科学研究費助成事業　　研究成果報告書

平成 ２６ 年　 ６ 月 ２３ 日現在

機関番号：３３９３４
研究種目：基盤研究(C)
研究期間：2011～2013
課題番号：２３５００２３０

研究課題名（和文）音声対話システムにおける音に着目した周囲状況推定技術の研究

研究課題名（英文）Estimation of environmental conditions based on acoustic signals for speech dialog systems

研究代表者
實廣　貴敏（JITSUHIRO, Takatoshi）

愛知工科大学・工学部・准教授

研究者番号：６０３９４９９６

交付決定額（研究期間全体）：（直接経費）　　4,100,000円、（間接経費）　　1,230,000円

研究成果の概要（和文）：音声対話システムでは周囲やユーザの状況はあまり考慮されない．そこで，音声自体をできるだけ利用し，周囲状況を推定する技術を提案する．単一マイクロホンにおいて，音声から音響伝達周波数特性を推定することで，発話者の口からマイクまでのおよその距離を推定する．距離ごとに音響伝達周波数特性のテンプレートを用意しておく．当面，雑音はほぼないと仮定し，あらかじめ用意した雑音のないクリーン音声モデルと入力音声との周波数特性の差分を推定する．この差分が距離に依存した特性となり，テンプレートとの照合により，最も近いものが示す距離が推定された距離となる．１ｍより近い／遠いの識別では8割程度の精度が得られた．

研究成果の概要（英文）：The conditions of users and environments are not considered much in speech dialog systems. We propose the estimation method of environmental conditions based on acoustic signals. The distance from a user to a system is estimated using estimated an acoustic transfer function. At first, templates of frequency characteristics depending on the distance are created from impulse responses for each distance. To estimate the transfer functions, we calculate the difference between input speech and clean speech using clean speech models. This difference means the transfer function depending on the distance. The distance can be estimated from the nearest template. The almost 80 % accuracy was obtained in the discriminative experiments that was to decide if the distance is closer or far than 1 m.

研究分野：総合領域

科研費の分科・細目：情報学・知覚情報処理・知能ロボティクス

キーワード：音声認識　音声対話システム　音源距離推定　単一マイクロホン　音響モデル　VQコードブック

１．研究開始当初の背景

音声対話システムでは，基本的に入力される音声が音声認識対象であると仮定して作られているのかが一般的である．認識対象でない音声が観測された場合，適切に処理できず，ユーザから見てちぐはぐな応答になることが多い．入力を検知し，音声認識したとき，そのとき周囲やユーザがどのような状況であったかの考慮があまりなされていないことが問題である．

そこで，音声自体をできるだけ利用して，周囲状況を推定し,システム全体でより適切な応答ができるような技術を検討する．また，ユーザとシステムのターン数が多い対話システムでは，対話内容から徐々に情報を得ていく手法も多い．しかし，１問１答程度の簡単な対話システムでは，言語的情報の利用は困難である．そこで，観測される音声から得られる情報に着目し，簡単なシステムであっても，より客観的に周囲状況を推定しつつ，適切に応答できる手法の確立に努める．

２．研究の目的

本研究では，単一マイクロホンにおいて，音声から空間伝達特性を推定することで，発話者の口からマイクまでのおよその距離を推定する．マイクから離れた音声には，周囲の壁等の反射も含まれ，その影響が大きい．マイクに近い音声では，直接波が強く，反射の影響が少ない．あらかじめ，対話システムを設置した場所の空間伝達特性を知っておく必要はあるが，およその距離はその特性から推定できる．また，マイクから近ければ，おそらくユーザは対話システムに話しかけていると考えられ，その音声は認識対象である可能性が高い．マイクから遠ければ，逆に認識対象である可能性が低い．その情報を元に，音声認識結果自体の信頼度と合わせ，その認識結果を信じて応答すべきか，再確認するような質問を発した方がいいのか，それとも反応しない方がいいのか，などの判断を行う．これにより，正確な応答が可能になる．

