解 説 音声処理技術とその応用

5. 認識技術の進展

Advanced Topics on Speech Recognition Technology by Satoshi NAKAMURA and Kiyohiro SHIKANO (Graduate School of Information Science, Nara Institute of Science and Technology).

中村 哲[1] 鹿野清宏[1]

1 奈良先端科学技術大学院大学情報科学研究科

1. はじめに

音声認識技術は，統計的音響モデルや統計的言語モデル，およびそれらを支える大規模な音声，テキストデータベースにより，近年，著しい進歩を遂げている．とくにこの10年の進歩は目覚しく，最近では，パーソナルコンピュータにバンドルソフトウェアとして装備されているものもあり，一般の消費者やシステム開発者の手元にも行き渡る技術となりつつある．これらは，現在の音声認識というものの技術レベルを知らせること，実際のフィールドから問題点を吸い上げてさらに改良するための調査に利用できるという意味で，大変重要な役割を担っている．これらのフィードバックの報告については，別報を待つとして，実際，現在の音声認識技術にはまだまだ多くの克服すべき課題が残されている．とくに，実環境での利用における大きな問題として，話者の問題，雑音の問題，通信回線の特性の問題，部屋の残響の問題がある．このうち，本稿では，環境が原因となる問題として，雑音の問題，通信回線の特性の問題，部屋の残響の問題についてその性質と現在研究されている方法を解説する．また，これらの発展として，マイクロホンから離れて発声した音声の認識技術であるハンズフリー音声認識について述べる．さらに，音声に加えて画像などを利用するマルチモーダルの方向の研究として，音声認識と自動リップリーディングの統合による方法についても触れる．

2. 実環境への適応

2.1 実環境の音声

実環境において音声認識性能を劣化させる要因としては，部屋や場所などの音場に起因する要因とその音場内に存在する雑音源に起因する要因がある．雑音源からの信号は，マイクロホンで受音する際に音声信号に独立に伝搬し加算的に受音されるため，加算性の歪みと呼ばれる☆．雑音の種類は多種多様であり，列挙することは難しい．端的にいうと，聞きたい音以外はすべて雑音といえる．たとえば，計算機雑音のように定常で比較的扱いやすい雑音から，非定常な雑音，他人の声など対処するのが困難な雑音まで種々存在する．雑音の大きさについては，雑音のレベル，または，信号と雑音のパワーの比であるSNR(Signal to Noise Ratio)を用いて表すのが一般的である．たとえば，自動車のダッシュボードに設置したマイクロホンからドライバの音声を収録した場合，90km/hの速度でのSNRは，－5dB以下となる．

これに対し，部屋や場所などの音場に起因する要因は，一種のフィルタとして作用し音声に歪みを与える．この歪みは，線形システムの場合，時間領域では要因を表すインパルスレスポンスと原音声との畳み込みで表され，スペクトル領域では乗算で表されるため，乗法性の歪みと呼ばれる．実際の音場では，壁の反射など種々の乗法性の歪みが存在し，さらに，発話者の位置の変化により発話者から受音点までの音響的な伝達関数が変化し歪みの性質が変わる．また，電話回線などを経由した音声を対象にする場合にも，回線の伝達特性が音声を歪ませ，認識率が劣化する．使用するマイクロホンが異なると，これも大きな乗法性の歪みとなる．乗法性の歪みについては，畳み込みにおけるインパルスレスポンスで規定することができる．とくに，残響時間☆2や直接音反射音比

☆ 雑音がある環境で実際に発声を行うと，自分の発声を聞き取りやすくするための発声変形(Lombard Effect)が生じるが，本稿ではこの影響については論じない．

☆2 インパルスレスポンスのパワーが60dB減衰する時間を残響時間(T_{60})とする．

が，乗法性の歪みの影響の尺度としてよく利用される．音声信号処理では，通常短区間分析に基づいて特徴抽出を行うので，残響時間が十分短かければ窓内での周波数特性の補正を行うことで対処でき，マイクロホンや電話回線の違いはこの種のものと考えられている．

このようなことを考慮すると，実環境のモデルとして，図-1 に示すモデルが考えられる．このモデルによると観測信号は，次の式で求められることになる．ここで，＋は加法性要因の和を，⊗は乗法性要因の畳み込みを表している．

$$O(t) = h_c(t) \otimes (S(t) \otimes h_1(t) + N(t) \otimes h_2(t)) \quad (1)$$

これらの要因による劣化を防ぐために，今までに種々の試みが行われてきた[1),2)]．音声入力部ではマイクロホンアレーなどによる雑音処理，分析部では雑音スペクトル減算法などによる雑音処理，頑健な音声区間判定法として連続照合や雑音モデルを付加した形での照合による雑音中の音声区間検出，雑音に強いスペクトル距離尺度，雑音混じりの音声のモデルの合成と適応方式などである．とくに，前処理としての観点から，雑音の抑圧のための雑音スペクトル減算法に加え，乗法性要因の正規化のためにケプストラムの長時間平均値を乗法性要因として差し引くケプストラム平均減算法が処理が簡単で効果的であるため多用されている．これに対し，音声認識におけるモデルを直接操作するモデル適応化の観点から，音声のモデルにこれらの要因を直接作用させ，精密に観測信号のモデルをつくり出すアプローチも提案されている[3),4)]．

