カクテルパーティ効果の工学的利用

眞溪歩
新領域創成科学研究科複雑理工学専攻

概要

カクテルパーティ問題の工学的解明では，これまで，逆フィルタ設計，学習理論，マイクロフォンアレイ計測・信号処理，独立成分分析など多方面から多大な努力が払われてきた．本研究では，これらの研究成果を横断的に統合し，現代的な問題設定下で人間の機能を模した予約語の認識を目指す．

ここでは，まず本研究の背景を述べ，次に関連する要素技術を整理し，最後に提案する手法の戦略について説明する．

1 背景

複数の人がそれぞれ自由に会話している環境において，人は特定の人の発話内容をある程度理解することができる．このような状況は，カクテルパーティ効果として，古くから研究的関心事として取り上げられてきた．

Cherry の 1953 年の聴覚心理実験では，ヘッドフォンによって両耳から異なる会話を聞かせ，片方の発話に注意するように指示し，聞かせ終わった後，注意を要求しなかった方の会話内容について質問をした．その結果は，注意しなかった発話内容はほとんどあるいはまったく理解されていなかった．その後，例外的に，被験者の名前の発話を非注意側に入れると一時的に注意が移動することも明らかになった．また，注意の研究は視覚など他の感覚モダリティにも波及し，注意の脳機能モデルやこれを検証するための巧みな実験課題が現れた．注意は人間に備わった機能であり，この機能を持った機械の開発が望まれている．

一方，カクテルパーティ効果は，工学的にも高い関心が持たれている．還元論にもとづく観察では，この特定発話抽出はいくつかのシステムによる実現と考えられる．すなわち，人間は，雑音，反響，興味のない人の発話などから興味のある人の発話をフィルタリングして抽出し，音声認識する多段処理が想像される．前者と後者は個々にも研究的関心がある．前者はさらに細かく分割しても，それぞれに関心がある．このため，カクテルパーティ効果の工学的実現という目標とは独立に，さまざまな要素技術が発展している．本研究では，このような要素技術をカクテルパーティ効果の工学的応用にフィードバックさせたい．

カクテルパーティ効果の工学的利用について，現代的な問題設定を考える．古くから，この工学的利用には，固定マイクロフォンアレイによる混合音声信号の多点同時を出発点にしている．しかし，現実的な場面を考えると，たとえば多くの人が携帯電話を持っていて，通常は利用していないものの移動マイクロフォンはいろいろなところに存在する．また，本研究の全体構想では，パーティ会場にロボットが存在することも想定されている．ロボットには移動マイクロフォンどころか移動する各種センサを取り付けたり，模擬信号を発生させたりすることもできる．このとき，ロボットの情報はリアルタイムに把握することができる．このようなことから，現代的な問題設定では，センサフュージョンも視野に入れていきたい．

2 要素技術の整理

本研究の要素技術を整理する．

2-1 残響除去

残響除去(dereverberation)は，音源からマイクロフォンアレイに至る室内の壁を介したマルチパスの影響で起こる残響を除去または軽減しようとする手法である．この種の研究の問題設定では，音源，マイクロフォンアレイ，部屋

の構造がわかっているとされる．しかし，現実的には，実験によって伝達関数は同定される．最も基礎となる考え方は，音源 1 個からマイクロフォン 1 個への離散系での伝達関数を $H(z)$として，その逆システム $1/H(z)$を設計しようというものである．このとき残響作用は時間遅れと波形歪みと考えられるので，$H(z)$は z の有理関数で近似して差し支えない．そこで，

$$H(z)=\frac{Q(z)}{P(z)}$$

とおく．このことから，逆システムが動的であるためには，$H(z)$が最小位相，すなわち $Q(z)$の零点がすべて単位円の内側に存在する必要がある．しかし，一般的な残響特性では最小位相特性は期待できない．理論的には，伝達関数 $H(z)$は，

$$H(z)=H_{\mathrm{MAX}}(z)H_{\mathrm{MIN}}(z)$$

のように最大位相部分 $H_{\mathrm{MAX}}(z)$と最小位相部分 $H_{\mathrm{MIN}}(z)$に一意に分解することができる．そこで，実際的に，この分解の推定と $H_{\mathrm{MAX}}(z)$の逆システムの近似について様々な方法が提案されている．また，この分野を離れると，近年小脳の学習モデルにこのような逆システムが提案されており，そこでもこの種の問題が議論されている．

