音声スタータ: “SWITCH” on Speech

北山 広治† 後藤 真孝†† 伊藤 克亘‡ 小林 哲則†

† 早稲田大学 理工学部 〒169-8555 東京都新宿区大久保 3-4-1
†† 産業技術総合研究所 〒305-8568 茨城県つくば市梅園 1-1-1
‡ 名古屋大学大学院 情報科学研究科 〒464-8603 名古屋市千種区不老町 1

E-mail: †{kitayama, koba}@tk.elec.waseda.ac.jp, ††m.goto@aist.go.jp, ‡itou@is.nagoya-u.ac.jp

あらまし 本稿では，非言語情報の一つである言い淀み (有声休止) を活用し，ユーザが音声認識を開始してほしいタイミング (発話区間の始端) を，言い淀むことによって明示的に指示できる「音声スタータ」という新しい音声インタフェースを提案する．通常の音声認識システムは，発話区間の切り出し後に音声認識を行うため，雑音下での切り出しミスが認識精度に悪影響を与え，頑健性を保証することが困難であった．我々は，有声休止が雑音下でも頑健に検出できると考え，常に有声休止の途中から音声認識を開始することで，信頼性の高い発話区間の検出方法を実現することを試みる．様々な雑音環境下で 4 種類の発話区間検出方法を比較実験した結果，音声スタータは他の検出方法に比べ，特に低 SNR(10dB 以下) の条件で高い性能が得られた．

Speech Starter: “SWITCH” on Speech

Koji Kitayama† Masataka Goto†† Katunobu Itou†† Tetsunori Kobayashi†

†School of Science and Engineering, Waseda University
††National Institute of Advanced Industrial Science Technology (AIST)
‡Graduate School of Information Science, Nagoya University

E-mail: †{kitayama, koba}@tk.elec.waseda.ac.jp, ††m.goto@aist.go.jp, ‡itou@is.nagoya-u.ac.jp

Abstract In this paper we propose a speech interface function, called *speech starter*, that enables noise-robust endpoint (utterance) detection for speech recognition. When current speech recognizers are used in a noisy environment, a typical recognition error is caused by incorrect endpoints because their automatic detection is likely to be disturbed by non-stationary noises. The speech starter function enables a user to specify the beginning of each utterance by uttering a filler with a filled pause, which is used as a trigger to start speech-recognition processes. Since filled pauses can be detected robustly in a noisy environment, reliable endpoint detection is achieved. Experimental results from a 10-dB-SNR noisy environment show that the recognition error rate with speech starter was lower than with conventional endpoint-detection methods.

1 はじめに

従来の一般的な音声認識システムは，入力信号から発話区間の始端と終端を同定した後に，音声認識を行っていた．そのため，発話区間の検出精度が低い環境では，音声認識システム全体の性能が大きく低下することがあった．典型的な発話区間の検出方法として，零交差数と短時間エネルギーの二つの特徴量を元に，発話区間を検出する方法 [1] が挙げられる．この方法は，静かな室内では正確な発話区間検出が可能となる．しかし，実環境で入力される音声は，利用者や環境によって音声や雑音のエネルギーが大きく異なり，検出精度が低下する問題が生じていた．また，咳払いや息，第三者との会話などの，ユーザがシステムに入力するつもりのない音が，認識対象の発話区間と誤検出される問題もあった．そのため，ユーザは咳払いなどをせずに誤りなく音声入力しなければならず，音声認識システムの操作性を損なっていた．

上記の問題に対処する代表的な方法として，ボタンを使う方法が挙げられる．ユーザが話している間，自分でボタンを押し続けてシステムに発話区間を指示する方法である．しかし，ボタン操作が前提なために，他のデバイスを使わないハンズフリーな音声認識は実現できない．さらに，実際にはユーザがボタンを押すタイミングが早過ぎたり遅過ぎたりすることがあるため，特に雑音環境下では，発話区間に雑音が含まれか