2.2 前処理による加法性・乗法性要因の正規化

加法性・乗法性要因に前処理として対処する方法として，雑音スペクトル減算法(SS: Spectral Subtraction)と長時間ケプストラム減算法(CMS: Cepstral Mean Subtraction)がある．雑音スペクトル減算法では雑音が定常であることを利用して，非音声区間の信号から雑音の特徴量を推定しておき，雑音混じりの音声の特徴量から雑音を取り除く．まず，サンプリングされた時刻 k の雑音混じりの信号を $o(k)$，真の音声信号を $s(k)$，雑音信号を $n(k)$ とすると，観測される雑音混じりの信号 $o(k)$ は，次のように表される．

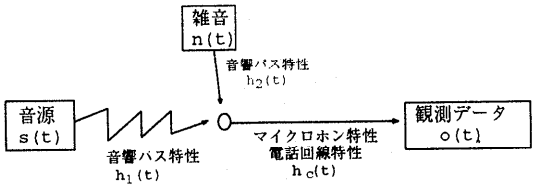


図-1 実環境のモデル

$$o(k) = s(k) + n(k) \quad (2)$$

ここで，窓位置 m で短時間分析を行い両辺のフーリエ変換をとると次の式が得られる．

$$S_o(\omega; m) = S_s(\omega; m) + S_n(\omega; m) \quad (3)$$

$S_o(\omega; m)$，$S_s(\omega; m) + S_n(\omega; m)$ は，周波数 ω の複素スペクトルを表す．音声の認識においては，人間の聴覚が位相に敏感でないことからパワースペクトルに注目すればよく，$s(k)$ と $n(k)$ が無相関とすると，真の信号のパワースペクトル $|S_s(\omega; m)|^2$ は，

$$|S_s(\omega; m)|^2 = |S_o(\omega; m)|^2 - |S_n(\omega; m)|^2 \quad (4)$$

により与えられる．ここで，雑音信号を定常と仮定し非音声区間において雑音のパワースペクトル $|\hat{S}_n(\omega; m)|^2$ を推定しておけば，

$$|\hat{S}_s(\omega; m)|^2 = |S_o(\omega; m)|^2 - |\hat{S}_n(\omega; m)|^2 \quad (5)$$

のように真の信号のパワースペクトルの推定値を求めることができる．この方法では，1)音声，非音声をどのように区別するか，2)非定常の雑音にどのように対処するか，3)減算することにより生じる負のスペクトルをどうするかが問題となる．この問題に対して，音声，非音声の区別を行った後，SNR に応じた雑音部の補正を行って雑音混じりの音声から減算する方法や，減算により推定された音声のスペクトルが雑音信号の平均値より小さい時はその値で置き換えることで改善を行う方法，逆に連続的に減算を行う連続スペクトル減算法などが提案されている．

乗法性要因についても類似の方法が長時間ケプストラム減算法として知られている．まず，時刻 k の乗法性の歪みを受けた信号を $o(k)$，真の音声信号を $s(k)$，乗法性の歪みを $h(k)$ とすると，観測される歪んだ信号 $o(k)$ は，次のように表される．

$$o(k) = s(k) \otimes h(k) \quad (6)$$

ここで，窓の位置を m で表した短時間分析によ

る両辺のフーリエ変換をとると次の複素スペクトルが得られる．

$$S_o(\omega; m) = S_s(\omega; m) \cdot S_h(\omega; m) \quad (7)$$

パワースペクトルを求めた後，さらに対数をとると対数パワースペクトルが求められる．

$$\log |S_o(\omega; m)|^2$$
$$= \log |S_s(\omega; m)|^2 + \log |S_h(\omega; m)|^2$$

