長岡技術科学大学大学院

工学研究科修士論文

題　　目

励起特徴と声道特徴を用いた
音声感情認識の精度向上に関する研究

指導教員

杉田 泰則　准教授

著　　者

工学専攻

電気電子情報工学分野

20317789　　西谷 洋哉

令和6年2月9日

目 次

第1章　序論

1.1　研究背景

1.1.1　音声感情認識

工学分野における感情認識とは，主に機械学習や人工知能技術などを利用してヒトの感情を認識するためのシステムを指す．これらのシステムは，文章や音声，画像，動画などのデータを処理し，そこから得られるさまざまな情報を基に感情要素を抽出することが可能である．

感情認識の一つである音声感情認識 (以下，Speech Emotion Recognition，SER) は，ヒトが発する音声信号から話者の感情を自動的にコンピュータで認識する技術のことを指す．この技術は，言葉の発音や発話のリズム，音の高さ，強さなど音声信号に含まれるさまざまな特徴を分析して話者の感情を推定する．

具体的に SER を行うプロセスを以下の図 1.1 に記載する．図 1.1 のように入力されたヒトの音声から波形を認識，その音声波形から感情認識に必要な特徴の抽出を行い，得られた特徴を基に感情の判定を行う．

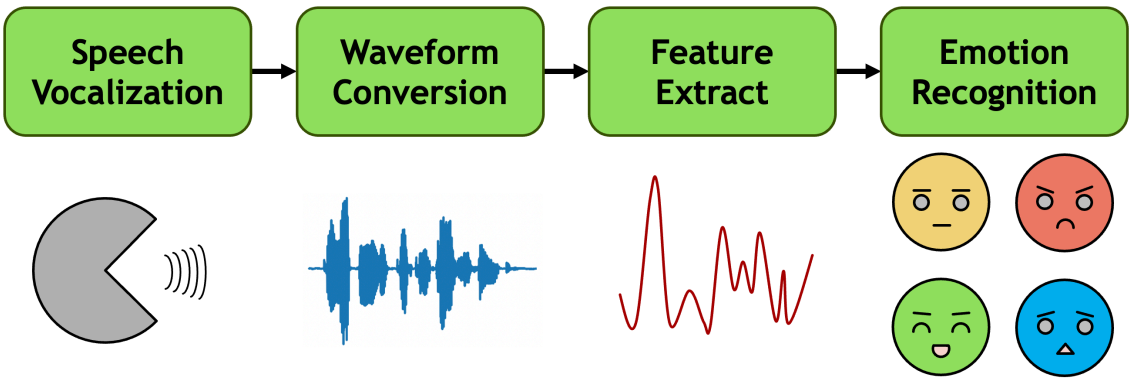

図 1.1: 音声信号から感情認識を行うまでのプロセス

1.1.2　SER の応用例

SER の応用例は多岐にわたる．例えば，学校の相談室や会社の会議室など，コミュニケーションが発生する場所に SER システムを設置することで利用者の現在の感情をリアルタイムでフィードバック可能となる．これにより，効果的なサポートや意思疎通が可能となり，コミュニケーションを向上させることが期待される．さらに，コールセンターなどの電話口にも SER システムを取り付けることで，顧客の感情を正確に把握し，適切な対応が可能となる [1].

このような応用例により，SER はさまざまな場所で活用され，音声アシスタントなどの観点からもヒトと機械のコミュニケーションにおいて新たな展望を切り拓いている [1]．そのため，SER の研究と実践は社会におけるコミュニケーションとサービスの向上に寄与していると言える．

1.1.3　SER の課題

しかし，現存している SER のシステムにはいくつかの課題が存在している．

まず，音素 (子音や母音など) を用いた SER [2] や韻律情報 (リズム感など) を用いた SER [3] には，話者依存が大きくなってしまうという問題がある．話者依存とは，音声信号や音声データが個々の話者に依存しているという性質を指している．これは話者ごとの発音，声の高さや音域，アクセント，発話スタイルなどが異なることから生じてしまう．この話者依存が大きいと，個々の話者ごとに音声データコーパスを作成する必要があり，これはコストが高いことを表している．例えば，前述のコールセンターを例に挙げる．新規の顧客から電話が来るたび，電話対応の前に毎回新規顧客の音声コーパスを作成する必要があるということになる．これは顧客も職員も負担が大きく，現実的ではない．そのため，現存している音声コーパスのみを使用して感情認識ができると非常に効果的である．これにより，一般的な言語パターンや感情の表現を捉えることが可能となり，話者ごとの個別のデータ収集が不要となる．特に，大規模で多様なコーパスが利用可能であれば，SER システムの話者依存性を軽減し，汎用性を向上させることが期待される．これにより，コストの削減やシステムの利便性，SER の実用性が向上し，異なる応用領域においても効果的な感情認識が可能となる．

