話者交替を考慮し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と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．これは特に Fig. 1 のようにシステムと複数の人が同時に存在するような環境で問題となる．これに対し，従来の研究ではユーザが意識して韻律特徴や言語特徴を変化させ入力する音声スポッタ [1] があるが，ユーザは自分の発話に不自然さを感じるという問題がある．人の発話は自然に話している場合でも，話し相手の反応によって音響・言語的特徴に差が生じる [2]．これは現在のカーナビのような機械的なインタフェースと人との会話の場合にはより顕著に表れる．我々は音声認識結果の言語的特徴を用いる手法 [3] や音響・言語的特徴を組み合わせる手法 [4] を提案してきた．これに対し本稿では音響的特徴と，発話前後での話者の交替 [5] に注目する．システムと複数の話者が同時に存在するような環境では，発話前後での話者の交替を考慮することで，より正確にシステム要求を検出することが可能となった．

2　本研究で用いたコーパス

まず，人間 2 人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．2 人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．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．Julius[6] の Adintool により発話区間を切り出した結果，全発話数は 1025 発話，システム要求発話が 108 発話であった．

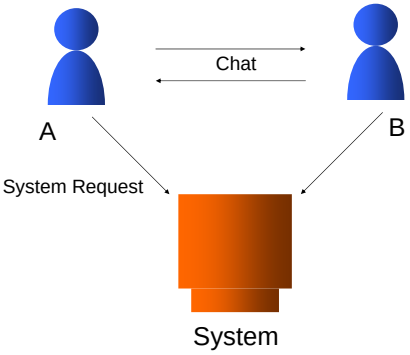

Fig. 1　One System + Two individuals dialog

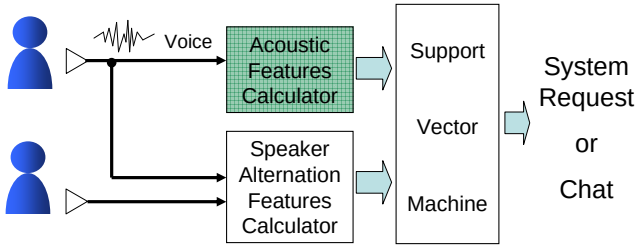

Fig. 2　System Overview

3　提案手法

本研究では Fig. 2 のように，それぞれの話者に取り付けた接話マイクを用いて，音響特徴量と話者交替特徴量を求める．その後，それぞれの特徴量を初期統合したうえで，サポートベクターマシンを用いて一発話毎にそれがシステム要求発話であるか雑談であるかを判別する．

3.1　音響特徴量

従来の音響特徴量は，一発話毎にパワーやピッチを求めていたが，システム要求発話と雑談の音響的な差は Fig. 3 のように発話の言い始めや言い終わりに現れることが多い．システム要求発話では発話の前後が無音になることが多いのに対し，雑談では言い始めや言い終わりがはっきりせず，切り出された発話区間の前後部分にも言い淀み等が残る．このため，本研究では切り出された発話区間からだけではなく，その前後にとったマージンからもそれぞれパワー・ピッチの平均・標準偏差・最大・最大-最小値差を求め，これら 8 次元 ×3 区間の 24 次元を音響特徴量として用いた．なお，マージンの長さは予備実験から判別精度の最も良かった 0.7 秒を用いている．

*Detection of System Request in Conversational Speech Considering Speaker Alternation. by Tomoyuki, YAMAGATA, Atsushi, SAKO, Tetsuya, TAKIGUCHI, Yasuo, ARIKI (Kobe University)

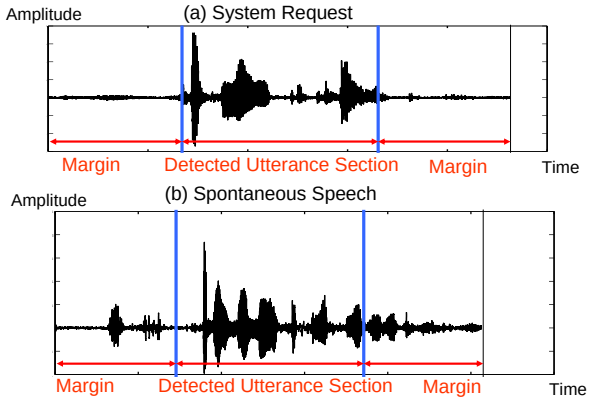

Fig. 3　The wave-form of a system request and a spontaneous speech.