さらに，両辺を逆フーリエ変換すればケプストラムが求められる．

$$C_o(t; m) = C_s(t; m) + C_h(t; m) \quad (8)$$

ここで，$C_o(t; m)$, $C_s(t; m)$, $C_h(t; m)$はそれぞれmフレームにおける観測信号，信の信号，乗法性の歪み成分の短区間ケプストラムである．ところで，$C_h(t; m)$は一定なので長時間平均により推定値$\hat{C}_h(t; m)$を求めることができる．これを用いると次のように真の信号のケプストラム推定値が求まる．

$$\hat{C}_s(t; m) = C_o(t; m) - \hat{C}_h(t; m) \quad (9)$$

問題は音声の長時間ケプストラムまで減算される点であり，乗法性の歪みのない音声にこの処理を施すと性能が劣化することなどが知られている．

2.3 HMMにおけるモデル適応化法

モデル適応化のアプローチは，あらかじめ学習済みの音声認識用モデルのパラメータを実際に認識しようとする環境での音声に適応化し，観測される音声に適合するモデルを作るアプローチである．これまで，観測データと元のモデルのパラメータの差ベクトルを利用する方法，学習用観測データへの尤度が大きくなるように最尤推定法や事後確率最大化法により推定する方法，雑音のモデルを学習後，元のモデルとの合成を行って観測データに適合するモデルを作るHMM合成法などが提案されている．モデル適応化による方法は，前処理法に比べると複雑であるが，モデルパラメータの分散も適応化できるので性能的には前処理法に勝る可能性がある．本節では，HMM分解・合成法について述べる．

通常の雑音減算法では1つのマイクロホンで音声を収録し，音声に対し雑音の加法性，定常性を仮定して，音声区間から無音区間で推定した雑音を減算する．ところが，実環境の雑音にはこのような定常性の仮定が成り立たない．そこで，複数の状態をもつ隠れマルコフモデルで雑音を複数の定常雑音の連結としてモデル化し，音声と雑音のHMMの適当な状態の出力の加算が観測データに近くなるように状態遷移を選択していくモデル分解法が提案された[5]．音声認識で有効な領域であるケプストラム領域では雑音と信号の単純加算が成り立たないという問題があり，この方法では対数演算を最大値演算により近似する方法($\max(x, y) \approx \log(x + y)$)を用いている．この手法は，機関銃雑音のような非定常な雑音に対処できるという点において非常に魅力的な方法といえるが，最大値近似のもたらす誤差が問題になる．これに対しあらかじめ音声HMMと雑音HMMのすべての状態の組合せを求めておく方法がモデル合成法として提案された．この方法では，指数変換操作後加算を行い，再び対数操作を行う方法[6),7)]により複数のHMMの直積を可能としている．ここでは，HMMのモデル合成法について述べる．まず，加法性雑音に対してHMMモデル合成を用いて対処する方法について述べる[7)]．HMM合成