次に，深層学習を導入した SER [4] は大量の音声データから特徴を学習する能力を有しているが，大規模なモデルは訓練データに過剰に適合しやすく，新しいデータに対して汎化性能が低下する可能性がある．特に感情認識のようなシステムでは，個々の話者や文脈に依存した過学習が発生しやすい．

加えて，これらすべての手法 [2] [3] [4] において感情ラベルが付与された音声コーパスが不足しているという課題も存在する．特に，深層学習を導入した SER は，学習を行う都合上コーパスを大量に使用するモデルもある．

既存の感情ラベルが付与された音声コーパスは，主に各国の俳優や声優が特定の感情を演じたもので構成されており，その作成には高いコストがかかる．また，収録された音声の感情評価を行う場合にどうしても主観評価になってしまう問題から，評価の妥当性も低いといえる．

以上の理由から，感情ラベルが付与された音声コーパスのデータを増加させることは非常に困難である．

1.2　研究目的

SER の理想的なシステムは，多数の音声コーパスを使用せず，かつ話者に依存しにくい性質を有し，かつ計算コストが低いものが望ましい．これらの要求を満たす有望な手法の一つは，励起特徴量を活用した SER であることが報告されている [5] [6] [7].

しかしながら，この励起特徴量を用いた SER にも欠点が存在している．この研究では，主として無感情，怒り，喜び，悲しみの四つの感情を認識しているが,「無感情と悲しみ」の認識精度が相対的に低いことが明らかになっている．そこで本研究の主な目的は，励起特徴量を活用した SER の利点を維持しつつ,「無感情と悲しみ」の認識精度を向上させることにある．

1.3　本論文の構成

本論文の構成は以下の通りである．第 1 章では研究背景および目的について述べた．第 2 章では従来手法である「励起特徴量を用いた SER」について説明を行う．第 3 章では提案手法である「励起特徴量と声道特徴量を考慮した SER」について詳細を示す．第 4 章では従来手法と提案手法の比較実験を行ない，提案手法の有用性を示す．第 5 章では，本研究の結論を示し，結びとする．

第2章　従来手法

この章では，先行研究である「励起特徴量を用いたSER」に焦点を当て，その重要性について示す．主に参考にしている先行研究は，文献 [5] である．

文献 [5] で使用されている「励起特徴量を用いたSER」は，発声機構や声帯の動きから抽出された瞬間基本周波数，励起の強さ，励起のエネルギーの三つの励起特徴量を用いて感情認識を行っている．三つの励起特徴量は声門の開閉運動に基づいた声帯の状態を定量的に捉え，それが感情表現にどう関与しているかを研究するものである．

2.1　従来手法の背景

文献 [5] では,「励起特徴量 (Excitation Feature)」と呼ばれる特徴量を用いて音声信号から感情の認識を行っている．

ヒトの声門は，発声過程において重要な役割を果たしている．声門は気道にある筋肉で構成され，定期的な開閉運動を行う．この開閉運動は，声帯が周期的に開いて閉じることを指し，これによって音声が生成される．特に，声門が閉鎖した瞬間には破裂音と呼ばれる音が発生し，この破裂音には音声生成における超文節が含まれている [8]．

超文節は，音声信号における声門の閉鎖状態に関連する情報を指し，超文節の抽出は声門の開閉運動を捉えることで，音声信号から特徴的な要素を抽出する手段となる．

励起特徴量は，この声門の開閉に基づいて声帯の状態を表す特徴量である．声門の開閉は音声信号において重要な情報を提供し，励起特徴量はその情報を用いて声帯の状態を定量的に捉える．この励起特徴量を用いることで，音声処理において効果的な信号処理や解析が可能となる．

簡潔に言えば，声門の動きから得られる情報を利用して，音声信号の特徴を抽出し，それを基に様々な音声処理が行われるのである．

以下の図2.1はヒトの声門を表す概略図である．図2.1左側は声門の閉鎖，図2.1右側は声門の開放を表している．図2.1において，声門の中心部で開閉している二本のひだ状の器官が声帯である．ヒトは肺から空気を押し出し，その呼気によって声帯を振動させることで音声が生成される [9]．