複数マイクロホンを使うことも可能ではあるが，規模が大きくなったり，コストが高くなったりする．単一マイクロホンで距離を推定することができれば，対話システムの認識精度向上，コスト削減やシステムの縮小化など様々な利点がある．本研究では，その条件下で,比較的容易な方法としてVQ (Vector Quantization)コードブックを利用して，話者までのおよその距離を推定する手法を提案する．

マイクロホンから離れて発話した音声は，部屋の壁などに反射され，歪みが生じる．この歪みは音源からマイクロホンまですべての音響的情報を含み，音響伝達特性として利用できる．

音響伝達特性は発話の距離に応じて変わる．あらかじめマイクロホンから距離別で音響伝達特性を調べ，テンプレートを作成し，入力音声の歪みとマッチングをすることで，音源の距離を推定する．

入力音声の歪みを推定するため，クリーン音声データを用いてk-means法によるVQコードブックを音響モデルとして用いる．入力音声スペクトルとVQコードブックに含まれるセントロイドとを比較し，最も近いセントロイドをクリーン音声とすることで，歪みを求める．

さらに，音声区間のみを利用するために，音声区間検出を利用したり，歪みの繰り返し推定を導入したりして精度向上を図る．

３．研究の方法

(1) VQ コードブックを用いた音響伝達特性の推定手法

提案法を説明する．あらかじめ，それぞれの位置からマイクロホンまでの音響伝達特性をテンプレート化しておく．入力音声から推定した音響伝達特性とテンプレートを比較し，最も近いテンプレートの位置が，推定された発話者の位置となる．

次に，入力音声から音響伝達特性を推定する方法について述べる.時刻tにおける入力音声のスペクトル$\boldsymbol{X}_t$は下記のように表される．

$$\boldsymbol{X}_t = \boldsymbol{H}_t \cdot \boldsymbol{S}_t + \boldsymbol{N}_t \qquad ①$$

ここで，$\boldsymbol{S}_t, \boldsymbol{H}_t, \boldsymbol{N}_t$はそれぞれクリーン音声のスペクトル，音響伝達周波数特性，加算性雑音のスペクトルである．ここでは，加算性雑音$\boldsymbol{N}_t$は無視できるほど抑圧できているとする．このとき，

$$\boldsymbol{X}_t \approx \boldsymbol{H}_t \cdot \boldsymbol{S}_t \qquad ②$$

と近似できる．両辺の対数を取ると推定される対数音響伝達特性は，

$$\log \widehat{\boldsymbol{H}}_t \approx \log \boldsymbol{X}_t - \log \boldsymbol{S}_t \qquad ③$$

と表せる．$\log \boldsymbol{S}_t$が推定できれば，$\log \widehat{\boldsymbol{H}}_t$を求めることができる．$\log \boldsymbol{S}_t$を推定するために，クリーン音声データベースから VQ コードブックを作成しておき，入力に最も近いセントロイド･ベクトル$\log \boldsymbol{C}_i$を探し,それを$\log \boldsymbol{S}_t$とする．ここでは距離として下記のように定義する．

$$d_i = (\log \boldsymbol{X}_t - \log \boldsymbol{C}_i)'(\log \boldsymbol{X}_t - \log \boldsymbol{C}_i) \qquad ④$$

‘’は転置を示す．この値が最も小さくなる$\log \boldsymbol{C}_i$を推定されたクリーン音声$\log \widehat{\boldsymbol{S}}_t$とする．

$$\log \widehat{\boldsymbol{H}}_t \approx \log \boldsymbol{X}_t - \log \widehat{\boldsymbol{S}}_t \qquad ⑤$$

$\log \widehat{\boldsymbol{S}}_t$は元の音声の特性と大まかなところでは一致すると考えられ，入力音声との差分がその空間での音響伝達特性となる．ただし，1 フレームごとの推定では不安定になると考えられる．そこで，対数音響伝達特性$\log \overline{\boldsymbol{H}}$は数秒程度の１発話においては一定と仮定できるので，全てのフレームに対する平均として推定できる．

$$\log \overline{\boldsymbol{H}} = \frac{1}{T}\sum_{t=1}^{T} \log \widehat{\boldsymbol{H}}_t \qquad ⑥$$