マイクロフォンアレイを用いた最も単純な残響除去手法は同期加算である．同期加算では，音源から各マイクロフォンまでの遅延を補正し，その出力を加算する．音声は広帯域信号であるため，マイクロフォンの設置間隔は波長に対し適切でなければならない．このため，マイクロフォンはオクターブ配置される．この手法は，直達音声のパワーを増強することにのみ注意が払われており，残響音の抑制には何ら対策が講じられていない．

残響音の抑制には，分散最小規範を導入した適応信号処理が用いられる．各マイクロフォンの信号は FIR フィルタ処理された後，加算され出力となる．FIR フィルタの係数は，アレイ数×遅延タップ数存在し，これをベクトル $\mathbf{w}$ とおく．また，フィルタ係数が乗じられる直前の信号も同様にベクトル時系列として $\mathbf{x}[n]$とおく．ここで，分散最小規範は，

$$\mathbf{w}^{\mathrm{T}}\left\langle \mathbf{x}[n]\mathbf{x}^{\mathrm{T}}[n]\right\rangle \mathbf{w}\rightarrow \min.$$

で表される．ただし，自明解 $\mathbf{w}=\mathbf{0}$ に陥らないために，

$$\mathbf{C}^{\mathrm{T}}\mathbf{w}=\mathbf{f}$$

なる拘束条件を定める．ここで，$\mathbf{C}$ は FIR フィルタのタップ数次元の単位行列を横にアレイ数並べた行列であり，$\mathbf{f}$ はシステム全体として見たときの理想 FIR フィルタ係数である．一般的に，$\mathbf{f}$ には音声帯域のバンドパス特性が期待される．この手法は，理想的なフィルタ特性を拘束条件として，遅延分も含めた信号の分散を最小にする射影方向を求めている．遅延分も含めることが残響も考慮した信号の分散を最小にしていることになり，残響音抑制効果が期待される．つまり，遅延も考慮した分散は自己相関関数的である．実際に適用してみると，残響音が長く続くとき，つまり残響システムのインパルス応答が長いテールを持つ場合には，あまり効果を期待できない．

残響システムのインパルス応答がわかっているときには，便宜的にそのもの自体を FIR フィルタとするマッチドフィルタを用いて残響抑制を行うことができる．しかし，マッチドフィルタで構成されるシステムは，完全な逆システムとはならない．いま，音源から m 番目のマイクロフォンへの伝達関数を $H_m(z)$とすると，全零型の完全な逆システム $G_m(z)$は

$$\sum_{m} G_m(z)H_m(z)=1$$

で与えられる．いま，すべての m に対し，$H_m(z)$より $G_m(z)$の次数が高く，かつ $G_m(z)$が互いに共通な零点を持たない場合，この条件を満足する $G_m(z)$は一意に定まる．このように設計された $G_m(z)$はマッチドフィルタより残響抑制性能は高いが，音源がフォーカス点から離れたときの性能劣化は著しい．この違いは，マッチドフィルタが双対性を考慮していないのに対し，逆システムでは考慮していることに起因している．

残響除去は，いずれの方法も何らかのコスト関数の最小化・最大化問題で記述できる．

2-2 到来方向推定

音声の到来方向(DOA: Direction Of Arrival)推定は，前述の残響除去と同じような枠組みでとらえられるが，若干異なる性質に配慮する必要がある．残響除去では，音声はひとつであり，その音声とコヒーレントな残響音をいかに抑

制するかが重要であった．到来方向推定では，音声は複数であり，互いにコヒーレントではないもののこれらを分離する必要がある．さらに，同時に個々の音声には残響音があり，コヒーレントなマルチパスにも配慮しなければならない．つまり，到来方向推定はより複雑な問題である．一般的には，より困難な問題にはより多くの情報が必要である．到来方向推定が積極的に行われている分野は，信号にも細工ができるセルラーコミュニケーションである．しかし，受動的に得られる信号のみから信号源を推定する問題は，音声のみならず，地質，天文，生体計測に多く見ることができる．これらの分野では，更なる情報として，任意の方向に信号源があった場合得られるであろうアレイ出力パターンをあらかじめ知っているという仮定がおかれる．このアレイ出力パターンは，ステアリングベクトルと呼ばれる．音声の到来方向推定では，マイクロフォンアレイに対してθの方向にある周波数ωの音源のステアリングベクトル $\mathbf{d}(\theta, \omega)$が必要となる．つまり，計測場のモデルか実測データが必要となる．また，一般にθは離散化される．

