SVMを用いたシステムへの問い合わせと雑談の判別*

◎山形知行，佐古淳，滝口哲也，有木康雄（神戸大）

1　はじめに

近年，様々な分野で音声によるインターフェースが実用化されつつある．特に，ロボットとのコミュニケーションや，カーナビのように手を使うことが困難な機器の操作への適用が顕著である．しかし，現在使用されている音声認識システムは入力された音声がシステムへの発話か周囲の雑談かを判別できないため，スイッチ等を用いなければ意図しない動作を湧き出させてしまう．これに対し，従来の研究では音声スポッタ[1]のようにユーザが意識して韻律特徴や言語特徴を変化させる方法があるが，これではユーザは自分の発話に不自然さを感じる．よりユーザに負担をかけない手法として，自然な発声の音響特徴を用いる方法[2] [3]が検討されている．また，音声認識結果を基にする方法[4]も提案されている．本研究では，発話毎の音響特徴に加え，各発話の前後部を切り分けて音響特徴量を求めることで識別精度の向上を行う．また，音響特徴と言語特徴を組み合わせることにより，より頑健なシステムを目指す．

2　本研究で利用したコーパス

まず, 二人以上の人間とシステムが同時に存在することを想定する．これは，ロボットを操作する際に周囲に人がいる場合や，カーナビを操作する際に助手席に同乗者がいる場合のように，自然な状況であると考えられる．本研究ではシステムとして音声コマンドにより移動するロボットを用いた．二人以上の人間が互いに会話を行いながら，任意にロボットへ「写真を撮って」,「こっちに来て」等のシステム要求発話を行う．また，現状のロボットで受理できないが，ユーザがロボットの動作を期待して発話した「付いてきてー」のような発話もシステム要求発話とした．収録は，二人の発話者それぞれの胸元に取り付けたマイクで行った. 発話数は 330 で, 内 49 発話がシステム要求発話であった.

3　従来手法

3.1 音響特徴量

発話区間のパワーとピッチの平均・標準偏差・最大・最大-最小値差の 8 次元を音響特徴量として用いる[2] [3].

3.2 言語特徴量

全発話の音声認識結果に含まれる単語の異なり数を次元としたベクトル空間を用意し，一発話内の単語の出現回数をベクトルの要素として用いる[4].

4　提案手法

対話中の発話を特徴づける要素は前節で述べた音響特徴量･言語特徴量だけに限らない. フィラーや会話の間と言った情報も含まれ，それらが組み合わされていると考えられる. そのため本研究では，まず音響特徴量を 4.1 節のように拡張し，次に音響的特徴と言語的特徴の組み合わせを行う．これらの特徴量を用いて SVM によりシステム要求判別を行う.

4.1 発話区間前後を考慮した音響特徴量

特に雑談では発話の前後部にフィラーや笑い声, 言い淀み等が入ることが多いのに対し，コマンド発話の前後は無音になることが多いと考えられる．このため，音響特徴量を発話前部・中部・後部の 3 区間に分け同様にパワー・ピッチを求め，24 次元の音響特徴量として識別に用いる.

4.2 音響特徴量と言語特徴量の組み合わせ

3.2 節及び 4.1 節で得られた特徴量のベクトルを連結することにより図 1 のように特徴量を求める(ただし α，β はスケーリング係数).

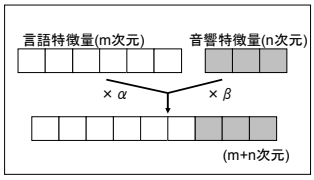

図 1 音響特徴量と言語特徴量の統合

* System Request Discrimination Based on SVM, by YAMAGATA Tomoyuki, SAKO Atsushi, TAKIGUCHI Tetsuya and ARIKI Yasuo (Kobe University)

5　実験

本稿ではまず音響特徴量のみを用い，発話区間を分けない場合と分けた場合で比較する．次に音声認識結果の言語特徴量を用いた場合，言語と音響特徴量を統合した場合について述べる．ただし，発話区間については音声認識で取り扱いやすいようにパワーを閾値に前後にマージンを取り，切り出しを行った．また，SVM の Kernel 関数には RBF Kernel を用い，評価は Leave-one-out により行った．