ここで，Tは総フレーム数である．これにより，安定に音響伝達特性を推定できる．

(2) 音声区間検出の利用

提案手法では，クリーン音声の VQ コードブックを利用し，入力音声フレームに最も近いクリーン音声セントロイドを用いる．すなわち，入力音声自体に音声が含まれていないと，無音区間の雑音などとマッチングし，不要な歪みを推定してしまう可能性がある．そこで，音声区間を検出しておき，その区間のみを用いて提案手法を用いる．VQ コードブックが十分に音声の特徴を含んでいれば，比較的近い特性を見つけられる可能性が高い．

(3) VQ コードブックを用いた音響伝達特性の繰り返し推定手法

さらなる精度向上のため，推定を繰り返し行う手法を提案する．(1)で述べた手法では④式での距離尺度において，入力音声特性$\boldsymbol{X}_t$に最も近いセントロイド$\boldsymbol{C}_i$を選択するが，$\boldsymbol{X}_t$自体は歪み$\boldsymbol{H}_t$を含んでいるため，その歪み特性によっては，不適切な$\boldsymbol{C}_i$を選択する可能性がある．そこで，1 度，音響伝達特性$\bar{\boldsymbol{H}}$を推定した後，クリーン音声$\tilde{\boldsymbol{S}}_t$を推定する．今度はその$\tilde{\boldsymbol{S}}_t$を$\boldsymbol{X}_t$の代わりに用い，もう一度，最も近いセントロイド$\boldsymbol{C}_i$，さらには，音響伝達特性$\boldsymbol{H}_t$, $\bar{\boldsymbol{H}}$を推定する．繰り返し得られる$\bar{\boldsymbol{H}}$の値の変化が小さくなったら，推定処理を停止する．以下に，この繰り返し手法をまとめる．

【VQ コードブックを用いた音響伝達特性の繰り返し推定手法】

Step. 1　初期化．$n=1$とおく．

Step. 2　すべてのフレーム$(t, 1 \le t \le T)$に対し，$\hat{\boldsymbol{H}}_t^{(n)}$ベクトルを以下のように推定する．

i)　$n=1$のとき，$\tilde{\boldsymbol{S}}_t^{(n-1)} = \boldsymbol{X}_t$とする．

ii) 最も近いセントロイドが見つかったら，それを新しいクリーン音声$\hat{\boldsymbol{S}}_t^{(n)}$とする．

$$\log \hat{\boldsymbol{S}}_t^{(n)} = arg \max_{\log C_i} \left(\log \tilde{\boldsymbol{S}}_t^{(n-1)} - \log C_i\right)' \times \left(\log \tilde{\boldsymbol{S}}_t^{(n-1)} - \log C_i\right)$$

iii) $\hat{\boldsymbol{H}}_t^{(n)}$を推定する．

$$\log \hat{\boldsymbol{H}}_t^{(n)} = \log \boldsymbol{X}_t - \log \hat{\boldsymbol{S}}_t^{(n)}$$

Step. 3　平均歪みを$\tilde{\boldsymbol{H}}^{(n)}$推定する．

$$\log \tilde{\boldsymbol{H}}^{(n)} = \frac{1}{T} \sum_{i=1}^{T} \log \hat{\boldsymbol{H}}_t^{(n)}$$

Step. 4　$\tilde{\boldsymbol{H}}^{(n)}$を用いて，クリーン音声$\tilde{\boldsymbol{S}}_t^{(n)}$を推定する．

$$\log \tilde{\boldsymbol{S}}_t^{(n)} = \log \boldsymbol{X}_t - \log \tilde{\boldsymbol{H}}^{(n)}$$

