社団法人 情報処理学会 研究報告
IPSJ SIG Technical Reports

2007－ＳＬＰ－69 (50)
2007／12／21

# 韻律及び話者交代情報を用いたシステム要求検出

山形 知行† 佐古 淳†† 滝口 哲也† 有木 康雄†

† 神戸大学大学院工学研究科 〒657-8501 兵庫県神戸市灘区六甲台町 1-1
†† 神戸大学大学院自然科学研究科 〒657-8501 兵庫県神戸市灘区六甲台町 1-1
E-mail: †{yamagata,sakoats}@me.cs.scitec.kobe-u.ac.jp, ††{takigu,ariki}@kobe-u.ac.jp

**あらまし** 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への音声インタフェースの適用が顕著である．しかし，現在主に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誤動作を湧き出させてしまう．一方，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韻律的特徴に差が生じる．本研究ではこの韻律的特徴の差と，発話前後での話者の交替からシステムへの問いかけと雑談を判別する．検出された発話区間の前後を考慮して韻律的特徴を求め，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 81.7%の精度で判別することが可能となった．また，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代を考慮することで，F 値で 85.1%まで判別精度が向上した．同時に，対数メルフィルタバンクと Gabor Wavelet 変換を用いた話し方の明瞭度を表す特徴量を検討し，韻律の変化や音素の変化を捉えることにより，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．

**キーワード** システム要求判別, Support Vector Machine, 韻律, 話者交代, Gabor Wavelet

# System Request Detection in Conversation Based on Prosodic and Speaker Alternation Features

Tomoyuki YAMAGATA†, Atsushi SAKO††, Tetsuya TAKIGUCHI†, and Yasuo ARIKI†

† Graduate School of Engineering, Kobe University
Rokkodaicho 1-1, Nada-ku, Kobe, Hyogo, 657-8501 Japan
†† Guraduate School of Science and Technology, Kobe University
Rokkodaicho 1-1, Nada-ku, Kobe, Hyogo, 657-8501 Japan
E-mail: †{yamagata,sakoats}@me.cs.scitec.kobe-u.ac.jp, ††{takigu,ariki}@kobe-u.ac.jp

**Abstract** For a hands-free speech interface sucha as robots and car navigation systems, it is important to detect commands in spontaneous utterances. The difference between spontaneous utterances and commands mostly appears in prosody. To discriminate commands from human-human conversations by prosodic features, it is efficient to consider the head and the tail of an utterance. Experiments show that using prosodic features result in 81.7% in F-measure thanks to Support Vector Machines. And also, considering the alternation of speakers using two channel microphones improved the performance to 85.1%. At the same time, we investigate utterance clearness features which obtained by using FBANK and Gabor Wavelet. Experiments also show that clearness features mark 92.6% in F-measures.

**Key words** System Request Discrimination, Support Vector Machine, Prosody, Speaker Alternation, Gabor Wavelet

## 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，現在主に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．これは特に図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．一方，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韻律や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韻律的特徴と，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．また，システムに対してコマンドを発する場合と人同士で雑談をする場合では，話し方の明瞭度が変わってくる．本研究では，話し方がはっきりしているかどうかを表すための特徴量を検討した．対数メルフィルタバンク (以降 FBANK と表記) の各次元に対し，時間方向に Gabor Wavelet [6] をかけその変化量を見ることにより，コマンドのような明確な発話を正確に検出できることが分かった．

## 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 (図2) を用いた．収録に際しては，ロボットを Wizerd of OZ [7] ではなく，実際に稼働させ録音している．ロボットには2つの全方位カメラ (障害物検知用と顔検出用) が装備されており，障害物を避けながら動いたり，人の顔写真を撮ることができる．システムと人との対話では視線方向の情報を用いて発話が誰に向けて行われたかを推定する研究 [8] も行われているが，カーナビ等では安全性の問題により視線を移すことは難しく，また，ロボットの場合も高解像度のカメラが必要となるため，本研究では視線情報は用いない．この他，ロボットには2チャンネルのマイクロフォン (収録に用いるマイクロフォンとは異なる) が装備されており，人がどちらから喋りかけたかを推定することができる．これらのセンサーにより，ロボットは人の方向や障害物を推定しながら部屋の中を移動することができる．ロボットの機能およびロボットの受理できるコマンドを表1にまとめる．