励起特徴量の取得において，最初のステップは音声信号から声門の閉鎖情報を抽出することである．声門の開閉運動は発声の基本的な構成要素であり，この動きによって気道を通る空気の流れが制御され，発声が可能となる．したがって，声門の開閉運動の正確な把握は，音声信号から励起特徴量を取得する上での重要な初期段階となる．

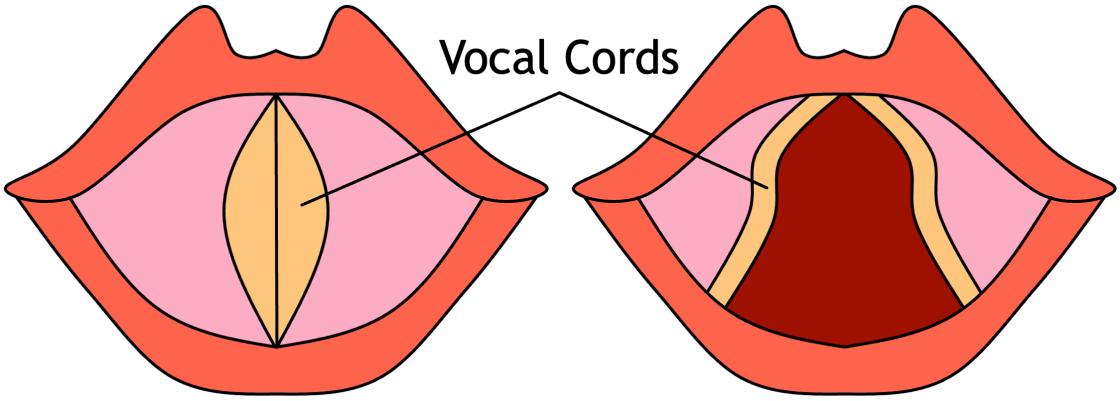

図 2.1: 声門の概略図 (左: 閉鎖時，右: 開放時)

声門の閉鎖情報が特に重要なのは，発話者が強調する部分や感情的な表現が含まれる可能性が高い瞬間を捉える役割を果たすからである．声門が閉じられる瞬間には，音声信号に破裂音や強弱の変化が現れ，これが発話者の感情や強調の度合いを反映する要素となる．つまり，声門の閉鎖情報を適切に取得することは，感情の微細なニュアンスを捉える上で不可欠な手段となる．

このような声門の閉鎖情報は，感情情報を含んでいると考えられている．感情は発声において音声信号にさまざまな形で現れるため，声門の開閉情報の取得は感情の把握に寄与する重要な手法となる．感情の微細な変化やニュアンスは，言葉だけでなく発声の質やリズムにも影響を与えるため，これを正確に捉えることは音声処理や SER において非常に有益となる．

総じて，声門の閉鎖情報の取得は励起特徴量の鍵となり，感情の豊かな表現や微細なニュアンスを捉える上で不可欠なプロセスである．

2.2　声門閉鎖瞬間の取得

声門閉鎖瞬間 (以下，Glottal Closure Instants，GCIs) の情報は通常，電気声門図 (以下，Electro Grotto Graph，EGG) 機器を使用して取得される．EGG 機器はヒトの首の表面に配置され，声門の開閉運動をモニタリングすることで GCIs のタイミングを測定する．しかし，現状では EGG 機器は医療従事者にのみ提供され，一般の人が手に入れることが難しい状況である．

この問題を解決するために，一般の人が手元の環境で GCIs 情報を取得できるように，一般的な録音機で収録された音声信号に対するフィルタ処理が提案されている [5] [6] [7]．このフィルタ処理は，音声信号中の GCIs に対応する特徴を強調したり，その他のノイズや周辺の情報を取り除いたりすることを含んでいる．これにより，GCIs の位置やタイミングを正確に検出できるようになり，EGG 機器に依存せずに音声信号から GCIs を取り出すことが可能となる．

以降は音声信号から GCIs を取り出す処理について詳しく説明をしていく．

2.2.1　音声信号の差分化

まずは，GCIs を取り出しやすくするために音声信号に前処理を行う．その中でも，一般的な前処理手法として差分化を利用する．以下の式 2.1 は，音声信号の差分化を表している [5].