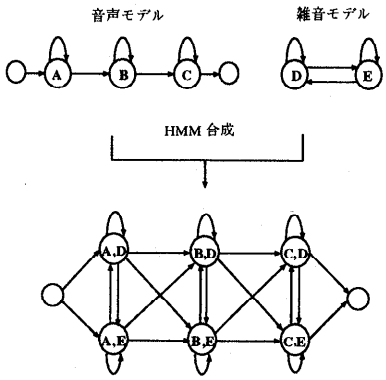


図-2 合成HMMの構造

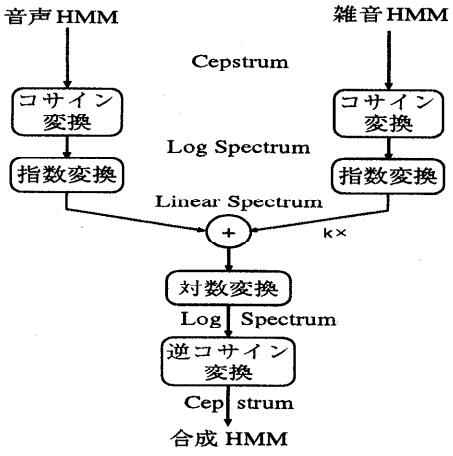


図-3 出力確率の合成法

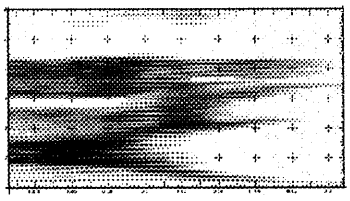


図-4 雑音，残響を含まないクリーン音声のスペクトログラム/ai/

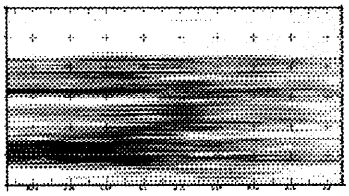


図-5 残響を含んだ音声のスペクトログラム/ai/: T_{60} = 600msec

とは，雑音を含まないクリーン音声を用いて学習したHMMと雑音のみを用いて学習したHMMを合成し，雑音重畳音声に対するHMMを作成する方法である．それぞれのHMMの状態は確率分布を有するので，その構造は図-2に示すようにそれぞれのHMMの直積で定義される．遷移確率は対応する遷移確率の積で求められる．出力確率については，それぞれのHMMを連続分布型HMMとし，与えられた音声と雑音の出力確率分布を単一正規分布 C_s: $\{\mu^s, \Sigma^s\}$, C_n: $\{\mu^n, \Sigma^n\}$とすると，合成HMMの出力確率分布はそれぞれの状態で結合され，合成出力確率分布 C_o: $\{\hat{\mu}, \hat{\Sigma}\}$が計算される．正規分布は再生性が保証されるため確率変数の和は分布の畳み込みとなり，合成された確率密度関数は各密度関数の平均，分散の和によって C_o: $\{\mu^s + \mu^n, \Sigma^s + \Sigma^n\}$のように与えられる．さらに，音声と雑音の加法性が成立するのはパワースペクトルの領域であるが，音声認識では特徴量がケプストラムで表現されているため，これらにコサイン変換および指数変換を行って，線形パワースペクトル領域に変換し，この結合(出力確率の計算)を行う必要がある．

$$S_o(w; m)$$
$$= \exp(\mathcal{F}(C_s(t; m))) + k \cdot \exp(\mathcal{F}(C_n(t; m)))$$

ここで k はSNRに応じた雑音重畳音声とのレベル調整係数である．この様子を図-3に示す．

3. ハンズフリーの音声認識

音声の特徴の1つは，多少の距離をおいて，手が塞がった状況でもコミュニケーションできることである．音声の中には，発声内容，話者などの情報のほかに，発声方向／距離などの空間における情報も含まれている．人間は，この情報を音声と雑音の分離や複数の話者の分離に利用している．このような情報を利用した実環境においてマイクロホンを意識せずに発話が可能なハンズフリー音声認識について簡単に述べる．音声認識の研究では，これまで1チャンネルの処理を行っていたので，発声者と雑音源の空間的な位相の異なりをあまり利用せず，雑音の定常性を仮定した方法が試みられてきた．しかし，雑音が指向性の場合，指向性を有するマイクロホンを用いることによりこれらの情報を分離できる．とくに，複数のマイクロホン素子を有し，信号処理により指向性を制御できるマイクロホンアレーを用いれば，任意の方向や位置からの発声された音声のみを取り出すことが可能になる．次の問題は，部屋の残響の問題である．一般にマイクロホンからはなれて部屋の中で発声すると，部屋の反射特性，遅延などにより音源からマイクロホンへ到達する間に音声が歪んでしまう．図-4，図-5にクリーンな音声と残響の長い部屋で発声した音声のスペクトログラムを示す．見てのとおり，元のスペクトログラムからはずいぶん違い，認識性能も大きく劣化する．一般にこの残響は，線形システムを仮定して音源からマイクロホンまでの音響パスのインパルスレスポンスで与えられる．残響と同種の乗法性の歪みを与えるものにマイクロホン，ハンドセット，電話回線の違いがある．しかし，これらの影響のインパルスレスポンスは部屋の残響などに比較して短く，短区間フレーム内で処理可能な場合が多い．これに対し，部屋の残響は，通常短区間フレーム分析の窓より長くなることもあり，扱いが難しい．現在までの研究としては，逆フィルタを構成して時間領域で処理するものがほとんどである．一般に部屋のインパルスレスポンスが最小位相にならないために逆フィルタが構成できないことが大きな問題であり，全域通過フィルタと最小位相フィルタに分けたり，多入力多出力系で逆フィルタを構成したり，ニューラルネットで逆フィルタを推定する研究が行われている．最近では，マイクロホンアレーの超指向性により直接波のみ

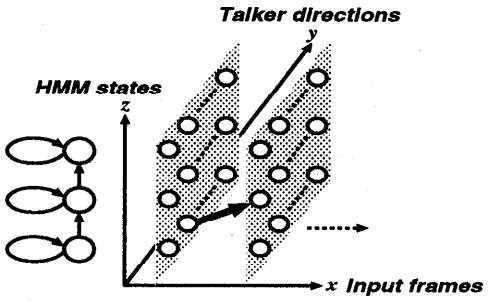


図-6 3次元トレリス空間のViterbi検索

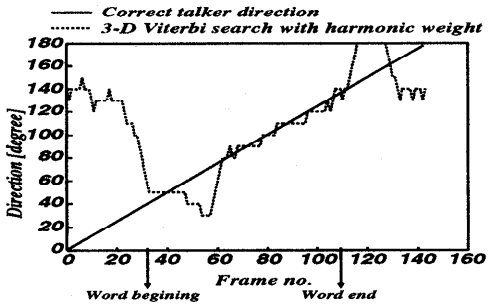


図-7 3次元Viterbi検索を用いた移動話者の方向追跡の例 (SNR: 20dB)

