，テキストだけでは伝わらない感情も視聴者に伝えることができるシステムを提案する．その実現に向けて，本稿では音響特徴量と言語特徴量の 2 種類の特徴量を用いて，音声に適した吹き出しを分類した結果を報告する．

2 アニメ音声と吹き出し

本稿では，アニメのデータを対象とする．アニメは，感情を多く含む音声で構成されている．また，アニメには対応するマンガが存在する場合が多い．マンガでは，様々な感情に合った吹き出しの形状を用いて話者の感情を表現されている．この吹き出しの技術を字幕にも導入することによって感情を伝えることが可能となると期待される．また，音声からの吹き出し分類技術を構築する上で重要となる学習データに関しても，アニメとマンガは大量に入手できるという利点がある．本稿では，怒りや，悲しみ，喜びなどの様々な感情を含むアニメとして，数ある作品の中でも世界的に有名なワンピースを選択する．

3 吹き出しの分類

音声に内在する感情に合った吹き出しを自動付与するために，音声の情報から吹き出しを分類する必要がある．これは機械学習の分野でよく見られるクラス分類問題である．吹き出しには様々な種類が存在するが，本稿では一般的に多く使用されている丸い吹き出しとギザギザの形状の吹き出しの 2 つのクラスに絞る．なお，本稿で用いるデータセットにおいては，丸い吹き出しが 72.98%、ギザギザの吹き出しが 27.02%であり、これら 2 つで全体の 100%を占める．音声からこの 2 つの吹き出しを分類するための特徴量として以下で述べる音響的特徴量と言語的特徴量を採用した．

3.1 音響的特徴量

音声での感情認識では，従来からピッチやエネルギーといった特徴量が使用されている [1]．近年では，それら以外にも音声分析で得られる様々な特徴量が用いられている．本研究では Schuller らにより提案され，INTERSPEECH 2009 Emotion Challenge[2](以下，IS09) と INTERSPEECH 2010 Paralinguistic Challenge[3](以下，IS10) で使用された音響的特徴量セットを使用する．各セットの詳細を表 1 および表 2 に示す．

Table 1 Emotion Challenge 2009 の特徴量

LLD(16・2)	Functionals(12)
ZCR	mean
RMS Energy	standard deviation
F0	kurtosis，skewness
HNR	extremes: value，rel. position，range
MFCC 1-12	linear regression: offset，slope，MSE

Table 2 Paralinguistic Challenge 2010 の特徴量

LLD(38・2)	Functionals(20)
PCM loudness	Position max/min
MFCC 0-14	arith mean std deviation
log Mel Freq Band 0-7	skewness，kurtosis
LSP Frequency	lin regression coeff
F0	lin regression error
F0 Envelope	quartile
Voicing Prob	quartile range
Jitter local	percentile
Jitter consec. frame pairs	percentile range
Shimmer local	up-level time

3.2 言語的特徴量

言語的特徴量として，形態素単位の unigram，bigram，trigram の頻度を使用する．本稿では，各発話を人手により書き起こしたテキストに対して，形態素解析を行うことで，形態素単位を抽出する．

4 実験的評価

前述した特徴量を用いて，音声から吹き出しの分類を行う．

4.1 実験条件

アニメのワンピースの中から，極力背景にノイズを含まないシーンの音声データを収集し，実験に使用した．収集された音声データは 2025 発話であり，アニメと対応したアニメと対応したマンガから吹き出しの正解ラベルを抽出した．音声から丸とギザギザ

*Evaluation of the generation of text balloon using the acoustic features. by, MATSUMIYA, Sho, SAKTI, Sakriani, NEUBIG, Graham, TODA, Tomoki, NAKAMURA, Satoshi (NAIST)

Table 3　吹き出し付与と字幕付与の有用性の比較

音声有り /音声無し		漫画の吹き出しの形状	
		丸	ギザギザ
実験動画表現	丸	83%/85%	20%/24%
	ギザギザ	14%/21%	84%/78%
	テキスト	35%/37%	42%/40%

の 2 つの吹き出しを 2 値分類するための機械学習手法として，LIBSVM[6] により実装されたサポートベクターマシン (SVM) を使用する．特徴量を抽出するために openSMILE[4] を用いた．各発話は MeCab[5] によって形態素解析された．

4.2　吹き出し付与の効果

吹き出しを字幕に付与する効果を調査する実験を行った．アニメーション動画に字幕のみのものと字幕と吹き出しを付与したものを実験動画として被験者 7 名に提示した．被験者は「場面に合った感情を感じるか」という問いに対して「はい」,「いいえ」で回答した．また実験動画は，音声ありと音声無しの 2 つを用いた．