$$x[n] = s[n] - s[n-1] \tag{2.1}$$

ここで，$x[n]$ は差分化された音声信号であり，$s[n]$ は元の音声信号，n はサンプル数を指す．この差分化は，時間的に変化する低周波部分の偏りを取り除くために用いられる．

差分化は，音声信号の各時点での変化を捉える手法であり，式 2.1 では，現在のサンプル $s[n]$ から一つ前のサンプル $s[n-1]$ を引くことで，連続するサンプル間の変化を反映した信号 $x[n]$ を得られる．この変化の情報が，GCIs に関連する振る舞いを捉えるのに役立つ．

特に，GCIs は音声信号において急激な変化をもたらすため，差分化を通じてこの急激な変化を強調することで，後続の処理段階で GCIs を効果的に抽出できるようになる．この前処理は，信号処理の手法を駆使して感情や発声の特徴を明示的にし，分析の対象となる声門の動きに焦点を当てる一環と言える．

2.2.2　ゼロ周波数フィルタの適用

次に，差分化した信号に対してゼロ周波数フィルタ (以下，Zero Frequency Filter，ZFF) を適用する．以下の式 2.2 は，ZFF を表している [5].

$$y_o[n] = \sum_{k=1}^{4} a_k y_o[n-k] + x[n] \tag{2.2}$$

ここで，$y_o[n]$ はフィルタ適用後の信号，$x[n]$ は差分化された音声信号，a_k はフィルタ係数を指す．具体的なフィルタ係数はそれぞれ $a_1 = +4$, $a_2 = -6$, $a_3 = +4$, $a_4 = -1$ とされている．

ZFF は，指定された周波数以下の信号成分を通し，それより高い周波数成分を遮断するフィルタである．図 2.2 の周波数特性からも分かるように，このフィルタは特に直流成分と低周波成分を強調し，高周波成分を抑える効果がある．直流成分は信号の基本的な成分であり，低周波成分は GCIs に関連するような変動を含んでいる [10].