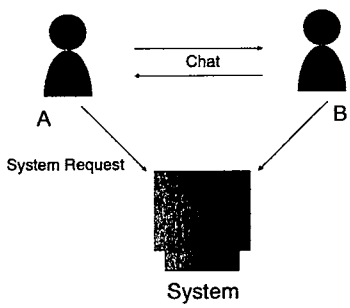


図1 システムと複数の人による対話
Fig. 1 One System + Multi individuals dialog

表1 本研究で用いたロボットの機能と受理できるコマンド例
Table 1 Functions of the robot and examples of commands.

| | |
|---|---|
| 機能 | CSP 法による音源到来方向推定<br>音源方向/反対方向への移動<br>障害物の回避<br>アームによるボトルの設置<br>顔写真の撮影 |
| コマンド例 | こっちに来て<br>向こうへ行って<br>写真を撮って<br>私について来て<br>場所 A にボトルを置いて |

表2 実験データの発話数
Table 2 The number of utterances in the corpus.

| 全発話数 | システム要求発話数 |
|---|---|
| 1024 | 110 |

収録は雑音の比較的少ない実験室内で行った．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する．システム要求以外の発話としては，通常の雑談に加え,「こっちに来て，とか言うと……」「こっちに来て，向こうへ行ってだけでは……」のようにコマンドを一部に含むような発話も含まれている．収録用のマイクロフォンは2人の発話者それぞれの胸元に取り付けた．

一回の収録は約30分である．話者4人が入れ替わり，3セット，計90分のデータを用意した．Julius [9] の Adintool により発話区間を切り出した結果，検出された発話数は表2のようになった．

図 2　本研究で用いたロボット
Fig. 2　An image of the robot.

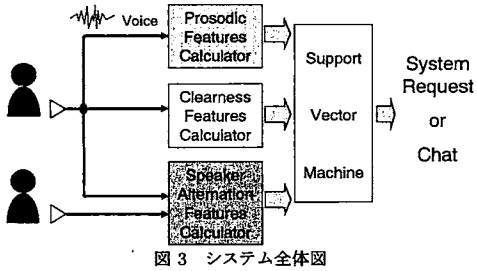


図 3　システム全体図
Fig. 3　System Overview

## 3. 提案手法

本研究では図 3 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，韻律特徴量，話者交替特徴量及び発話明瞭度特徴量を求める．その後，Support Vector Machines (以降 SVM と表記) [10] 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．

### 3.1　韻律特徴量

従来，システムへの要求と雑談を判別するためには一発話毎にパワーやピッチ等の韻律的特徴量を求めていたが，図 4 に示すようにシステム要求発話と雑談の韻律的な差は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表 3 の特徴量を求め，これら 8 次元 ×3 区間の 24 次元を韻律特徴量として用いることとした．

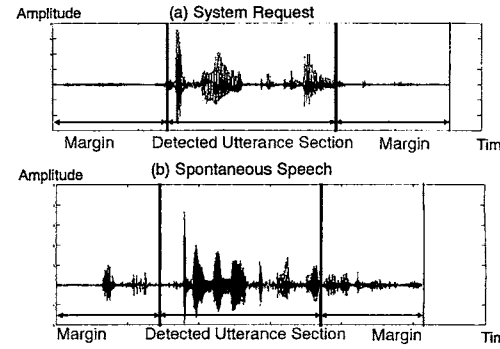


図 4　システム要求発話・雑談の波形
Fig. 4　The wave-form of a system request and a spontaneous speech.

