FBANK と Gabor Wavelet 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，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での音声インタフェースの利用が顕著である．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] や韻律・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や韻律的特徴，またその変化に注目する．音素のような短周期の変化から韻律のような長周期の変化までを捉えるため，対数メルフィルタバンク (以降 FBANK) の時間軸に対して時間方向にスケーリングしながら Gabor Wavelet を畳み込んだ特徴量を用いた．Support Vector Machines によりシステムへの問いかけと雑談の判別を行った結果，F 値 92.6%の精度でシステムへの問いかけと雑談を判別することができた．

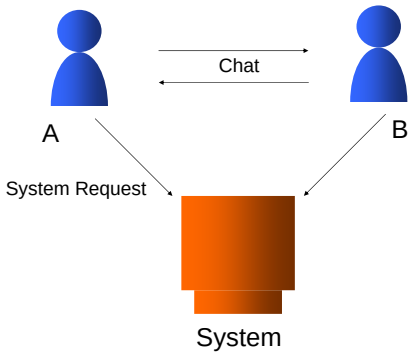

Fig. 1　One System + Two individuals dialog

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 人が自由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 (コマンド) 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[5] の Adintool により発話区間を切り出した結果，全発話数は 1024 発話，システム要求発話が 110 発話であった．

3　従来手法

従来，我々は発話区間やその前後のマージンからパワー・ピッチの統計情報を求め，これらを特徴量としてシステム要求判別を行ってきた [4]．検出された発話区間のみからではなく，その前後に残る言い淀み等からも特徴量を求めることでシステム要求判別精度が向上した．システム要求発話の発話区間前後が無音になることが多いのに対して，雑談のような自由発話では言い淀み等が残る事が多い．前後のマージンからも特徴量を求めることで，発話の立ち上がり・立ち下がりといった韻律的な情報を考慮することができるようになるため，より高精度でシステム要求判別を行うことができたと考えられる．

4　提案手法

本研究では図 2 のように，話者に取り付けた接話マイクからの入力を用いて，音響特徴量や Gabor 特徴量を求める．その後，SVM を用い，一発話毎にそれぞれの特徴量から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．

4.1　音響特徴量

3 章で述べた発話単位での長周期的な韻律の変化の他に，より短周期での変化や音響的な特徴に差が現れるかをまず調査する．特徴量には FBANK を使い，各次元の発話区間でのパワーを発話時間で正規化し

*Detection of System Request in Conversational Speech Using FBANK and Gabor Wavelet. by Tomoyuki, YAMAGATA, Atsushi, SAKO, Tetsuya, TAKIGUCHI, Yasuo, ARIKI (Kobe University)

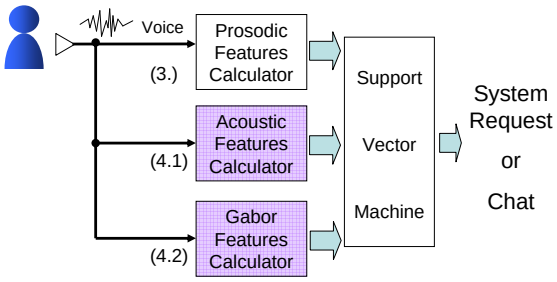

Fig. 2　System Overview

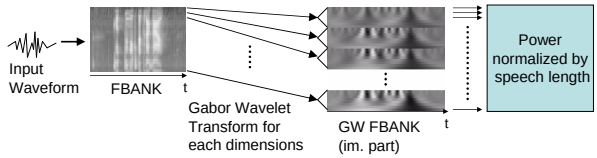

Fig. 3　Flowchart of Calculating Gabor Features

て用いる．実験結果より，FBANK を用いても F 値 80.3%の精度でシステム要求判別が行えている．しかし，FBANK の短時間変化 (ΔFBANK) の発話長正規化パワーを加えることにより，F 値 87.4%とより高精度でシステム要求判別が行えている．

4.2　Gabor 特徴量

本研究では 3 章と 4.1 章の結果を踏まえ，音素のような短周期での音響特徴量の変化から，韻律のような長周期での変化までを統合して取り扱える特徴量を考える．特徴量の抽出手法は Fig.3 に示すように，FBANK をベースに時間軸方向に 1 次元の Gabor Wavelet をスケーリングさせながら畳み込んでいく．

$$\Phi(t) = \frac{1}{\sqrt{2\pi}\sigma} exp(-\frac{t^2}{\sigma^2}) exp(j2\pi ft) \qquad (1)$$

σ は韻律的な変化を見るために比較的長い 512(ms) から始め，音素等のより細かい変化まで見るため $1/\sqrt{2}$ スケーリングで 8 段階の分析を行った．最も小さい σ は 45(ms) となる．3 章と 4.1 章より，韻律の変化量や音素の変化量を用いるのが有用であると考えられるため，Gabor Wavelet のは実部 (Fig.4 赤線)・虚数部 (Fig.4 青線) 両方を用いる．これは Gabor Wavelet の虚数部が奇関数であるため，微分型になっていると考えられるためである．最後に，各周期でのパワーと時間変化量を求めるために，それぞれ実部・虚部の各次元から発話長で正規化したパワーを算出し，これを特徴量としてシステム要求判別を行う．

5　実験

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．従来発話区間のみから特徴量を求めていた場合 (Prosodic8) に比べ，前後のマージンを考慮すること (Prosodic24) で判別精度が上がっている．また，FBANK の発話長正規化パワーを用いた場合 (FBANK_E) に対し，短時間 Δ の発話長正規化パワーを加えた場合 (FBANK_E_D) の判別精度が高くなっている．これに対し，FBANK に Gabor Wavelet を用いた特徴量は全体的に高精度で判別が行えている．Gabor Wavelet の実・虚部両方を用いることにより，F 値 92.6%の精度でシステム要求判別が行えた．

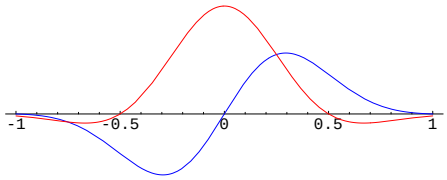

Fig. 4　Gabor Mother Wavelet

Table 1　The accuracy of system request detection.

	Precision	Recall	F-measure
Prosodic8	0.584	0.806	0.667
Prosodic24	0.756	0.889	0.817
FBANK_E	0.719	0.909	0.803
FBANK_E_D	0.866	0.882	0.874
Gabor Re.	0.906	0.873	0.889
Gabor Im.	0.933	0.891	0.912
Gabor Re.+Im.	0.943	0.909	0.926

6　おわりに

本稿では FBANK の時間軸方向に対して様々な大きさの Gabor Wavelet をかけ合わせることにより，韻律のような長周期から音素のような短周期の変化までを取り出せる特徴量を提案した．この特徴量を用いて SVM によりシステムへの問いかけ発話と雑談発話の判別を行った結果，高い精度で判別することができた．今後の課題としては，ノイズ環境下での評価や，カーナビ等の別のタスク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] http://julius.sourceforge.jp/