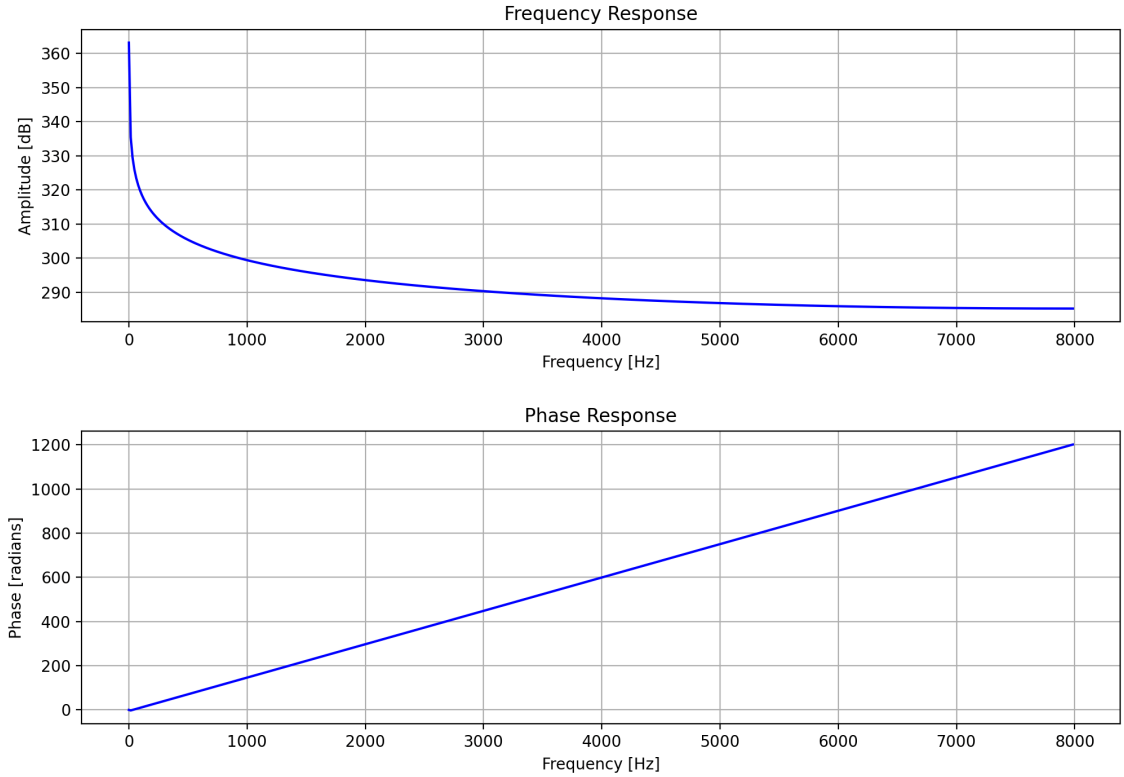

図 2.2: ZFF の周波数応答と位相応答

式 2.2 では，直前の 4 つのサンプルを用いて現在のサンプルを計算している．これにより，ZFF は過去の信号の影響を考慮しつつ，信号を滑らかに変化させる．よって，ZFF の処理を用いることで GCIs が反映された信号が強調される．ただし，図 2.2 の位相応答から分かるように，出力信号には大きなトレンド成分が含まれているため，これを除去する必要がある．

2.2.3　局所平均によるトレンド除去

ZFF を通過した信号 $y_o[n]$ は，長期的に増加または減衰のトレンドを持つ．このようなトレンドが残ると，今後の処理において GCIs の検出が難しくなってしまう問題がある．そのため，このトレンドを取り除くために，局所平均処理が行われる．以下の式 2.3 は，局所平均を用いたトレンドの除去を表している [5].

$$y[n] = y_o[n] - \frac{1}{2N+1}\sum_{i=-N}^{N} y_o[n+i] \tag{2.3}$$

ここで，$y[n]$ はトレンドが取り除かれた信号を表し，$y_o[n]$ は ZFF を通過した信号，N はサンプルの数を指す．式 2.3 では，各サンプルに対してその周囲のサンプルを用いて局所平均を計算し，これを $y_o[n]$ から差し引いている．

具体的には，各サンプルの前後 N 個のサンプルを含む窓を設定し，その窓内のサンプルの平均を計算する．この平均値を用いて $y_o[n]$ から引くことで，長期的な増加や減衰のトレンドを取り除き，信号の局所的な変動がより強調されるようになる．

$y[n]$ において負から正へのゼロ交差部分は，元の音声信号 $s[n]$ における GCIs に対応しており，これを利用して GCIs のタイミングを特定する．

2.2.4　GCIs の抽出

グラーツ工科大学のピッチ追跡データベース (PTDB-TUG) [11] 内の気導マイクから収録された音声信号と，EGG 機器から測定された信号を用いて，式 2.1，式 2.2，式 2.3 によって得られる GCIs の位置が妥当かどうかを調査した．