ねないため頑健な音声認識が困難となる．

他の解決法として，発話区間検出に用いる音響特徴量の使用方法を改善する研究もなされてきた [2, 3, 4]．しかし，非定常な雑音に対しては，まだ性能向上の余地がある．また別の解決法として，発話区間を明示的に検出せずに，連続して音声認識を行う手法 [5, 6] が提案されている．この方法は，発話を検出し損なう可能性は少ないが，ユーザが入力するつもりのない発話 (咳払いや，第三者へ対する発話など) や，背景雑音を全て認識対象とする誤りが生じやすく，これらの発話を適切に棄却しなければならない．一方，カメラで捉えた話者の顔の動きに基づいて発話区間を切り出す方法 [7] も研究されており，雑音環境下で頑健に発話区間を検出することに成功している．しかし，カメラが利用できない環境では，この方法を適用できない．

そこで我々は，上記の問題を解決する新しい音声インタフェース「音声スタータ」を提案する．人間は話し始めるときに，しばしば有声休止 (言い淀みに含まれる，母音の引き延ばし) を含む「えー」や「あのー」といったつなぎ語を言うことがある．音声スタータでは，この音声の非言語情報の一つである有声休止に着目し，音声認識を開始するトリガとして活用する (音声認識器は，有声休止から認識処理を開始する)．有声休止は，自然対話中で頑健に検出できることが報告されている [8]．有声休止がパワーの安定した母音を伴うことから雑音下でも検出可能と考えられ，これを発話区間の始端とみなせば，頑健な発話区間検出が期待できる．音声スタータでは，ユーザが発話開始時に常に有声休止を発声することをルール化することが重要となる．こうすることで，ユーザが認識してほしい発話区間を，その先頭で故意に言い淀むことによって明示的に指示できるという利点が得られる．

本稿では，以下，音声スタータの概要と利点を紹介する．次に，音声スタータに基づく音声認識システムの具体的な実現方法を説明し，システムの実装について述べる．そして，7 種類の雑音環境下での評価実験を通じて，音声スタータが他の発話区間検出方法に比べて頑健であることを示す．最後に，まとめを述べる．

2　音声スタータ

音声スタータは，ユーザが他のデバイスを使わずに，音声だけで音声認識開始点を明示的に指示できる音声インタフェース機能である．ユーザが，認識して欲しい各発話の先頭で必ず有声休止を発声して言い淀むことで，音声認識器の認識開始点を思い通りに制御可能となる．発話の先頭で言い淀む際には，有声休止を含む様々なつなぎ語を用いることができる．例えば，ユーザが「藤井フミヤ」を入力したい場合,「えー、藤井フミヤ」や「あのー、藤井フミヤ」のように，故意に言い淀んだ後に単語を発声すればよい．

音声スタータは，以下の 3 つの利点を有する．

1. 雑音に頑健な発話区間検出
   母音のエネルギーは音声信号中で高い傾向にあるため，母音の引き延ばしから成る有声休止も同様に高いエネルギーを持ち，雑音環境下で頑健に検出しやすい．雑音下で音声認識を行う際には，雑音から認識を開始して誤認識する問題が生じるが，音声スタータでは常に安定した母音から認識を開始するため，そういった問題を回避できる．また，有声休止が検出されるまでは認識を開始しないため，非定常な突発的雑音を誤って発話区間と検出することがない．
2. 利用者に負担の少ない操作方法
   音声スタータは，特別な訓練をせずに使うことができ，ユーザの負担も少ない．音声スタータでは，ユーザは故意に有声休止を発声してから話し始めなければならないが，日常会話でも言い淀んでから話し始めることがよくあるため，そのような話し方には馴れているといえる．
3. マイク以外の装置が不要
   音声スタータは音声のみで利用可能なため，ボタンやカメラなどのマイク以外の装置を必要としない．そのため，音声スタータを組み込んだアプリケーションシステムはコンパクトに実現することができ，作成者と利用者の双方にとって負担が少なくなる．