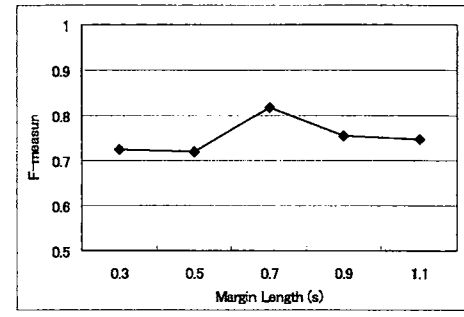


図 5　マージンの長さとシステム要求判別精度
Fig. 5　The relation of the margin length and the accuracy of System Request Detection.

表 3　韻律特徴量
Table 3　Prosodic Features.

| | | | | |
|---|---|---|---|---|
| パワー | 平均 | 標準偏差 | 最大 | 最大 - 最小 |
| ピッチ | 平均 | 標準偏差 | 最大 | 最大 - 最小 |

マージンの長さは短すぎると発話前後の特徴を正確に捉えられず，長すぎると前や次の発話を含んでしまうと考えられる．このため，予備実験により最適なマージンの長さを求めた．マージンの長さを変えながら韻律特徴量を求め，システム要求判別を行った結果を図 5 に示す．予備実験により 0.7 秒のマージンを用いた場合が一番判別精度が高かった．

### 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．次に喋る話者が，今の話者が喋り終わるまでに重ねるように話し始める現象はラッチングと呼ばれ，人間同士の自然な会話では日常的に起こる．しかし，ラッチングは韻律特徴量を求める上で，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という問題を引き起こす．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．一方，逆に考えれば，ラッチングが起こる場合は自然な会話，つまり雑談であることが多いと考えられる．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，図6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[11]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係数は$x_1, x_2$をマイクロフォン入力とし，その短時間フーリエ変換結果$X_1, X_2$に対し，(1)式で表される．CSP係数のピーク値は図7のように現れるため，ピーク値の現れ方を調べることにより，どちらの話者が喋っているのか(もしくは両方の話者が喋っているのか)が分かる．

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

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

窓関数の大きさを$N$とし，$\Sigma_A: 0 < k < \frac{N-1}{2}$及び$\Sigma_B: \frac{N}{2} < k < N$でそれぞれピークの値$\tau_A, \tau_B$を求める．(ただし短時間分析でのCSP係数はノイズが大きいので，250msで平滑化を行った．)例えば，$\Sigma_A$区間に大きなピークが存在するということは，話者A側のマイクに先に音が到達していることを示している．逆に，$\Sigma_B$区間にピークが存在するということは，話者B側のマイクに先に音が到達していることを示す．また，予備実験により，二者が同時に喋ったときにも，値は小さくなるがピークが出現することが分かっている．これにより，例えば$\tau_A$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)，$\tau_A, \tau_B$共に大きければ両者が喋っているという事が分かる．このどちらの話者が喋っているかを表す$\tau_A, \tau_B$を3.1章で求めた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．

### 3.3 明瞭度特徴量

カーナビや今回用いたロボットのように，コマンドを受理するタイプのシステムを対象とする場合，システムに対する発話ははっきりとした喋り方になる場合が多い．発話の明瞭度は3.1章で述べた韻律特徴量にも現れるが，本章では発話の明瞭度をより詳細に捉えるための特徴量を提案する．

まず，本研究では発話の明瞭度を韻律の明瞭度のような長周期の変化と音素の明瞭度のような短周期の変化と

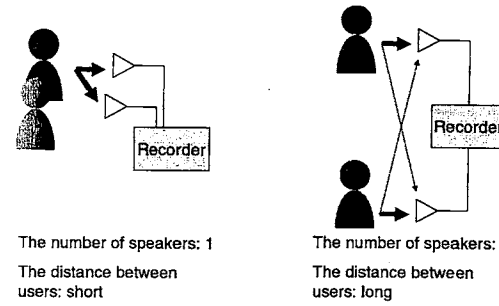


図6 パワーを元にした特徴量で判別できない例
Fig. 6 An example in which the power based system cannot detect speakers correctly.