ステアリングベクトルを用いると，残響除去での分散最小規範に対応する手法は，

$$\mathbf{w}^{\mathrm{T}}(\theta,\omega)\left\langle \mathbf{x}[n]\mathbf{x}^{\mathrm{T}}[n]\right\rangle \mathbf{w}(\theta,\omega) \to \min.$$

$$\text{s.t. } \mathbf{w}^{\mathrm{T}}(\theta,\omega)\mathbf{d}(\theta,\omega)=1$$

となる．これは，minimum variance beamformer と呼ばれ，最適重みは

$$\mathbf{w}(\theta,\omega)=\frac{\left\langle \mathbf{x}[n]\mathbf{x}^{\mathrm{T}}[n]\right\rangle^{-1}\mathbf{d}(\theta,\omega)}{\mathbf{d}^{\mathrm{T}}(\theta,\omega)\left\langle \mathbf{x}[n]\mathbf{x}^{\mathrm{T}}[n]\right\rangle^{-1}\mathbf{d}(\theta,\omega)}$$

となる．ここで最小化するコスト関数の計量行列は$<\mathbf{x}[n]\mathbf{x}^{\mathrm{T}}[n]>$であるが，これを

$$\sum_{\theta}\mathbf{d}(\theta,\omega)\mathbf{d}^{\mathrm{T}}(\theta,\omega)$$

に置き換えれば，厳密には少し異なるものの，$\mathbf{w}(\theta, \omega)$は重みつき最小ノルム解，あるいは最小 2 乗解となる．この置き換えは，$\mathbf{x}[n]$がコヒーレントであると分離効果が向上し，そうでなければあまり効果がない．

その他，MUSIC(Multiple Signal Classification) やステアリングベクトルを必要としない ESPRIT(Estimation of Signal Parameters via Rotation Invariance Technique)が提案されている．

到来方向推定は，いずれの方法も何らかのコスト関数の最小化・最大化問題で記述できる．

2-3 独立成分分析

統計的に独立な音声信号が線形混合されている場合，独立性以外の先験情報を利用することなく，混合信号を分離する手法が盛んに研究されている．このような手法は，独立成分分析 (ICA: Independent Component Analysis)と呼ばれている．

ICA の問題設定を以下に示す．いま，統計的に独立で時間平均零な源信号を $\mathbf{s}[n]$，観測される信号を $\mathbf{x}[n]$とすると，この線形混合は

$$\mathbf{x}[n]=\mathbf{A}\mathbf{s}[n]$$

と表すことができる．ここで混合行列 $\mathbf{A}$ については何ら先見情報を持たない．次に，適当な行列 $\mathbf{B}$ を用いて，

$$\mathbf{y}[n]=\mathbf{B}\mathbf{x}[n]$$

とする．このとき $\mathbf{y}[n]$の要素ができる限り統計的に独立となるように $\mathbf{B}$ を選ぶ．どのように $\mathbf{B}$ を選ぶかについては，後述するさまざまな手法が提案されているものの，最終的に ICA は線形作用である．さて，理想的には，$\mathbf{AB}=\mathbf{I}$ となるが，ICA ではそこまでは要求せず，スケール倍と順序の不定性は許容している．つまり，$\mathbf{AB}$ の各列の要素はどれかひとつだけ非零であり，その要素は各列で排他的である．実際には，必ずしもそのような状況を直接目指すわけではなく，結果的にその状況に近いものが手に入ることを期待している．

$\mathbf{B}$ の選び方の代表的な手法 3 つを示す．

第 1 に，Infomax では，まず，$\mathbf{y}[n]$の各要素ごとに作用する非線形関数を $\mathbf{g}(\cdot)$とし，

$$\mathbf{u}[n]=\mathbf{g}(\mathbf{y}[n])$$

を考える．$\mathbf{g}(\cdot)$が $\mathbf{y}[n]$の各要素の累積確率密度関数であれば，$\mathbf{u}[n]$は一様分布となる．そこで，$\mathbf{u}[n]$のエントロピーが最大となるように，$\mathbf{B}$ を最適化する．ここで，$p(\cdot)$は確率密度関数であり，$p_1(\cdot)$は一様分布の確率密度関数である．