このような音声スタータを実現するには，まず，リアルタイムに有声休止を自動検出する必要がある．次に，検出結果に基づいて音声認識器の認識開始点を決定し，実際の認識処理を始めなければならない．最後に，音声認識器の認識終了点を，音声入力中の刻々と変化する認識結果に基づいて決定する必要がある．以下では，これらの具体的な方法を順に説明する．

2.1　有声休止の検出

言語的な制約を一切用いずに有声休止を検出するために，後藤らのリアルタイム有声休止検出手法 [8] を用いる．この手法は，有声休止の始端と終端を，基本周波数の変動，スペクトル包絡の変形，の両方が小さいという音響的特徴から検出するボトムアップな信号処理手法であり，あらゆるつなぎ語中の母音の引き延ばしを検出可能である．

2.2　発話始端の検出

検出した有声休止区間に基づいて，発話始端 (音声認識開始位置) を決定する様子を図 1 に示す．有声休止区間の途中に発話始端が位置するようにし，有声休止を含めて音声認識を行う．これは，有声休止区間の

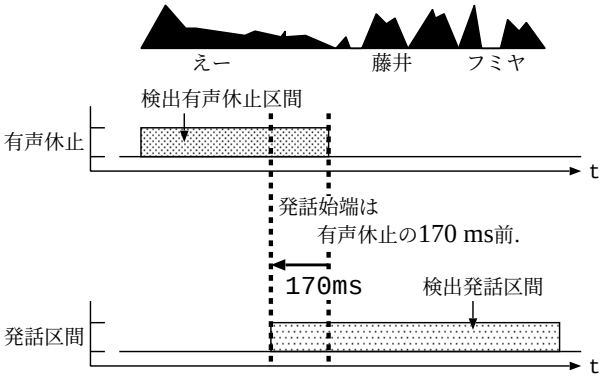


図 1: 発話始端検出

終端付近に音素遷移の過渡的な現象が現れることがあり，終端から音声認識すると誤認識を招く可能性があるからである．有声休止区間の途中ならば安定した母音であることが分かっているので，上記のようにそこを発話始端とすることで，確実に母音から認識対象の発話部分へと続く区間を音声認識できる．ただし，有声休止区間の始端近くの方を発話始端とすると，音声認識対象となる有声休止の部分が長くなり，突発的な雑音の混入により誤認識を招く可能性が増えるため，必要以上に前にしない方がよい．これらを考慮して，現在の実装では，有声休止区間の終端から 170 ms 手前の時点を発話始端としている．

2.3 発話終端の検出

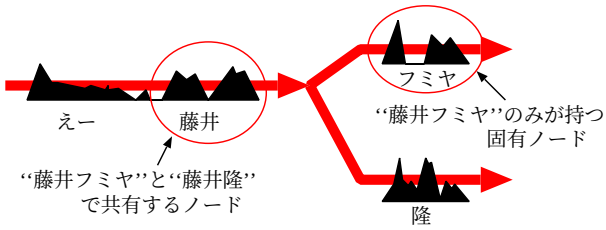


図 2: 発話終端検出: 一定時間固有ノードに停留した場合に認識を終える．

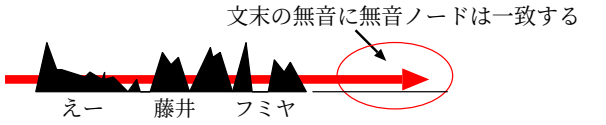


図 3: 発話終端検出: 一定時間無音ノードに停留した場合に認識を終える．

発話区間の終端は，音声認識器が認識を開始した後に検出する．発話区間の終端検出には，井ノ上らの方法 [9] と内藤らの方法 [10] を基に行う．具体的には，音声認識途中の最尤仮説を使って行う．毎フレームの最尤仮説を調べ，以下に挙げる二つのノードのいずれかに，一定時間以上停留している場合，そのフレームを発話終端とする．この停留時間の閾値は，現在の実装では，予備実験の結果から 200 ms としている．