結果を表 3 に示す．

実験動画の「表現」は，実験動画に付与されているものが丸い吹き出し，ギザギザの吹き出し，テキストのみのいずれかを表している．また「漫画の吹き出しの形状」は，実験動画に使用したシーンに対応している漫画の吹き出しの形状（正解ラベル）を示している．表中の数字は，回答が「はい」であった割合を表している．正解ラベルが付与されている実験動画は，高確率で場面に合った感情を感じるという結果となり，不正解ラベルが付与されている実験動画に対しては，逆の結果となった．また，テキストのみの場合では，場面に合った感情を感じる割合が低い結果となったが，誤った吹き出しを付与するよりは高くなる結果となった．この実験結果から，テキストのみよりも吹き出しを付与することでより適切な感情表現が可能であり，正解ラベル通りの吹き出しを付与することが有用であることが分かった．

また，音声有りと音声無しの結果を比較すると，音声なしの場合を見ても，正解ラベルの吹き出しが付与されている動画に対して感情を感じる割合は，80%ほどの精度を保つことが分かった．このことから，音声が音声が聞き取れない場合においても自動吹き出し生成の有用性が確認できた．

4.3　吹き出し分類の評価

音声からの自動吹き出し分類実験を行った．言語的特徴量のみ，音響的特徴量のみ，さらにそれら 2 つの特徴量を組み合わせたものを使用した．4 交差検定により最終的な分類率を求めた．

分類結果を図 1 に示す．横軸に使用した特徴量を示す．縦軸に分類精度を示す．言語的特徴量の結果で最も精度の良かったものは bigram の 82.34%であり，音響的特徴量の結果では IS10 で 85.22%であった．また両特徴量を組合せることによって精度の改善がみ

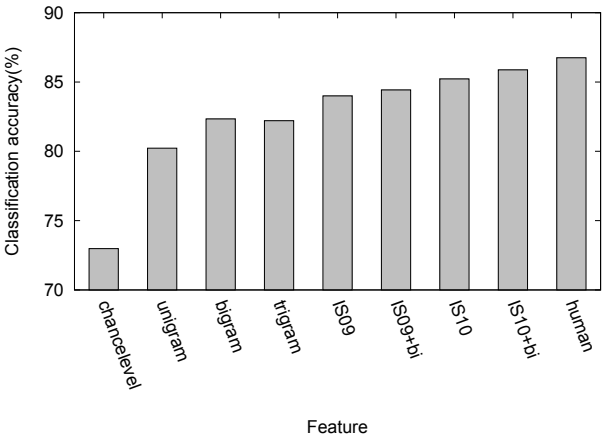

Fig. 1　吹き出しの分類率

られた．全体的に，言語的特徴量よりも音響的特徴量の方が高い精度が得られる傾向が見られた．言語的特徴量のみでは精度が低くなった原因として，セリフデータが少なかったために十分に学習が行えなかったことが考えられる．また，人間による分類実験 (図 1 の human で示す) では，88%の精度が得られた．

5　おわりに

本稿では，音声からの吹き出し分類を試みた．高次元の音響的特徴量を素性として導入することによって，高い精度で分類をすることに成功した．また，言語的特徴量と音響的特徴量を組み合わせることによって精度の改善が確認された．今後は，人間による分類実験の精度の数値を目標に手法の改良を目指す．

謝辞 本研究の一部は JSPS 科研費 24240032 の助成を受け実施したものである。

参考文献

[1] 白澤敏行，山村毅，田中敏光，大西昇，”音声に込められた感情の判別”，**信学技法**，HIP96-38，pp. 79–84，1997.

[2] Bjorn Schuller, Stefan Steidl, and Anton Batliner, ”The INTERSPEECH 2009 Emotion Challenge,” *Proc. Interspeech,* 2009.

[3] Bjorn Schuller, Stefan Steidl, Felix Burkhardt, Laurence Devillers, Christian Muller, and Shrikanth Narayanan, ”The INTERSPEECH 2010 Paralinguistic Challenge,” *Proc. Interspeech,* 2010.

[4] Florian Eyben, Martin Wollmer, and Bjorn Schuller, ”openSMILE The Munich Versatile and Fast Open-Source”

[5] 工藤拓，山本薫，松本裕治．”Conditional random fields を用いた日本語形態素解析,” **情報処理学会自然言語処理研究会**, SIGNL-161, Vol.47, pp. 89–96, 2004.

[6] Rong En Fan, Pai Hsuen Chen, and Chin Jen Lin. ”Working set selection using second order information for training SVM,” *Journal of Machine Learning Research,* Vol. 6, pp. 1889–1918, 2005.