Step. 5　前回の歪み$\tilde{\boldsymbol{H}}^{(n-1)}$と新たに推定された歪み$\tilde{\boldsymbol{H}}^{(n)}$との差を計算する．

$$D = \left(\log \tilde{\boldsymbol{H}}^{(n)} - \log \tilde{\boldsymbol{H}}^{(n-1)}\right)' \times \left(\log \tilde{\boldsymbol{H}}^{(n)} - \log \tilde{\boldsymbol{H}}^{(n-1)}\right)$$

Step. 6　もし，$D < \epsilon$（ϵは小さな正の数）であれば，この繰り返し推定は終了し，Step. 7 へ行く．そうでなければ，$n = n+1$として，Step. 2 へ行く．

Step. 7　$\log \tilde{\boldsymbol{H}}^{(n)}$と最も近いテンプレートが示す位置を推定距離とする．

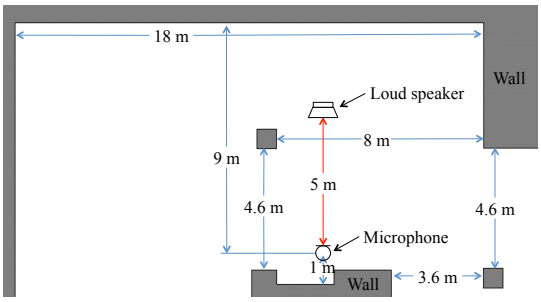

図 1　収録環境

4．研究成果

(1) 評価実験内容

評価実験として，下記の 3 種類を行った．

①音声区間検出あり／なしでの距離認識

②各距離，各コードブックサイズに対する距離認識率

③遠距離別での距離認識率

④繰り返し歪み推定手法の評価

(2) 実験条件

図 1 に収録した環境の見取り図を示す．大学内のロビーの一部で，部屋ではなく，開かれた空間になっている．音声データの再生録音でデータを作るとよいが，収録に時間がかかり、環境が変動する可能性があるため，ここでは，インパルス応答を推定，クリーン音声に畳み込むことで評価データを作成した．インパルス応答の推定には TSP 信号を用いた．なお，実環境であるが，比較的静かな時間帯で収録した．マイクロホンからの位置 0.20, 1, 2, 3, 4, 5 m において，TSP 信号を再生し，録音した．各位置で 4 回，TSP 信号を収録した．1 つのインパルス応答を評価データ用とし，残りをテンプレートとして用いた．

クリーン音声の VQ コードブック作成用や評価データとして日本音響学会新聞記事読み上げ音声コーパス(JNAS)を用いた．学習データ，評価データは音声認識に使われるものと同じ発話を用いた．音声データは 16 kHz サンプリング周波数であり，フレーム長は 25 ms，フレーム周期は 5 ms とした．特徴量としては，256 次元対数スペクトルを用いた．VQ コードブックの学習データとして，31,617 発話を用いた．評価データは IPA-98-TestSet で，23 名分の 100 発話を用いた．その評価データに各位置でのインパルス応答を畳込んで，各位置での距離推定用の評価データとした．VQ コードブックは k-means 法を用いて作成し，サイズを 128, 256, 512, 1024, 2048, 4096 のものを用意し，評価した．