1. 最尤仮説が，木構造辞書の複数単語に共有されていない固有ノード (一単語のみのノード) に停留している [9]．図 2 参照．
2. 最尤仮説が，文末の無音にあたる，無音ノードに停留している [10]．図 3 参照．

3 実装

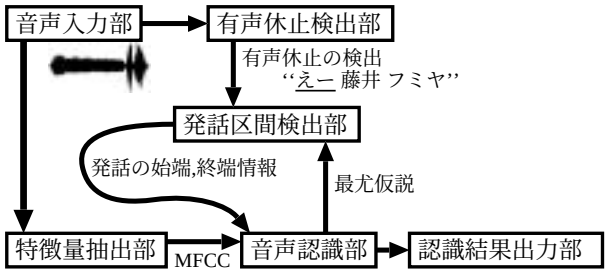


図 4: システム構成

図 4 に音声スタータを用いた音声認識システムの構成を示す．図中の，枠線で囲われた部分は，それぞれ異なるプロセスを示す．各プロセスは LAN を介して複数の計算機上に分散して配置され，RVCP(Remote Voice Control Protocol) [11] を使い協調して動作する．音声認識部は CSRC の日本語ディクテーション基本ソフトウェア (julian 3.3beta)[12, 13] を，毎フレームの単語仮説を発話区間検出部に送信することができるように手を加えて用いた．

音声入力部に入力された音声信号は，有声休止検出部と，特徴量抽出部の両者に平行して処理される．発話区間検出部は有声休止の終端情報を受け取り，発話始端を決め，音声認識部にその情報を送信する．音声認識部は特徴量抽出部から MFCC パラメータを受け取り，検出された発話始端から認識を行う．そして，単語仮説を毎フレームごとに発話区間検出部に送る．発話区間検出部は受け取った単語仮説を元に，発話終端条件を十分満たしているかどうかを判別し，満たしているならば終端を検出し，終端情報を音声認識部に送信する．最後に，認識結果出力部が認識結果を表示する．

4 孤立単語認識実験

音声スタータが雑音環境下で頑健に機能することを確認するために，様々な雑音を混合した音声データを

用いて，以下の 4 つの発話区間検出手法を比較評価する実験をした．

1. 音声スタータ
2. 零交差数と短時間エネルギーに基づいて発話区間を検出する方法 [1]
3. 発話区間を検出せずに音声認識を行うための，julian[12, 13] に実装されているショートポーズでセグメンテーションを行う方法 [6]
4. 発話区間の先頭で常にキーワードを発声することをルールとし，キーワードに基づいて発話区間の始端を決定する方法．

キーワードを用いる 4. の方法では，音声スタータと同じように，発話区間の始端をユーザが明示的に指示する．しかし，有声休止の代わりにキーワードを定め，「もしもし、藤井フミヤ」のようにキーワード (ここでは「もしもし」) を常に発声する点が音声スタータとは異なる．システムはキーワードを検出して認識処理を開始する．キーワードの検出 (キーワードを受理するかの判定) は以下のように実現する．まず，3. の方法で音声認識を行う．次に，キーワードと認識された区間に対して，文法の拘束を無くし (全ての音素で遷移を許可)，もう一度認識処理を行い，尤度を得る．この結果を対立候補と考え，2 つの尤度の差を閾値と比較する．閾値を越えていたならば，キーワードに続く認識結果を受理する．そうでなければ棄却する．キーワードとして用いる単語は，この方法に関する過去の研究事例が乏しいため，親しみやすいという点から，上記で例にあげた「もしもし」を使う．

認識対象は孤立単語に限定する．ただし，ここでの単語は，音声認識器の単語辞書上 (言語モデル上) の 1 単語とする．本実験では，日本のポピュラー音楽のヒットチャートから作成した単語辞書 [14] を用い，曲名 (342 語) とアーティスト名 (179 語) をそれぞれ 1 単語として登録した．