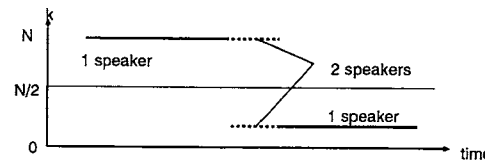


図7 CSP係数によるピークの現れ方
Fig. 7 CSP Peak Trace

考えた．つまり，3.1章で述べたような発話ごとの韻律的な特徴と，一つ一つの音素がはっきりと発話されていることを表す特徴であり，我々はこれら両方を取り扱えるような特徴量を考える．このため，本研究ではまずベースの特徴量として音素の特徴を表せるようにFBANKを用いることとした．システム要求発話と雑談のFBANK(パワーを含めた25次元)の具体例を図9に示す．FBANKを見るとシステム要求発話は全体的にパワーが強く，また時間方向にメリハリがあることが分かる．しかし，これら一発話ごとのFBANKを直接システム要求判別に用いることは難しい．コマンド例をあらかじめ記憶しておきパターンマッチングを行ったり，音声認識をかけ言語情報から判別を行う手法[3]は考えられるが，これらはシステムが受理できるコマンドの事前知識が必要となる．本研究ではあくまで人が機械に対して喋りかける場合と，人間同士で話し合う場合の自然な話し方の差から判別を行いたいので，このFBANKの結果から，一発話ごとの明瞭さを表す特徴量を求めたい．

一つの手法として，FBANKの各次元に対して，発話ごとのパワー平均を求める方法が考えられるが，実験結果(図11)よりこれだけでは3.1章の韻律特徴量と同程度の判別精度しかでないことが分かった．そこで我々は，明瞭度とはパワーが大きいことに加え時間変化量が大きいことと考え，音素のような短時間での変化から韻律のような比較的長時間の変化までを同時に取り扱える手法として，FBANKを基に時間軸方向にウェーブレット変

換をかけることを考える．Mother Wavelet としては，図 8 に示す Gabor Wavelet を用いた．

$$\Phi(t) = \frac{1}{\sqrt{2\pi\sigma}} exp(-\frac{t^2}{\sigma^2}) exp(-j2\pi ft) \qquad (3)$$

韻律的な変化を捉えるため，$\sigma$ は比較的長い 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．特に時間変化に関しては虚数部に現れると考えられる．これは Gabor Wavelet の虚数部が奇関数であるためである．システム要求発話と雑談それぞれに 8 段階にスケーリングした Gabor Wavelet をかけた結果 (虚数部) を図 10 に示す．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，これを明瞭度の特徴量とする．

## 4. 実　　験

実験では 3. 章で求めたそれぞれの特徴量を用い，1 発話ごとにシステムへの問いかけであるか雑談であるかを SVM を用いて判別する．SVM の Kernel 関数には RBF (Gaussian) Kernel を用い，10-fold のオープンによる評価を行った結果を図 11 に示す．またそのときの適合率と再現率を表 4 に示す．それぞれ特徴量，Prosodic8 は 3.1 章で述べた発話区間全体から求めた 8 次元の韻律特徴量，Prosodic24 は 3 区間に分けた発話区間から求めた 24 次元の韻律特徴量，Prosodic24+S.A. は Prosodic24 特徴量に 3.2 章で述べた話者交替の 6 次元を付加した特徴量，FBANK は 3.3 章で述べた 24 次元+エネルギーの計 25 次元 FBANK から求めた発話区間での平均パワー 25 次元，Gabor Re. は FBANK に Gabor Wavelet をかけた実部のパワー 200 次元による明瞭度特徴量，Gabor Im. は FBANK に Gabor Wavelet をかけた虚部のパワー 200 次元による明瞭度特徴量，Gabor は実・虚部両方を用いた 400 次元による明瞭度特徴量を表している．