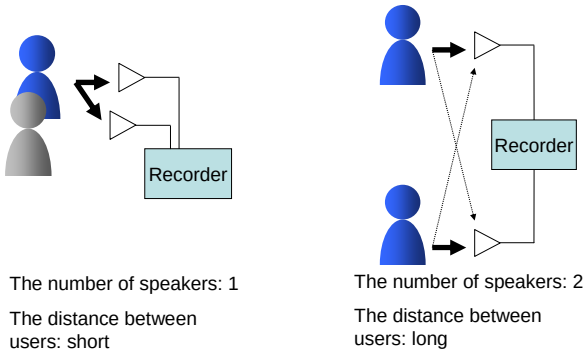

Fig. 4　An example in which the power based system cannot detect speakers correctly.

3.2　話者交替特徴量

発話区間前後のマージンには言い淀み等の他に，隣の話者の発話が入る場合がある．このため，マージンの部分に残っている波形が言い淀みによるものなのか，隣の話者の声が入った物なのか区別ができない．これは接話マイクを用いた場合でも，2 者の距離が十分に離れていなければ問題になる．このため本研究では，それぞれの区間でどちらの話者が喋っているのか示す特徴量を用意する．入力音声のパワーを見るだけでは Fig. 4 のように話者は 1 人だがマイク間距離が近い場合と，マイク間距離は遠いが話者が 2 人の場合の区別ができない (どちらも両方のマイクのパワーが大きくなり，どちらのユーザーが喋っているか区別ができなくなる)．このため，本研究では CSP 係数 (1) を用い，どちらのマイクに先に音声が到達しているかを求める．

$$CSP[k] = IDFT(\frac{X_1[n]X_2^*[n]}{|X_1[n]||X_2[n]|}) \quad (1)$$

$$\tau_i = \max_{k\ in\ \Sigma_i}(CSP[k]) \quad (2)$$

窓関数の大きさを N とし，Σ_A : $0 < k < \frac{N-1}{2}$ 及び Σ_B : $\frac{N}{2} < k < N$ でそれぞれピークの値 τ_A, τ_B を求める．これにより，例えば τ_A のみ大きければ話者 A が喋っている (話者 A に取り付けたマイクに先に音が到着している)，τ_A, τ_B 共に大きければ両者が喋っているという事が分かる．これら τ_A, τ_B を 3.1 で求めた 3 区間それぞれで求め，計 6 次元を話者交替の特徴量とする．

4　実験

実験では切り出された発話区間のみから特徴量を求めた場合，前後のマージンを含め 3 区間で特徴量を求めた場合，話者交替を考慮した場合の 3 種類で評価を行った．SVM の Kernel 関数には RBF (Gaussian) Kernel を用い，10-folds のオープンによる評価を行った．初期統合の重み付けは実験的に行い，最も結果が良くなった場合を Table 1 に示す．従来発話区間のみから特徴量を求めていた場合に比べ，前後のマージンを考慮することで判別精度が上がっている．また，話者交替の特徴量を含めることで，さらに性能が上がっていることが分かる．これは特に，雑談の場合によく起こるラッチング (前の人が完全に喋り終わるまでに次の人が喋り始める現象) 等をより正確に判断できるようになったからであると考えられる．

Table 1　The results of system request detection.

	F-measure
Acoustic 8 dim. (1 section)	0.677
Acoustic 24 dim. (3 sections)	0.817
Aco. 24 dim. + Speaker Alternation	0.851

5　おわりに

本稿では 2 チャンネルのマイクロフォンを用い，発話前後の話者交替を考慮し，システム要求発話と雑談を判別する手法を提案した．実験結果から，検出された発話区間の前後からも音響特徴量を求めると共に，話者の交替を考慮することでより効果的にシステム要求発話と雑談を判別できることが分かった．

今後の課題としては，ノイズ環境下での評価や，カーナビを用いての実験があげられる．

参考文献

[1] Goto et al, ICSLP, 8, 1533-1536, 2004.
[2] 山田 他, SIG,-SLP-61(2), 7-12,2006-5.
[3] 佐古 他, SIG-SLP-64, 19-24, 2006-12.
[4] 山形 他, 音響論 (春), 185-186, 2007.
[5] 大須賀 他, JSAI, 論文誌 Vol.21, No.1, 2006
[6] http://julius.sourceforge.jp/