4.1 実験条件

図 5 は 4 手法の評価に用いた音声データの作成方法を示している．179 語の単語と，有声休止 (「えー」) を 11 発話，キーワード (「もしもし」) を 11 発話，を男性話者一人から収録した．音声スタータには，収録した有声休止と単語を結合して一つの単位発話とする．キーワードを用いる方法に使用するデータも同様に，キーワードと単語を結合して，一つの単位発話とする．発話の検出性能を評価するために，各発話の間に 5 秒間の無音 (背景雑音) を挿入する．

音声データは，下記 7 種類の実環境雑音 [15] を 5 種類の SNR (0, 10, 20, 30, 40 dB) で混合する．SNR は，単位発話 179 個の全区間の平均エネルギーと，それと同区間の雑音の平均エネルギーから計算した．

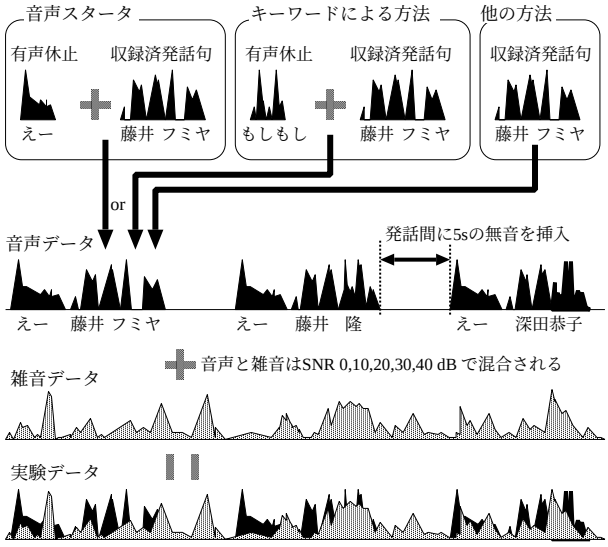


図 5: 実験データの作成方法

- 走行自動車内 [1500cc クラス] (低域にエネルギーが集中している)
- 展示会場 [ブース内] (時々人の話し声が聞こえる)
- 展示会場 [通路] (時々人の話し声が聞こえる)
- 交差点 (車の通行音が聞こえる)
- 列車 [在来線] (断続的に列車の激しい走行音がする)
- 計算機室 [ワークステーション] (定常的なファンの音が聞こえる)
- エレベーターホール [百貨店] (人の雑踏や話し声が聞こえる)

本実験で用いた音声特徴量は，MFCC 12 次元 + ΔMFCC 12 次元 + Δpower 1 次元の計 25 次元 とした．CMN (Cepstrum Mean Normalization) は行わない．音響モデルは，ASJ-JNAS，ASJ-PB の男性話者 133 人分 (計 20414 文)[16] から学習し，混合数 16，状態数 2000 のトライフォンモデルとした．

図 6 は音声スタータの文法を，図 7 はキーワードを用いる方法の文法を，図 8 はそれ以外の方法で用いた文法を示している．音声スタータの文法は,「えー」「んー」「うー」などの有声休止を含む代表的な 14 語のつなぎ語から始まる．キーワードを用いる方法の文法は，キーワード「もしもし」から始まる．それ以外の文法は，無音から始まる．

零交差数と短時間エネルギーによって発話区間を検出する方法で用いるエネルギーの閾値とキーワードを用いる方法の閾値は，評価データとは異なり，発話間の無音は 3 秒間にし，展示会場 [ブース内] を SNR 20 dB で混合して作成した．実験対象の 0-40 dB どの場合にも対応できるように，0-40 dB の中間値にあたる，SNR 20 dB とした．こうして作成したデータの発話検出率が最もよくなる値を閾値とし，以後の実験でこ