実際に音声信号に，順に式 2.1，式 2.2，式 2.3 を適用させたものを示す．図 2.3 には，以下の要素が示されている．

a. 元の気導音声信号

b. 気導音声信号に式 2.1 を適用した信号

c. 式 2.1 を適用した信号に式 2.2 を適用した信号

d. 式 2.2 を適用した信号に式 2.3 を適用した信号

e. EGG 機器を用いて測定した信号．

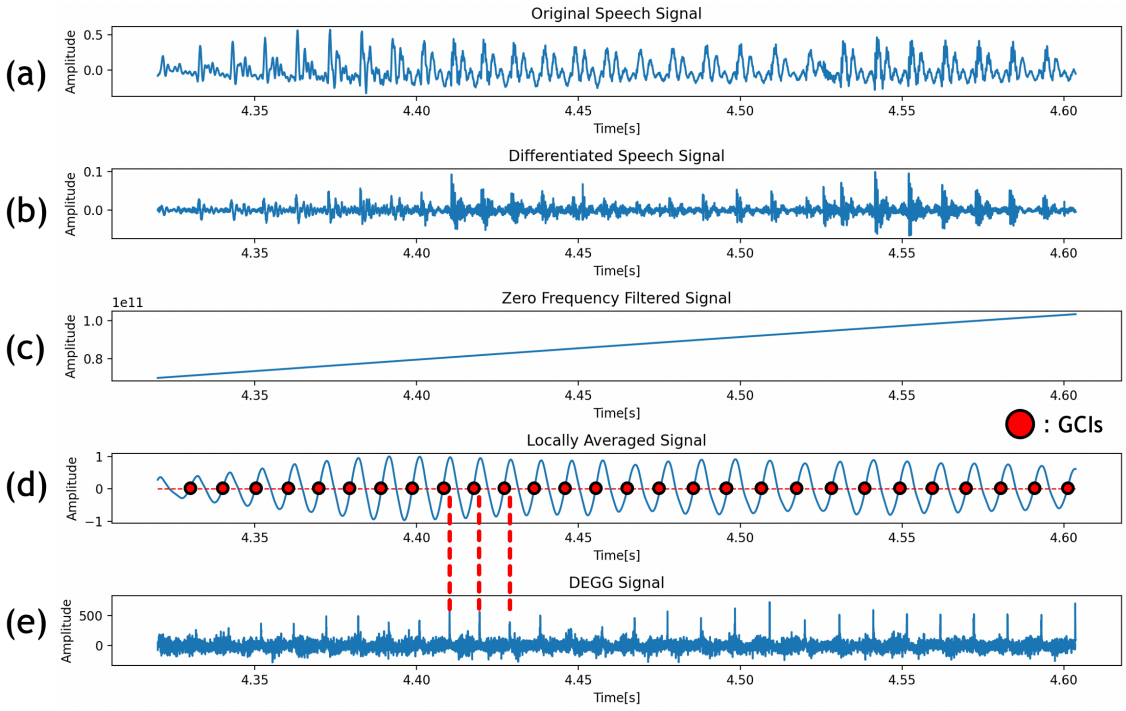

図 2.3: (a) 元の音声信号，(b) 差分化した a の信号，(c)ZFF を通過した b の信号，(d) 局所平均を行った c の信号，(e)EGG 機器を用いて測定した信号

図 2.3 の (d) の信号に記されている赤い丸印は，信号が負から正へゼロ交差している部分を表しており，同時に GCIs の位置を表している．また，図 2.3 の (e) の信号は EGG 機器で測定された信号であり，信号のピーク位置は GCIs の位置を示している．ここから (d) と (e) を比較した際，(d) の赤い丸印と (e) の信号のピークの位置と個数が一致していることが観察された．

この結果から，気導マイクから収録された音声信号に対してフィルタ処理を行うだけで，専用の EGG 機器を使用せずとも GCIs を取り出すことが可能であることが分かった．

2.3　励起特徴量の導出

検出された GCIs の位置を表すサンプル番号を $g_1, g_2, \cdots, g_M$ として，瞬間基本周波数，励起の強さ，励起のエネルギーの三つの励起特徴量を計算する．

2.3.1　瞬間基本周波数

瞬間基本周波数 (以下，Fundamental Frequency，F0) は GCIs における声帯振動の周期性を表す励起特徴量であり，以下の式 2.4 によって取得が可能である．

$$F_{0_{g_c}} = \frac{F_s}{(g_c - g_{c-1})}, c = 2, 3, \cdots, M. \tag{2.4}$$

ここで，$F_{0_{g_c}}$ は GCIsg_c における F0 を示し，F_s はサンプリング周波数を表す．式 2.4 では，各 GCIs 間の時間差を用いて，GCIs の F0 を計算している．

具体的には，g_c と g_{c-1} のサンプル番号の差を取り，サンプリング周波数 F_s をこれで割ることで，その時間差における基本周波数を得る．この F0 は，声帯振動の周期性や発声の基本的な情報を捉えるのに役立つ．

2.3.2　励起の強さ

励起の強さ (以下，Strong of Excitation，SoE) は声門閉鎖の継続時間を示す励起特徴量の一つであり，以下の式 2.5 によって取得が可能である．

$$SoE_{g_c} = |y[g_c + 1] - y[g_c - 1]|, \quad c = 1, 2, \cdots, M. \tag{2.5}$$

この式では，隣接するサンプルの振幅差の絶対値が励起の強さとして取られる．SoE_{g_c} は GCIsg_c において，その前後のサンプルの振幅差を絶対値で求めることで得られる．この励起の強さは，声門閉鎖がどれだけ強いかを示す指標として用いられ，振幅差が大きいほど声門閉鎖が強調されていると解釈できる．

2.3.3　励起のエネルギー

励起のエネルギー (以下，Energy of Excitation，EoE) は元の音声信号のエネルギー量を示す励起特徴量の一つであり，以下の式 2.6 によって取得が可能である．

$$EoE_{g_c} = \frac{1}{2K+1} \sum_{i=-K}^{K} HE^2[g_c + i], c = 1, 2, \cdots, M. \tag{2.6}$$