第 2 に，FastICA では，まず，$\mathbf{y}[n]$のホワイトニング

$$\mathbf{y}'[n]=\mathbf{\Lambda}^{-1/2}\mathbf{V}^{\mathrm{T}}\mathbf{y}[n]$$

を行う．ここで，$\mathbf{V}$，Λは，$\mathbf{y}[n]$の分散共分散行列の対角化

$$\langle \mathbf{y}[n]\mathbf{y}^{\mathrm{T}}[n]\rangle = \mathbf{V}\mathbf{\Lambda}\mathbf{V}^{\mathrm{T}}$$

によって与えられる．ホワイトニングによって $\mathbf{y}'[n]$の各要素は 2 次までの統計量が無個性となる．次に，適当な非線形関数 $f(\cdot)$を用いて，

$$f\left(\mathbf{w}^{\mathrm{T}}\mathbf{y}'[n]\right)$$

が極大・極小となる単位ベクトル **w** を求める．$f(\cdot)$には非正規性を量る関数が用いられることが多い．**w** は

$$\mathbf{b} = \mathbf{w}^{\mathrm{T}}\mathbf{\Lambda}^{-1/2}\mathbf{V}^{\mathrm{T}}$$

によって，**B** の行 **b** を構成する．

第 3 に, JADE(Joint Approximate Diagonalization of Eigenmatrix)では，独立性の必要条件を用い，

$$\langle \mathbf{x}[n]\mathbf{x}^{\mathrm{T}}[n+m]\rangle$$

を異なる m に対し同時対角化できる **B** を求める．

ICA の応用例はインターネット上で簡単に入手することができ，音声や音楽に対する単純な線形混合の場合，人が実際に聞いてみると驚くほどの分離結果となる．一方，残響音響・歪み下では，単なる線形混合ほどの分離性能を得ることはできない．

なお，先のマイクロフォンアレイによる到来方向推定も，混合された音声の分離という点で ICA と同じ目的を持つ．2 種類の独立信号の分離という点で，これらの 2 手法から得られる分離行列が酷似することが指摘されている．このことは次のように考えると，容易に理解される．線形混合の分離の場合，理想的には，混合行列 **A** の各列の双対基底が分離行列 **B** の各行に現れればよい．つまり，混合行列 **A** のある列は，分離行列 **B** の対応する特定の行以外とはすべて直交していればよい．このことは，2 手法とも特定の音声以外に零点を配置するアルゴリズムとなっていることを意味している．

独立成分分析は，いずれの方法も何らかのコスト関数の最小化・最大化問題で記述できる．

2-4 音声認識

現代的な音声認識は，隠れマルコフモデルを用いたケプストラムのマッチングによって行われることが多い．また，予約語の検出はワードスポッティングと呼ばれる．

音声認識においても, 何らかのコスト関数の最小化・最大化問題で記述できる．

2-5 聴覚心理

ICA で線形混合に用いた行列と分離に用いた行列の積は単位行列とは程遠い行列であることが多い．しかし，人間には音声は分離されて聞こえる．このことは，人間の聴覚機能の複雑性を象徴している．つまり，混合音の分離という観点において，人の知覚上，実効的に逆行列としてよい集合は単位行列を含むかなり大きな分布を持っていることを示唆している．

3 提案するシステム

本研究は，マイクロフォンアレイで計測された不特定多数の音声から予約語を抽出することを目的とする．第 2 章で説明した要素技術を組合せる戦略では，混合音分解(到来方向推定，または ICA)→残響除去(逆システム)→ワードスポッティングというシステムのシーケンスが考えられる．この戦略は，あるシステムの出力を次のサブシステムに都合のよい入力にするというリレー方式である．しかし，それぞれのシステムの働きを担う要素技術は，第 2 章で説明したように，すべて何らかのコスト関数の最小化・最大化問題で記述できた．このため，各システムにおいて順次最適化せず，まとめてコスト関数を設定することが考えられる．つまり，最終出力までの中間段階の出力は，単に内部状態変数に過ぎず，その意味や効果を評価する必要はないといえる．人間の聴覚機構や注意機構も，個々に評価可能な目的を持ったシステムのリレー方式ではないかもしれない．

最後に本システムの概念図を示す．ロボットは，マイクロフォン，位置センサ，擬似雑音を発生させる機能を有する．これにより，マイクロフォンアレイの一部は固定ではなくなる．また，擬似雑音によってステアリングベクトルの更新も行う．