の値を用いた．

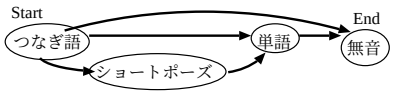


図 6: 音声スタータの文法

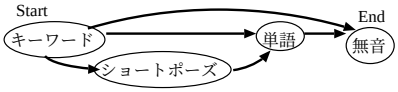


図 7: キーワードを用いる方法の文法

図 8: 他手法の文法

4.2 評価方法

実験データの発話区間の出現位置とその発話の正解単語を記述した正解単語ラベルと，システムの出力 (発話区間と，それの認識結果) を比較する．比較した結果の，正解と認識結果との適合具合を F 値を用いて評価する．F 値は再現率 (R) と適合率 (P) を統合させた評価尺度である．F 値，及び再現率，適合率は式 (1), (2), (3) により計算する．また，F 値の計算は再現率と適合率を等価に扱うため $\beta = 1$ とした．

$$F\text{ 値} = \frac{(\beta^2+1)PR}{\beta^2 R + P} \tag{1}$$

$$R = \frac{\text{正しく認識した単語数}}{\text{正解の発話総数 (179)}} \tag{2}$$

$$P = \frac{\text{正しく認識した単語数}}{\text{検出された発話の総数}} \tag{3}$$

ただし，単語に先行するつなぎ語を，別のつなぎ語と誤認識した場合，誤りと数えなかった (例えば,「えー」を「あー」と誤認識しても誤りとしなかった).

4.3 実験結果

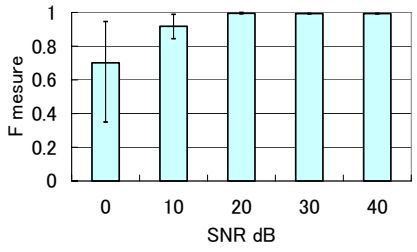


図 9: 有声休止検出率の平均値

まず，発話区間検出手法の比較実験に先だって，有声休止が雑音下でどの程度頑健に検出できるかを調査した．有声休止のリアルタイム検出手法 [8] による検出率を，7 種類の雑音環境においてそれぞれ求め，それらを平均した値を図 9 に示す．棒グラフが平均値，高低線が最大値と最小値を意味する．図 9 によれば，SNR 20-40 dB ではほぼ完全に有声休止を検出でき，SNR 0, 10 dB の高雑音環境下でも比較的高い性能で検出できていることが分かる．

次に，4 種類の発話区間検出手法による音声認識結果の比較を，図 10-17 に示す．図 10-16 は，7 種類の雑音環境における音声認識結果を，図 17 は，7 種類の雑音環境における認識結果の平均を示している．図 17 によれば，音声スタータが SNR 0, 10, 20 dB において性能が高く，他の方法に比べ有効であることが分かる．特に，SNR 10 dB 以下で他の手法との差が大きく，雑音の大きい環境で有利に働いている．零交差数と短時間エネルギーを用いる方法は，閾値を SNR 20 dB で決めたので，10 dB 以下で著しく性能が悪くなっている．また，個別の結果である，図 10-16 をみると，音声スタータは，図 14，15 の二つの雑音以外では，性能が高かった．図 10 はエネルギーが低域に集中する特徴を持つ雑音，図 11，12，16 は人の声や音楽による雑音の結果であり，こうした雑音に対して音声スタータは頑健な傾向があることが分かる．

これらの結果から，音声スタータは，雑音の強い環境下で，頑健に発話区間を検出することが，十分に可能であり，特に車内雑音に有効に働くことが分かった．

5 おわりに

本稿は，音声の非言語情報である有声休止を積極的に利用し，ユーザが発話区間の始端を音声だけで指定できる「音声スタータ」という音声インタフェースについて述べた．音声スタータは，特別な訓練をせずに利用でき，マイク以外の装置が不要なだけでなく，雑音下で頑健に発話区間を検出可能という利点も持つ．実際に，従来の発話区間検出方法と比べ，雑音の大きい環境下において有効であることを確認した．