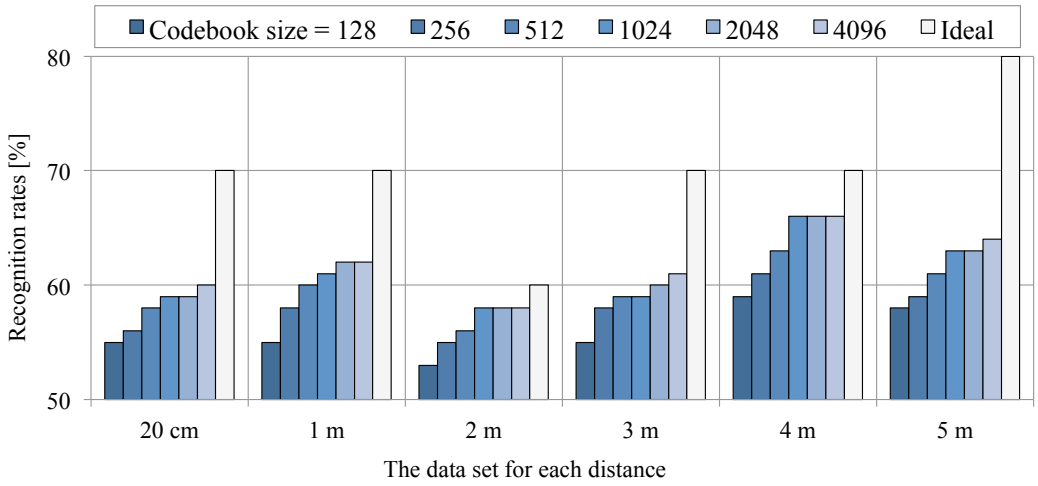

図 3　各距離，各コードブックサイズに対する距離認識率

(3) 評価実験

① 音声区間検出あり／なしでの距離認識

音声区間検出あり／なしでの評価を行った．距離別のテンプレートは 20 cm と 5 m のみを用い，評価データもその 2 種類のみを評価した．コードブックサイズは 256 とした．

図 2 に評価結果を示す．音声区間検出を用いると，位置 20 cm で 8%，5 m では 6%認識率が上昇した．音声区間のみを用いると精度が高いことが分かる．

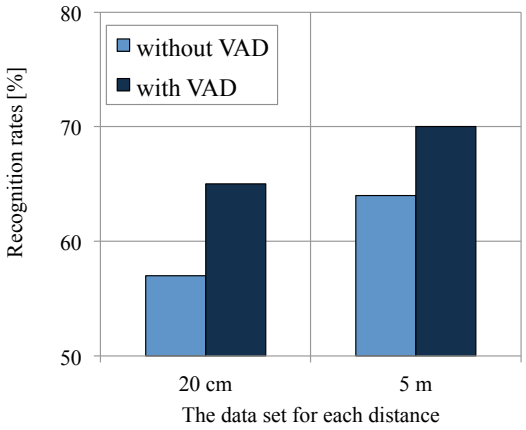

図 2　音声区間検出あり／なしでの評価

② 各距離，各コードブックサイズに対する距離認識率

実験条件に示した各距離とコードブックサイズ 128，256，512，1024，2048，4096 での評価を行った．図 3 に各距離，各コードブックサイズに対する距離認識率を示す．“Ideal”は理想的な場合の結果である．本提案手法での精度は 55～66%であった．コードブックサイズが大きいと精度は高くなった．照合させる音声のパターンが多く含まれるためである．4 m で精度が高いのは付近にあるは柱での反射が特徴的な特性を与えているためと考えられる．

③ 遠近距離別での距離認識率

対話システムでは厳密に位置を推定する必要は必ずしもない．ここでは近距離を 20 cm，1 m とし，遠距離を 2～5 m として 2 クラス識別問題として評価した．図 4 に結果を示す．近距離では 67～75%，遠距離では 69～80%の精度が得られた．

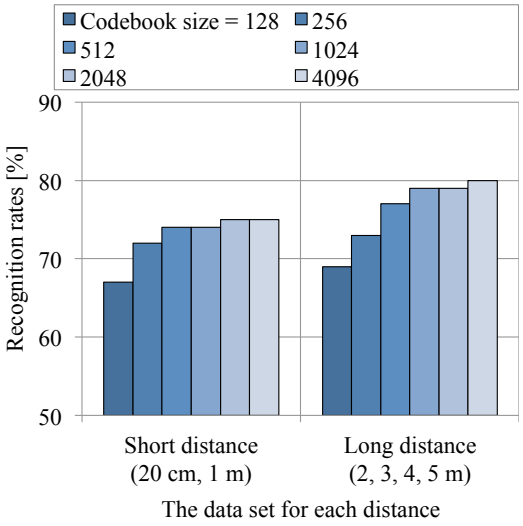

図 4　遠近距離別での距離認識率

④ 繰り返し歪み推定手法の評価

繰り返し歪み推定なし／ありでの距離認識実験を行った．図 5 に代表的な結果としてコードブックサイズ 4096 での距離認識率を示す．提案法により，どのコードブックサイズでも 1～3%の精度向上が見られた．特に，コードブックサイズ 4096 では，60～69%の精度が得られた．