を取り出す方法，聴覚特性を模倣した帯域フィルタごとにバイアスを取り除く方法(RASTA)，残響を受けた観測信号をモデル領域でつくり出して認識を行うアプローチが音声認識の分野でも始められている[8]．本稿では，マイクロホンアレーを用いる方法とHMMのモデル合成法に基づく方法による試みについてさらに以下の節で述べ，実環境におけるハンズフリー音声認識の可能性や問題点について明らかにしていきたい．

3.1 マイクロホンアレーの利用

実環境に存在する雑音や残響に前処理レベルで対処する方法としては，マイクロホンアレーを用いる方法がある[9]．先に述べたように，指向性の雑音のみならず無指向性の雑音についても，超指向性をもつマイクロホンを用いれば抑圧することができ，話者の信号のみをうまく取り出すことができる．また，超指向性により直接波のみを取り出すことができれば残響の影響も低減することができる．さらに，話者の方向を同定することができれば移動する話者に追従することができる．マイクロホンアレーを用いれば，この話者の超指向性を任意の方向に形成したり，方向同定を行う機能を実現できる．上述したようにマイクロホンアレーの機能としては，音源の方向，位置同定(Source Localization)と超指向性の形成(Beam Forming)がある．音源の方向，位置同定は，一般には低周波数域に帯域制限を行い，4～8個程度のマイクロホンを広い間隔で配置し空間分解能をあげる方法がよく用いられる．音源の位置，角度の計算は，マイクロホンに入力される信号が同相となるマイクロホン信号間の遅延を正確に測定することで行われる．これまで，相互相関による方法，Cross-power Spectrum Phaseによる方法などが提案されている．しかし，実際には，実環境ではかなりSNRが低く，負になることもあるため音声の継続性や音源の周期性により生じる調波性などの特徴を利用する必要がある．一方，超指向性の形成には，基本的に多数のマイクロホンが必要となり，その間隔も対象周波数の半波長となるように設計する必要がある．500チャンネル以上を用いた研究や，音声認識性能を最適化するマイクロホン配置の研究が行われている[8]．

3.2 時間，方向，HMM状態の3次元トレリス空間探索法

マイクロホンアレーは，これまでも音声認識の前処理としてその有効性が検討されてきたが，従来のシステムでは音源の同定，超指向性の形成，音声認識がまったく独立に順番に処理されていた．したがって，音源同定の誤りが，直接音声認識性能の劣化につながっていた．しかし，人間の知覚を考えてみるといろいろな音源がある場合，聞き終ってから，あの方向からの音は私に話かけている音声だったと理解することがある．このように，音源の同定にも音声の言語的知識を入れ，統合的に処理する必要がある．著者らは，比較的多くのチャンネルを有し空間分解能の高いマイクロホンアレーの指向性をいろいろな方向に向けることにより方向ごとの信号を取り出し，方向，時間，HMM状態の3次元からなる空間を尤度最大基準で探索し，最適音素列，移動方向系列の組合せを求めるViterbi探索法を提案している[10]．これにより，音声認識用の文法で記述された言語制約の下に尤度が最大となる音源方向，音素系列を求めることができる．また，この方法によれば，移動する話者の音声も認識することができ，さらに，N-best探索法への拡張により複数の音源を

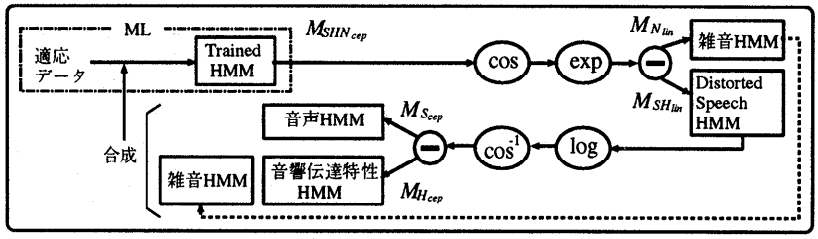


図-8 HMM 合成・分解によるパラメータ推定法

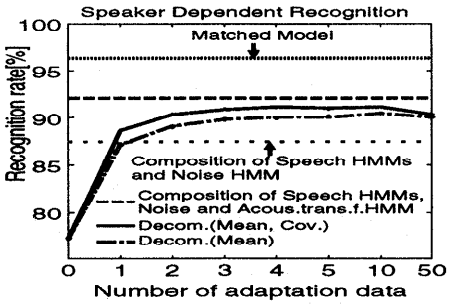


図-9 実環境下での音声に対する 500 単語認識

一度に認識する可能性もある．図-6 に 3 次元トレリスの例を示し，図-7 に移動する音源を認識した場合の例を示す．図-7 は，0 ～ 180 度まで移動しながら 1 単語発話した音声の方向同定結果である．単語区間では，正確に音源方向を同定していることがわかる．