5.1 音響によるシステム要求判別

3.1 節及び 4.1 節で述べた特徴量をもとに識別を行う．ただし，パワーは RMS(Root Mean Square)，ピッチは F0 を用いた．発話前後部の特徴量は，それぞれ 0.7 秒間のデータを元に算出した．また，特徴量はそれぞれ決定木学習ツール C4.5 で生成された木の順序を元にスケーリング[5]を行った．実験の結果，F 値が最大となったケースを表 2 に示す．発話区間を 3 つに分割することにより精度が向上していることが分かる．

5.2 言語によるシステム要求判別

まず，音声認識の条件を述べる．ベースラインの音響モデルは CSJ モニター版のうち，男性話者 200 名の講演音声を用いて作成した．音響分析条件と HMM の仕様を表 1 に示す．さらに，MLLR+MAP により音響モデル適応を行った．適応データの分量は約 10 分である．音響モデル適応はテストセットを含めたクローズ，言語モデル適応は話者 B の発話を用いて話者 A の認識用言語モデルを作成することによりオープンとした．この条件の下，Julius により音声認識を行った結果，単語正解精度 42.1%となった．この音声認識結果を基に 3.2 節の言語特徴量を生成した結果，566 次元のベクトルとなった．これを用いてシステム要求判別を行った結果が表 2 である．音響特徴量を使った場合よりも高精度でシステム要求判別を行うことができることが分かる．

5.3 音響及び言語特徴量の組み合わせによるシステム要求判別

音響特徴量と言語特徴量の計 590 次元で判別を行う．ただし，言語特徴量と音響特徴量のスケーリングは最も識別率が高くなるものを実験的に求めた．実験結果より言語特徴量のみで識別した場合より高い精度が得られているのが分かる．

表 1 音響分析条件と HMM の仕様

音響分析	サンプリング周波数	16kHz
	特徴パラメータ	MFCC(25 次元)
	フレーム長	20ms
	フレーム周期	10ms
HMM	窓タイプ	ハミング窓
	タイプ	244 音節
	混合数	32 混合
	母音(V)	5 状態 3 ループ
	子音+母音(CV)	7 状態 5 ループ

表 2 システム要求発話判別結果

	適合率	再現率	F 値
音響	0.71	0.61	0.66
音響(3 分割)	0.80	0.92	0.86
言語	0.94	0.94	0.94
言語+音響	0.94	0.96	0.95

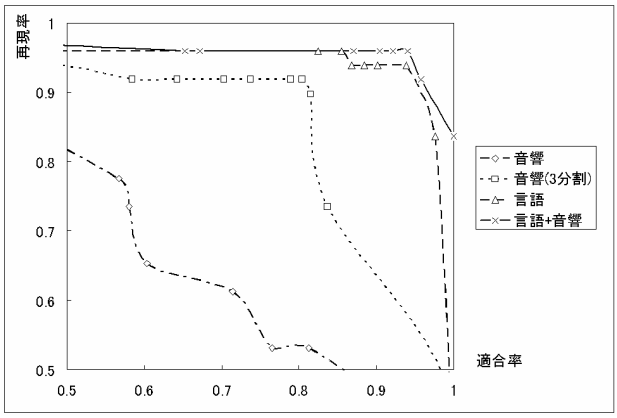

図 2 システム要求発話判別結果

6　まとめ

本稿では発話区間の前後部を切り分けシステム要求判別の音響特徴量を求める手法，及び音響特徴量と言語特徴量を組み合わせる手法について述べた．発話区間全体からではなく，3 区間に分けて特徴量を求めることで判別精度が向上することが分かった．また，音響特徴量と言語特徴量との組み合わせにより高い識別精度が得られた．

今後の課題としては，大規模なコーパスでの実験，システム要求発話を特徴づける発話区間前後部の長さの自動推定があげられる．

参考文献

[1] Goto *et al*, ICSLP, 8, 1533-1536, 2004.
[2] 山田, SIG,-SLP-61(2), 7-12,2006-5.
[3] 杉本, 信学会, 総合大会, D-14-9, 133, 2006.
[4] 佐古, SIG-SLP-64, 19-24, 2006-12.
[5] C.-W. Hsu *et al.*, http://www.csie.ntu.edu.tw /~cjlin/papers/guide/guide.pdf