これまでにインタフェース機能として非言語情報を積極的に活用した研究として,「音声補完」[11, 14] や「音声シフト」[17] が提案されている．音声補完では，有声休止をキーボードの TAB キー (補完トリガキー) とみなし，音声シフトでは，声の高さをキーボードの SHIFT キー (モード切替えキー) とみなして活用した．音声スタータは，非言語情報の新しい活用法を切り拓くという意味で，これらの研究に続くものとして位置づけられる．音声スタータでは，有声休止によってマイクのスイッチの機能を実現し，ユーザが自分の望むタイミングで音声入力が可能となる利点を持つ．

今後は，音声スタータをカーナビゲーションシステムに組み込み，ユーザビリティの評価を行う予定である．

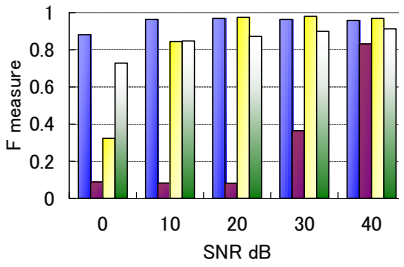


図 10: 走行自動車内 [1500cc クラス]

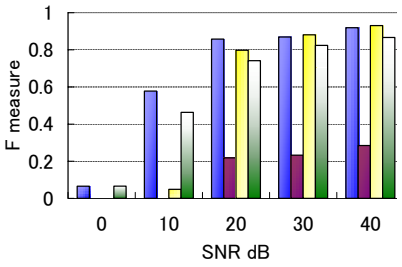


図 11: 展示会場 [ブース内]

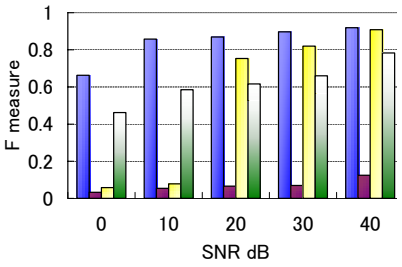


図 12: 展示会場 [通路]

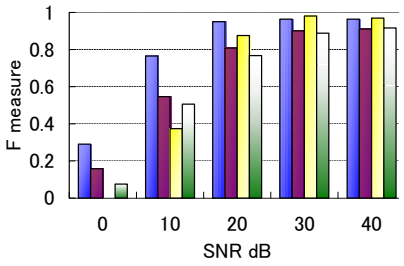


図 13: 交差点

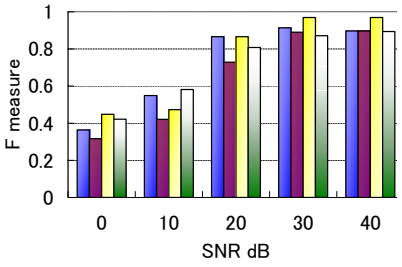


図 14: 列車 [在来線]

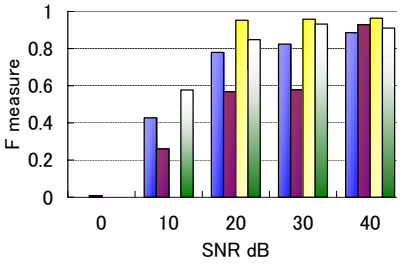


図 15: 計算機室 [ワークステーション]

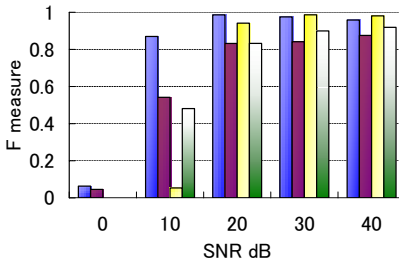


図 16: エレベーターホール [百貨店]

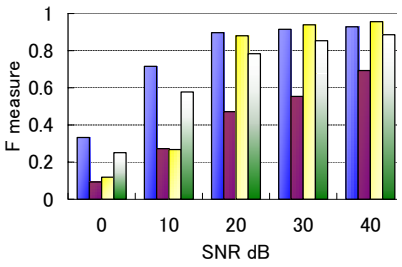
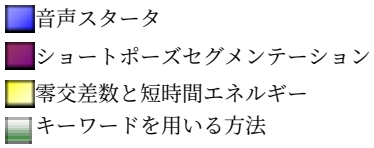


図 17: 実験結果の平均値



参 考 文 献

[1] L.R.Rabiner and M.R.sambur, “An Algorithm for Determining the Endpoints of Isolated Utterances,” The Bell System Technical Journal, vol.54, no.2, pp.297-315, Feb. 1975.