3.3 モデル合成による方法

一般に，使用環境に近い環境で得られた音声を用いることができる場合，その音声を用いてモデルを学習することが最も効果的であるといえる．しかしながら，一般にモデルを学習するほどの大量の音声を使用環境で集めることは困難であり，そのためモデルの適応化というアプローチが必要となる．本節では，著者らが行っている部屋の伝達関数のモデル化と HMM 合成法による観測信号のモデルの合成法，推定法について述べる．このような環境を式(1)をさらに簡略化してモデル化してみると次のように書ける．

$$O(t) = S(t) \otimes h_1(t) + N(t) \qquad (10)$$

実際の部屋においてマイクロホンで音声を収録した場合，ここで，$h_1(t)$は，部屋の伝達関数で，場所により異なり，また，話者が移動するものとして時間の関数となる．$N(t)$は，雑音である．情報源を HMM でモデル化するという手法に基づいて，無音区間の観測信号を用い雑音源を HMM でモデル化する．$h_1(t)$についても，正規分布をもつ HMM でモデル化ができる．あらかじめ部屋がわかっているとすれば，部屋の代表的な場所からマイクロホンへのインパルスレスポンスを測定し平均値とし，さらにその周りもカバーできるように分散を決める．部屋全体をモデル化するためには，この代表的な場所(分布)を状態に割り当てた Ergodic HMM ☆3 を作成しておけばよい．このように，音声，雑音，伝達関数の HMM が作成できるので，それぞれ加法性の成り立つ領域に変換して HMM の合成を行い最終的にケプストラム領域でモデル化する．実験により，この方法を用いれば，粗い近似ながらも，場所のモデル化ができることが示された[11]．実際には，部屋が変わるごとにインパルスレスポンスの測定をすることは実用的でないため，最尤推定を基本とした HMM 分解合成法を用いて学習により推定する方法を提案している[12]．図-8 に，処理のブロック図を示す．適応学習用データを用いて，環境に適合した HMM を学習した後，スペクトル領域に変換し，HMM を尤度最大基準で無音区間で学習した雑音 HMM と残響歪みを受けた音声 HMM に分解する．さらにケプストラム領域に変換し，再度尤度最大基準で音声 HMM と伝達特性 HMM に分解する．図-9 に残響時間 180msec，SNR が約 15dB の実験室で行った特定話者の実験結果を示す．この方法によれば，非常に少ない学習データ(2 ～ 4words)で，最尤推定法により伝達関数のモデルパラメータ推定ができることが示されている．しかしながら，その環境における音声で学習した Matched Model の結

☆3 すべての状態が初期状態，最終状態になり，すべての状態間の遷移が可能な HMM を指す．

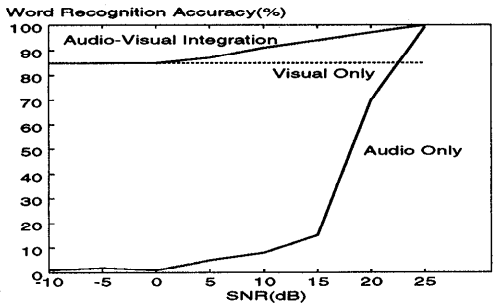


図-10 音声・唇画像の統合による単語認識

果に及ばない．これは，現在のところ短区間フレーム分析内の処理が基本となっているためで，短区間分析フレームを超えた影響への対処が今後の問題である．

4. 計算機によるリップリーディングと音声認識の統合

これまで，音響信号のみを用いたハンズフリー音声認識の試みについて述べたが，実際には，マルチモーダルな情報処理が利用できる．たとえば，顔や唇，ジェスチャーなどの画像情報である．このような音響信号と画像信号の統合は，人間においても存在することが示されている．特筆すべきは，画像の情報が音響的な雑音の影響をまったく受けないことである．ここでは，計算機による自動リップリーディングについて簡単に触れる．最近，自動リップリーディングの研究もHMMやニューラルネットワークの利用により大きな進歩がみられている[13]．著者らの研究によれば，分布に制約をもたせた音素単位のHMM(Tied-mixture HMM)を用いれば，唇の画像情報のみで特定話者の日本語100単語認識に対し85.0％の単語認識の性能が得られることがわかっている．また，音声と画像を統合すれば，それぞれ単一の情報のみの結果より高い性能が得られることが示された[14]．この結果を図-10に示す．統合方法としては，それぞれのHMMの結果を統合する結果統合による方法を用いている．しかし，実際には，カメラから離れて移動する状況では，画像と音声の情報との協調による位置同定と唇部分の切り出し，さらに画像の大きさ，照明，角度などの正規化がきわめて重要な技術になると考えられる．