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 (Prosodic24) ことでシステム要求判別精度が F 値で 15.0%上がっている．また，話者交替の特徴量を含める (Prosodic24+S.A.) ことで，さらに判別性能が 3.4%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．

一方，FBANK をそのまま用いた特徴量 (FBANK) も韻律特徴量 (Prosodic24) と同程度の高い判別精度が出ているが，時間方向に Gabor Wavelet をかけることにより，判別精度が大幅に向上している．Gabor Wavelet の実部のみをかけた場合 (Gabor Re.) でも FBANK より判別精度が 8.6%も高いのは，比較的長周期のウェーブレットをかけているため，発話の中での長周期的な変化が影響しているのではないかと考えられる．しかし，Gabor Wavelet の虚部を用いた特徴量 (Gabor Im.) の方が実部を用いた場合より 2.3%判別精度が高くなっている．これは，図 8 のように虚数部は奇関数になっており，虚数部

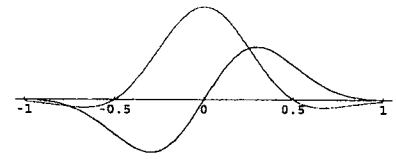


図 8　Gabor Mother Wavelet (赤:実数部 青:虚数部)
Fig. 8　Gabor Mother Wavelet


図 9　システム要求と雑談の FBANK
Fig. 9　FBANK of a System Request and a Chat

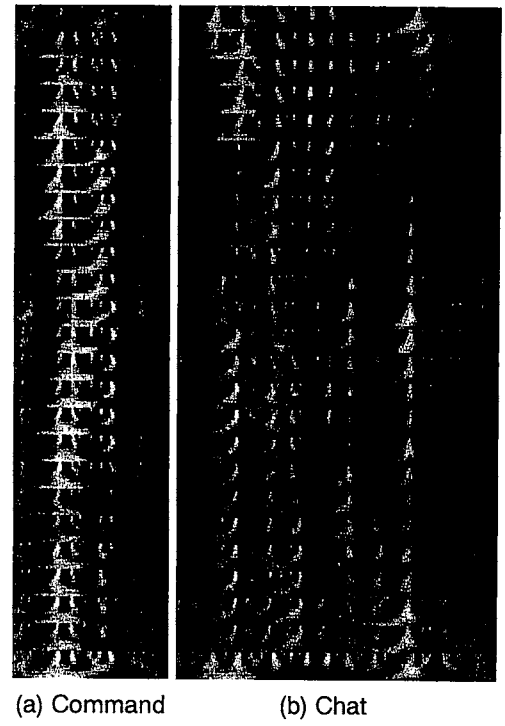


図 10　Gabor Wavelet 変換後の FBANK
Fig. 10　FBANK after calculating Gabor Wavelet

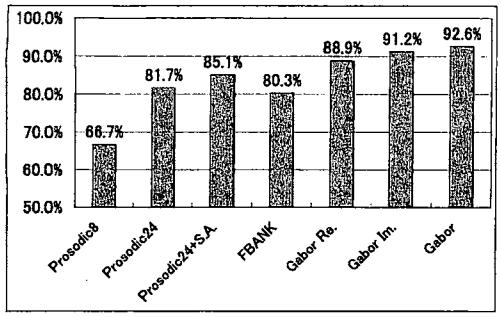


図 11 システム要求判別結果 (F 値)

Fig. 11 F-Measure of System Request Detection.