ここで，K は 1[ms] 間のサンプル数，HE は $x[n]$ の線形予測残差から取り出したヒルベルト包絡を表す．式 2.6 では，ヒルベルト包絡の各サンプルの二乗和を 1[ms] ごとに平均している．

得られたエネルギーは，声門閉鎖によって生じる振動の強さや音声信号のエネルギーの変動を表す．

2.4　励起特徴量を用いた SER

文献 [5] では，ここまでの処理で取り出した F0，SoE，EoE の三つの励起特徴量を用いて感情認識を行っている．以下の図 2.4 は，各感情の音声ファイルから取り出した三つの励起特徴量を三次元特徴空間にプロットしたものである．

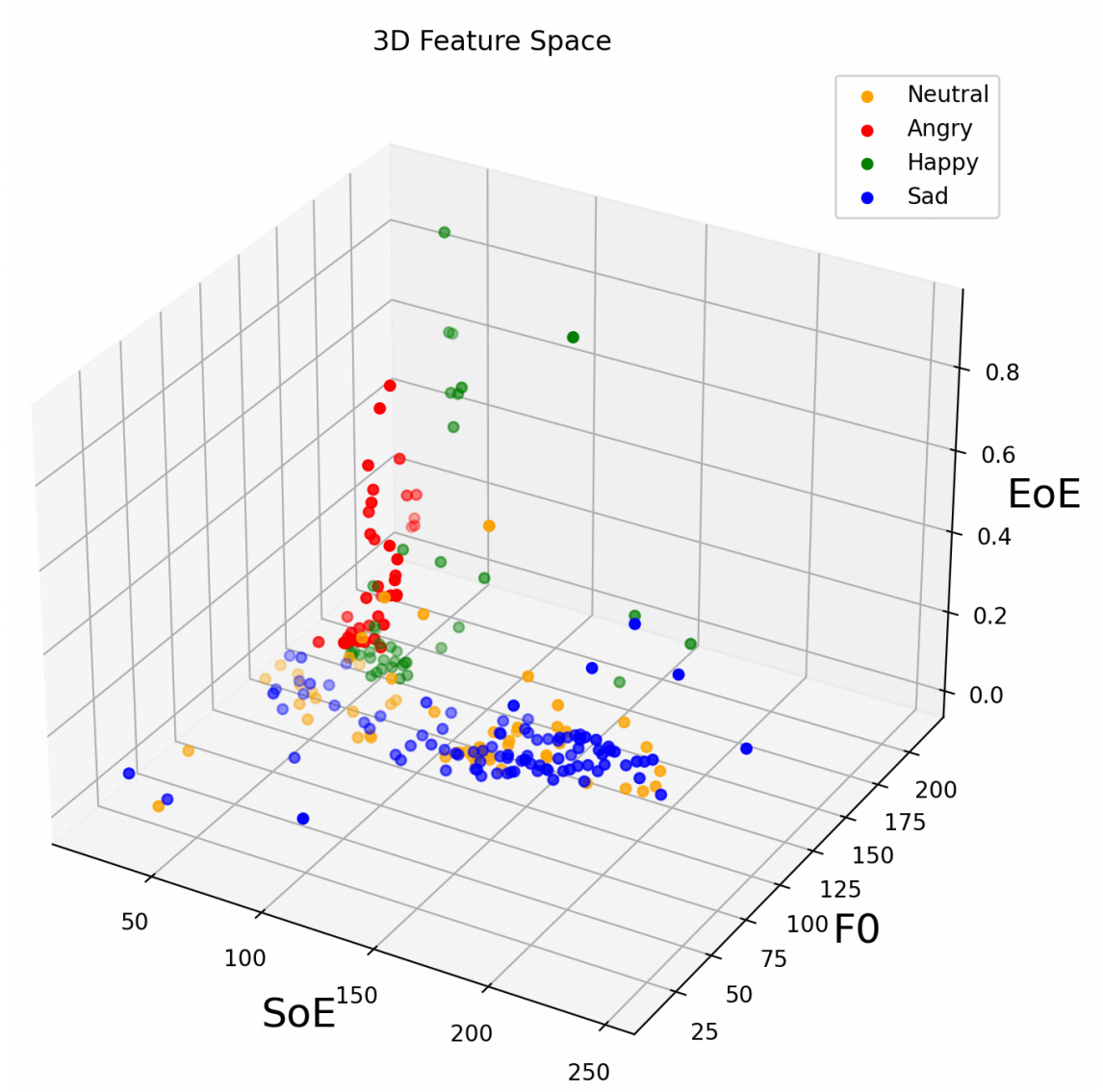

図 2.4: 四感情の励起特徴量の三次元特徴空間
(橙: 無感情，赤: 怒り，緑: 喜び，青: 悲しみ)