[2] Sahar E. Bou-Ghazale and Khaled Assaleh, “A Robust Endpoint Detection of Speech for Noisy Environments with Application to Automatic Speech Recognition,” Proc. ICASSP 2002, pp.3808–3811, 2002.

[3] A. Martin, D. Charlet and L. Mauuary, “Robust Speech / Non-speech detection using LDA applied to MFCC,” Proc. ICASSP 2001, pp.237–240, 2001.

[4] L. sheng Huang and C. ho Yang, “A Novel approach to robust speech endpoint detection in car environments,” Proc. ICASSP 2000, pp.1751–1754, 2000.

[5] O. Segawa, K. Takeda and F. Itakura, “Continuous Speech Recognition without End-point Detection,” Proc. ICASSP 2001, pp.245–248, 2001.

[6] 河原 達也, 加藤 一臣, 南篠 浩輝, 李 晃伸, “話し言葉音声認識のための言語モデルとデコーダの改善,” 情処技報 2001-SLP-36, pp.15-22, 2001.

[7] Murai, K. Kumatani, K. Nakamura, S.: “A Robust End Point Detection by Speaker’s Facial Motion,” Proc. HCI International 2001, pp.199–202, 2001.

[8] 後藤真孝, 伊藤克亘, 速水悟, “自然発話中の有声休止箇所のリアルタイム検出システム,” 信学論 (D-II), Vol.J83-D-II No.11, pp.2330-2340, 2000.

[9] 井ノ上 直己, 中村 誠, 酒寄 信一, 山本 誠一, 谷戸 文廣, “単語固有セルでのゆう度判定を用いた音声認識処理の高速化手法,” 信学論 (D-II), Vol.J79-D-II No.12, pp.2110-2116, 1996.

[10] 内藤 正樹, 黒岩 眞吾, 山本 誠一, 武田 一哉, “部分文仮説のゆう度を用いた連続音声認識のための音声区間検出法,” 信学論 (D-II), Vol.J80-D-II No.11, pp.2895-2903, 1997.

[11] 後藤 真孝, 伊藤 克亘, 速水 悟: “音声補完: “TAB” on Speech,” 情処研報, 2000-SLP-32-16, pp.81-86, 2000.

[12] 鹿野 清宏, 伊藤 克亘, 河原 達也, 武田 一哉, 山本 幹雄 編著, “IT Text 音声認識システム,” オーム社, 2001

[13] A. Lee, T. Kawahara and K. Shikano, “Julius — an open source real-time large vocabulary recognition engine,”Proc. Eurospeech2001, pp.1691–1694, 2001

[14] 後藤 真孝, 伊藤 克亘, 秋葉 友良, 速水 悟: “音声補完: 音声入力インタフェースへの新しいモダリティの導入,” コンピュータソフトウェア (日本ソフトウェア科学会論文誌), Vol.19, No.4, pp.10-21, 2002.

[15] 電子協騒音データベース, “http://www.milab.is.tsukuba.ac.jp/corpus/noise_db.html”

[16] K. Itou, M. Yamamoto, K. Takeda, T. Takezawa, T. Matsuoka, T. Kobayashi, K. Shikano, S. Itahashi: “The Design of the Newspaper-Based Japanese Large Vocabulary Continuous Speech Recognition Corpus,” Proc. ICSLP98, pp.3261-3264, 1998.

[17] 尾本 幸宏, 後藤 真孝, 伊藤 克亘, 小林 哲則, “音声シフト: “SHIFT” on Speech,” 情処技報 2002-SLP-40, pp.13-18 , 2002.