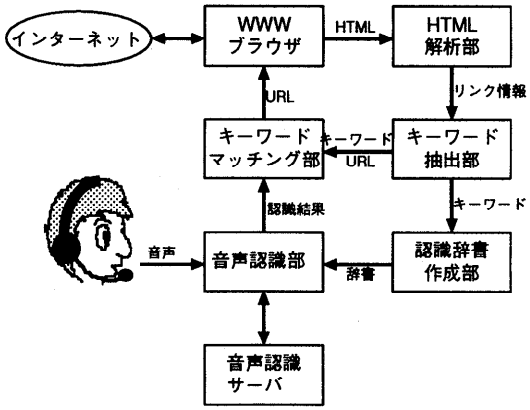


図-11 システム構成図

5. インターネットにおける音声認識の利用

最後に，音声認識の応用の1つとしてインターネッドにおける音声認識の応用について紹介する．インターネット上には，膨大な情報が存在する．これらの情報の検索のために検索エンジンなどが開発されているが，なかなか的確に必要な情報をみつけることができない．自然言語の音声入力，音声対話ができれば，格段に自然性，効率が改善できると考えられる．この方向の研究として，最近，英語のWWWブラウザと音声を結びつけたシステムが提案されている[15]．また，日本語によるシステムも研究されはじめている[16]．しかし，ドメインが限定されていたり，文法をあらかじめ書いておく必要があったり，マウスやキーボードを単に音声に置き換えただけであったりと，ユーザインタフェースの観点からはいろいろな問題が残されていた．本章では，著者らが開発しているページのリンク項目を動的に取得し，日本語のキーワードを音声認識することで無駄なマウスクリックを省き，効率的なネットサーフィンを実現するシステムについて述べる[17]．

このシステムでは，多様なリンク項目を簡潔に音声認識させるための「キーワードによるリンク項目の認識」，現在表示されていないページの情報を動的に取得し，音声によって効率的なネットサーフィンを実現可能とする「ページの先読みによる動的な認識語彙の獲得」により構成されている．

ページ中のリンク項目を音声認識する場合，リ

ンク項目をそのまま読みに直し，認識辞書としてことが考えられるが，リンク項目が長くなったり，読みにくい字が出てきた場合には逆に不便である．そこで，リンク項目中からキーワードを抜き出し，キーワード集合を発声することによりリンク項目が認識できるようにする．本システムでは形態素解析システム「茶筌」[18]を用い，リンク情報中の名詞を抽出し，音素列に変換したものを，キーワード集合として並列に認識できるようにしている．また，現在表示されているページのリンク項目を音声認識するだけでは，1回のマウスクリックが1度の音声発話に置き換えられただけで，効率がよくなるとはいえない．本システムでは，通常目的のページに到達するために複数回のマウスクリックを必要とするところを，1度の音声発話で置き換える．そのために，あらかじめある程度の深さまでのページを先読みし，キーワードを取得して認識辞書を作成している．**図-11** にシステム構成を示す．実際にこのシステムを用いたネットサーフィンの実験を行い，ページ到達速度，実行手数の観点から有効性を確認している．

6. おわりに

実環境において音声認識を利用する際に問題となる点，さらに現在研究されているそれらの要因への対処方法について述べた．また，ハンズフリー，マルチモダリティへの展開，インターネットのネットサーフィンへの利用について述べた．現在，音声認識の研究は，大語彙連続音声認識技術と実環境における認識技術の2つの大きな方向で急速に進んでいる．また，一方でインターネットやカーナビゲーションなどへの応用展開も急速に進んでいる．このような状況は，音声認識技術の実用化にとってきわめて絶好の機会といえる．今度こそ，研究室のレベルを出て，実世界で有用な技術となることを期待してやまない．