表 4 システム要求判別結果

Table 4 The Accuracy of System Request Detection.

| | Precision | Recall | F-measure |
|---|---|---|---|
| Prosodic8 | 0.584 | 0.806 | 0.667 |
| Prosodic24 | 0.756 | 0.889 | 0.817 |
| Prosodic24+S.A. | 0.832 | 0.870 | 0.851 |
| FBANK | 0.719 | 0.909 | 0.803 |
| Gabor Re. | 0.906 | 0.873 | 0.889 |
| Gabor Im. | 0.933 | 0.891 | 0.912 |
| Gabor | 0.943 | 0.909 | 0.926 |

をかけることにより変化量を取り出しているためと考えられる．また，Mother Wavelet をスケーリングすることにより，音素のような短周期の変化量から韻律のような長周期の変化量までうまく捉えているためと考えられる．実・虚部両方を用いた場合 (Gabor) が一番判別精度は高いが，虚部の変化成分は定常ノイズの影響をほとんど受けないと考えられるため，車内のようなノイズ環境下でも判別精度が落ちにくいと期待している．この点に関する評価は今後の課題としたい．

## 5. ま と め

本稿では，まずシステム要求と雑談を判別するために効果的な韻律特徴量を取り出すため，発話区間にマージンを設け，3 区間から特徴量を求める手法を提案した．次に，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．また，機械に対して話しかける場合と人間同士で雑談する場合の特徴の差が話し方の明瞭さに表れると考え，FBANK の時間方向に Gabor Wavelet をかけることにより，明瞭さの指標とする音素や韻律の変化量を取り出す手法を提案した．長周期から短周期までの様々な変化成分の大きさを求めることで，発話の明瞭度から高い精度でシステム要求判別ができることが分かった．

今後の課題としては，ノイズ環境下やカーナビを用いての実験等の様々な環境での評価があげられる．我々はこの他にも言語情報を用いたシステム要求判別手法等も提案しており，環境によってどのような特徴を用いるのが効果的なのか評価していきたい．

## 文 献


[1] Masataka Goto, Koji Kitayama, Katunobu Itou, and Tetsunori Kobayashi, "Speech Spotter: On-demand Speech Recognition in Human-Human Conversation on the Telephone or in Face-to-Face Situations", Proceedings of ICSLP-2004, pp.1533-1536, October 2004.

[2] Shinya Yamada, Toshihiko Itoh and Kenji Araki, "Linguistic and Acoustic Features Depending on Different Situations - The Experiments Considering Speech Recognition Rate", Proceeding of Interspeech 2005, pp.3393-3396, Sep. 4-8, 2005.

[3] 佐古 淳, 滝口 哲也, 有木 康雄, "AdaBoost を用いたシステムへの問い合わせと雑談の判別", 第 8 回音声言語シンポジウム, SIG-SLP64, pp.19-24, 2006-12

[4] 山形 知行, 佐古 淳, 滝口 哲也, 有木 康雄, "SVM を用いたシステムへの問い合わせと雑談の判別", 日本音響学会 2007 年春季研究発表会, 1-P-31, pp.185-186, 2007-03

[5] Tomoko Ohsuga, Masafumi Nishida, Yasuo Horiuchi, Akira Ichikawa, "Investigation of the Relationship between Turn-taking and Prosodic Features in Spontaneous Dialogue", Proceedings of Eurospeech2005, pp.33-36, 2005.9

[6] 佐藤 雅昭, "ウェーブレット理論の数学的基礎 第 I 部「非直交ウェーブレット」", 日本音響学会誌, vol.47, no.6, 405-415, 1991

[7] N.M.Fraser and G.N.Gilbert, "Simulating Speech Systems", Computer Speech and Language, vol.5, no.1, pp.81-99, 1991

[8] 山畠 利彦, 藤江 真也, 小林 哲則, "対話システムのための視線方向認識", MIRU 2006, IS3-38, pp.1237-1242, 2006

[9] "大語彙連続音声認識エンジン Julius", http://julius.sourceforge.jp/

[10] "SVM-Light Support Vector Machine", http://svmlight.joachims.org/

[11] M. Omologo and P. Svaizer, "Acoustic source location in noisy and reverberant environment using CSP analysis," Proc. ICASSP, pp. 921-924, 1996