マルチモーダル情報を用いた運転中におけるシステム向け発話の推定

情報科学科　澤田 優希　　指導教員：入部 百合絵

1　はじめに

近年急速に音声対話システムが普及しているが, 自動車の運転に取り入れるにあたってはいくつかの課題がある. その一つとして, 車内に同乗者がいる場合, 対話システムに向けての発話か同乗者への対話であるのか判別する必要がある. 人間とロボットとの対話における受話者推定や応答義務推定の研究では，音声の韻律的情報や発話スタイルの他に顔の向きも有効であると報告されている [1][2].

本研究では人間への対話とシステムへの対話の特徴の差異を検出するために, 音声の韻律情報の他に顔の向き情報や視線情報を含むマルチモーダル情報を運転中のドライバから取得し, ドライバの発話行為の特性について明らかにする. また, それらの特徴量を用いて識別器によりシステム向け発話の推定を行うことで, 抽出した特徴量が有用なものであるのかを調査し, 運転環境下のシステム向け発話の推定精度を向上させる.

2　運転中におけるドライバからの特徴量抽出

運転中のシステム向け発話の推定に用いるため, ドライバからマルチモーダル情報を収集した. 運転中に対話システムに話しかけることを想定し, 被験者は運転席に着席し運転しながら対話システムと会話をした. また, 人間向け発話とシステム向け発話の差異を検出するため, 助手席に協力者に座ってもらい話の受け手となってもらった. 安全性の問題より運転は実車ではなくドライビングシミュレーターを代用し, 運転コースは高速道路とした. 本実験における「対話システム」では, 予め音声合成器 OpenJTalk で生成した音声をスピーカから流し,Wizard-of-Oz 法 (WOZ 法) にて被験者はシステムと対話してもらうこととした. スピーカは実車にてカーナビゲーションシステムが設置されている付近の位置に取り付けた (図 1[a]).

被験者に 2 度発話をしてもらうタスク (タスク 1), 被験者に 1 度発話をしてもらうタスク (タスク 2) に加え, ドライバからの問い掛けに対し意図的にシステムが応答しないというタスク (タスク 3) を設定した. タスク 3 の目的は, システムからの反応が無い場合にドライバがどのような特性を示すかを明らかにすることである. そのため, タスク 3 は対話システム向けの発話を判定する上で重要な判断材料となると考えられる. 抽出した特徴量を比較することでドライバの運転時の特性分析を行った.

解析を行う特徴量は先行研究で有用であるとされている基本周波数, ラウドネス等の韻律情報に加え, 顔の向きと視線の向きのマルチモーダル情報である.

3　特徴量の解析およびシステム向け発話の推定

被験者 10 名 (男 6 名, 女 4 名) 分のデータより, 運転中における発話中の各特徴量について人間相手とシステム相手での差異

[a] 実験で用いたスピーカの位置

[b] 発話中の視線分布

図 1　データ収集実験

表 1　システム/人間向け発話における抽出特徴量の有意差の有無

	発話中		発話前後 20 フレーム		
	平均	最大	平均	最大	最小
基本周波数	○		-	-	-
ラウドネス	○	○	-	-	-
視線 X 座標				○	
視線 Y 座標	○		○		
頭部ピッチ					
頭部ヨー	○		○	○	○

(○:有意差有, 空欄:有意差無, -:抽出不可)

表 2　識別器によるシステム向け発話推定結果 (%)

	Precision	Recall	F-Measure
韻律のみ	65.2	39.5	49.2
視線, 顔向きのみ	72.1	61.3	66.3
韻律 + 視線顔向き	76.0	67.9	71.7

を調べた. これらの特徴量を全発話区間, 発話開始または終了前後 20 フレーム区間の 2 つの区間においてそれぞれ t 検定を行い, 有意差が認められた特徴量を表 1 に示す. ドライバは運転中でも発話行為の際に発話相手によらず左方向に視線と顔を向ける傾向があったが, 特に人間相手の方がシステム相手よりも視線が動く傾向が明らかとなった (図 1[b]).

次に, 基本周波数とラウドネスの韻律情報の特徴量 (2 次元), 視線 (X 座標,Y 座標) と顔向き (ピッチ, ヨー) の特徴量 (4 次元), 全ての特徴量 (6 次元) をそれぞれ Support Vector Machine(SVM) を使用した識別器にかけ, システム向け発話と人間向け発話の推定を行った. システム向け発話の推定結果を表 2 に示す. 正解率も韻律のみが 60.9%, 視線顔向きのみが 70.2%, 全ての特徴量が 74.4% となり, 適合率, 再現率，F 値, 正解率全ての項目において韻律と視線顔向けの特徴量を使った推定が最も精度が高い結果となり, マルチモーダル情報は運転中のシステム向け発話の推定に有用であるといえる.

4　おわりに

本研究では, ドライバから取得したデータから有用な特徴量を抽出し, それらの特徴量を用いて識別器によりシステム向け発話の推定を行った. その結果, 運転中におけるシステム向け発話と助手席の人間向け発話では, 先行研究で有用とされていた基本周波数, 発話パワーの韻律情報に加え, 視線の座標や頭部回転のヨー方向に差異が認められた. また識別器による推定では, 韻律情報のみ, 視線と顔向き情報のみに比べ, 全てを含めたマルチモーダル情報を使った場合が最も精度の良い結果となった.

今回の推定では 70 %以上の正解率を実現したが, 推定精度を更に高めるために特徴量を増やして検証することが今後の課題として挙げられる.

参考文献

[1] 馬場 直哉, 黄 宏軒, 中野 有紀子: 人対会話エージェントとの多人数会話における頭部方向と音声情報を用いた受話者推定機構, 人工知能学会論文誌 28(2), 149-159, 2013

[2] 杉山 貴昭, 船越 孝太郎, 中野 幹生, 駒谷 和範: 多人数対話におけるユーザの状態に着目したロボットの応答義務の推定, 人工知能学会論文誌 31(3), C-FB2_1-9, 2016