参考文献

1) Acero, A.: Acoustical and Environmental Robustness in Automatic Speech Recognition, Kluwer Academic Publishers, Boston, MA (1993).
2) Gong, Y.: Speech Recognition in Noisy Environments: A Survey, Speech Communication, 16, pp.261-291 (1995).
3) Furui, S.: Recent Adavances in Robust Speech Recognition, ESCA-NATO Workshop on Robust Speech Recognition for Unknown Communication Channels, pp.11-20(Apr. 1997).
4) Stern, R. M. et al.: Compensation for Environmental Degradation in Automatic Speech Recognition, ESCA-NATO Workshop on Robust Speech Recognition for Unknown Communication Channels, pp.33-42(Apr. 1997).
5) Varga, A. P. and Moore, R. K.: Hidden Markov Model Decomposition of Speech and Noise, Proc. IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP90, pp.845-848 (1990).
6) Gales, M. J. F. and Young, S.: An Improved Approach to the Hidden Markov Model Decomposition of Speech and Noise, Proc. IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP92, pp.233-236 (1992).
7) Martin, F., Shikano, K. and Okabe, Y.: Recognition of Noisy Speech by Composition of Hidden Markov Models, 電子情報通信学会技術報告, SP92-96, pp.9-16 (1992).
8) Omologo, M.: On The Future Trends of Hands-Free ASR: Variabilities in The Environmental Conditions and in The Acoustic Transduction, ESCA-NATO Workshop on Robust Speech Recognition for Unknown Communication Channels, pp.67-73 (Apr. 1997).
9) Flanagan, J. L., Mammone, R. and Elko, G. W.: Autodirective Microphone System for Natural Communication with Speech Recognizers, 4th DARPA Workshop, pp.4.8-4.13 (Feb. 1991).
10) 山田武志，中村 哲，鹿野清宏：マイクロホンアレーによる3次元ビタビ探索に基づく移動話者の音声認識，電子情報通信学会音声研究会，SP97-22 (June 1997).
11) 滝口哲也，中村 哲，鹿野清宏：雑音と残響のある環境下でのHMM合成によるハンズフリー音声認識法，電子情報通信学会D-II，Vol.J79-D-II, No.12, pp.2047-2053 (1996).
12) 滝口哲也，中村 哲，Huo, Q.，鹿野清宏：HMM分解に基づいたモデル適応化法による雑音・残響下での音声認識，日本音響学会研究発表講演論文集，1-6-17 (Mar. 1997).
13) Stork, D. G. and Hennecke, M. E.: Speechreading by Humans and Machines, NATO ASI Series, Springer (1995).
14) 中村 哲，山本英里，永井 論，鹿野清宏：HMMを用いた音声と唇画像の統合による音声認識と唇画像生成，情報処理学会研究会，音声言語情報処理，pp.15-17 (Feb. 1997).
15) Hemphill, C. T., Thrift, P. R. and Linn, J. C.: Speech-Aware Multimedia, IEEE Multimedia, Vol.3, No.1, Spring (1996).
16) Kondo, K. and Hemphill, C. T.: Surfin' The World Wide Web with Japanese, Proc. IEEE

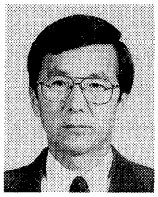
International Conference on Acoustics, Speech and Signal Processing, ICASSP97, pp.1151-1154 (1997).
17) 桂浦 誠, 中村 哲, 鹿野清宏：キーワードを用いた音声によるネットサーフィン, 日本音響学会研究発表講演論文集 (Sep. 1997).
18) http://cactus.aist-nara.ac.jp/lab/nlt/chasen.html

(平成9年9月3日受付)

中村 哲 (正会員)

昭和33年生. 昭和56年京都工芸繊維大学工芸学部電子工学科卒業. 昭和56年～平成6年シャープ(株)中央研究所および情報技術研究所に勤務. 昭和61年～平成元年ATR自動翻訳電話研究所に出向. 平成6年奈良先端科学技術大学院大学情報科学研究科助教授. 平成8年3月～8月 Rutgers University CAIP Center 客員教授. 音声情報処理, 音声対話システム, マイクロホンアレー, マルチモーダル信号処理の研究に従事. 京都大学博士 (工学). 平成4年日本音響学会粟屋学術奨励賞受賞. IEEE, 電子情報通信学会, 日本音響学会, 人工知能学会各会員.

e-mail:nakamura@is.aist-nara.ac.jp



鹿野 清宏 (正会員)

昭和22年生. 昭和45年名古屋大学工学部電気学科卒業. 昭和47年同大学院工学研究科修士課程修了. 同年電電公社武蔵野電気通信研究所入所. 昭和59年～昭和61年カーネギーメロン大客員研究員. 昭和61年～平成2年ATR自動翻訳電話研究所音声情報処理研究室長. 平成4年NTTヒューマンインタフェース研究所主席研究員. 平成6年奈良先端科学技術大学院大学情報科学研究科教授. 音情報処理学講座を担当, 主として音声・音情報処理の研究および研究指導に従事. 工学博士. 昭和50年電子情報通信学会米沢賞, 平成3年IEEE SP 1990 Senior Award, 平成6年日本音響学会技術開発賞受賞. IEEE, 日本音響学会各会員.