また，遠近距離別での距離認識率を図 6 に示す．遠近距離，それぞれに対し，繰り返し推定なし／ありをコードブックサイズ 1024，2048，4096 に対する結果を示す．どの場合にも繰り返し推定により 2～3%の精度向上が見られた．最終的に，認識精度はコードブックサイズ 4096 のとき，近距離で 77%，遠距離で 82%が得られた．

より多くの特徴量や手法を検討し，精度向上を検討していく．

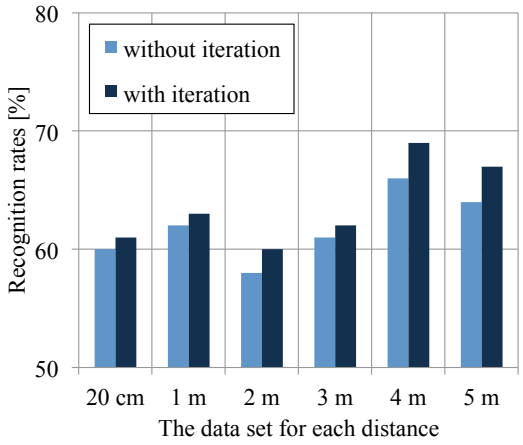

図 5　繰り返し推定での距離認識率（コードブックサイズ 4096）

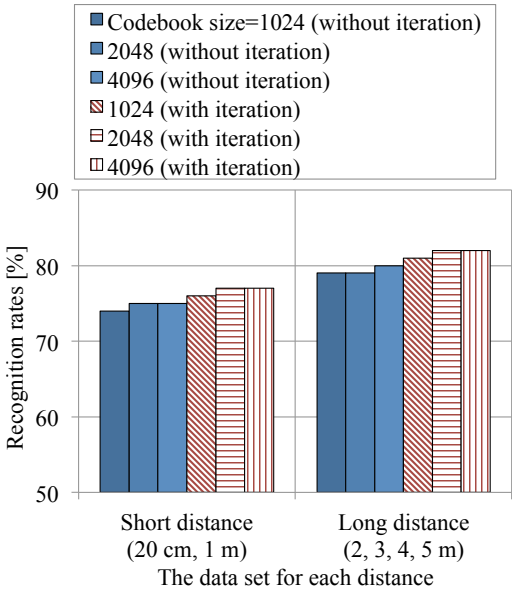

図 6　繰り返し推定での遠近距離別評価

５．主な発表論文等
（研究代表者、研究分担者及び連携研究者には下線）

〔学会発表〕（計 2 件）
① 李津，實廣貴敏，武田一哉，“単一マイクロホンによる音響モデルを用いた発話者までの距離推定”，日本音響学会講演論文集，2013 年 3 月 13 日，東京工科大

② 李津，實廣貴敏，“音響モデルを用いた発話者までの距離推定”，平成 24 年度 電気系学会東海支部連合大会，2012 年 9 月 25 日，豊橋技科大

〔その他〕
ホームページ等
愛知工科大学：實廣研究室：研究外部資金による研究
http://www1.aut.ac.jp/~jtlab/AUT_JTLAB/yan_jiu_zi_jin.html

６．研究組織
(1)研究代表者
實廣　貴敏（JITSUHIRO, Takatoshi）
愛知工科大学・工学部・准教授
研究者番号：　６０３９４９９６

(2)連携研究者
武田　一哉（TAKEDA, Kazuya）
名古屋大学・大学院情報科学研究科・教授
研究者番号：　２０２７３２９５