図 2.4 から分かるように三次元特徴空間において，各感情の分布の仕方に固有の特徴が生まれている．この固有の特徴をコンピュータ内で捉えることで感情の認識を行うことが可能である．

第3章　提案手法

文献 [5] では励起特徴量を取り出すためにいくつかの処理を行なっていた．その処理の中で GCIs を取り出す上で必要な処理である「局所平均によるトレンド除去 (式 2.3)」があったが，使用するサンプル数を決定する窓長は固定して決定していた．しかし，音声信号によって適切なサンプル数が変化することから，窓長が常に固定であると，GCIs がうまく取れない可能性がある．そこで，本研究では自動で適切な窓長を決定する効率的な処理方法を提案した．

また，本研究の目的である「無感情と悲しみの認識精度向上」を達成する提案手法として，2 章で説明した三つの励起特徴量に新たな特徴量である「声道特徴量」を加えた，四つの特徴量を用いて SER を行う方法を提案した．

3.1　トレンド除去の処理改善案

本論文の第 2 章でトレンド除去を行うことで GCIs を取り出す処理を紹介した．しかし，この局所平均を行う上で重要となるのが，局所平均処理の適用回数と使用するサンプル数である．

文献 [7] において，この局所平均処理を合計で三回行っていることが明記されていることから，本研究でも同数の局所平均処理を行うこととする．

サンプル数 $2N+1$ の適切な決定については，使用する音声信号 $s[n]$ によって異なるため，適切な窓長の決定が非常に重要である．このサンプル数が少なすぎると，不要なゼロ交差部分が大量に検出され，逆にサンプル数が多すぎると必要なゼロ交差部分が検出されなくなる可能性がある [7].

しかし，人間の手でこの窓長を毎回決定するのは非効率的である．そこで，サンプル数を適宜変更して得られた結果を参考に適切なサンプル数を決定する方法を用いる．この手法では，GCIs を抽出するために処理された信号が正弦波に近い概形であることを利用する．取り出す手順は以下のようになる．

1. トレンドが取り除かれた信号 $y[n]$ からピークを検出する．
2. ゼロ線を境目として，ピークの奇数番目 (偶数番目) が，偶数番目 (奇数番目) のピークの時間位置に合うように正側 (負側) の信号をシフトする (図 3.1)．
3. 正側の信号のサンプルと，負側の信号のサンプルの絶対値を取得し，二つのサンプルの値の相関係数を計算する．
4. 導出された相関係数を基にして，対称性を自動で判断する (相関係数が 1 に近ければ対称性が強い)．

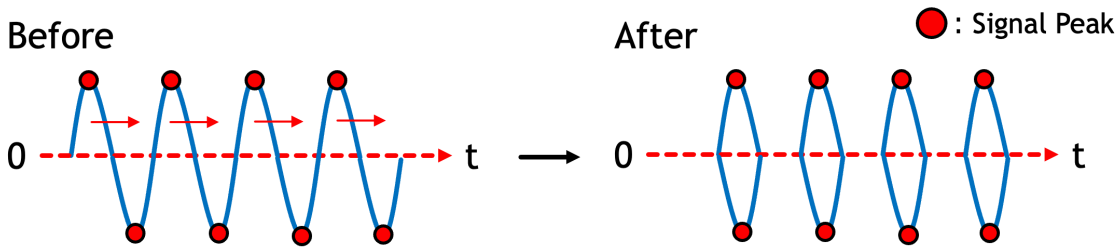

図 3.1: ゼロ線を境目として奇数番目と偶数番目のピークの
時間位置が合うようにシフトされた信号

これらの手順を通じて，サンプル数の変更に伴って得られる信号の特性を観察し，最終的には対称性を利用して適切なサンプル数を自動的に決定する．この手法は，人間の目視や手作業でサンプル数を選択するのが難しい場合に，効率的に適切なサンプル数を見積もる手段を提供する．

図 3.2 は，窓長を変更することで，フィルタ処理された信号に対するトレンド除去がどのように変化するかを示している．

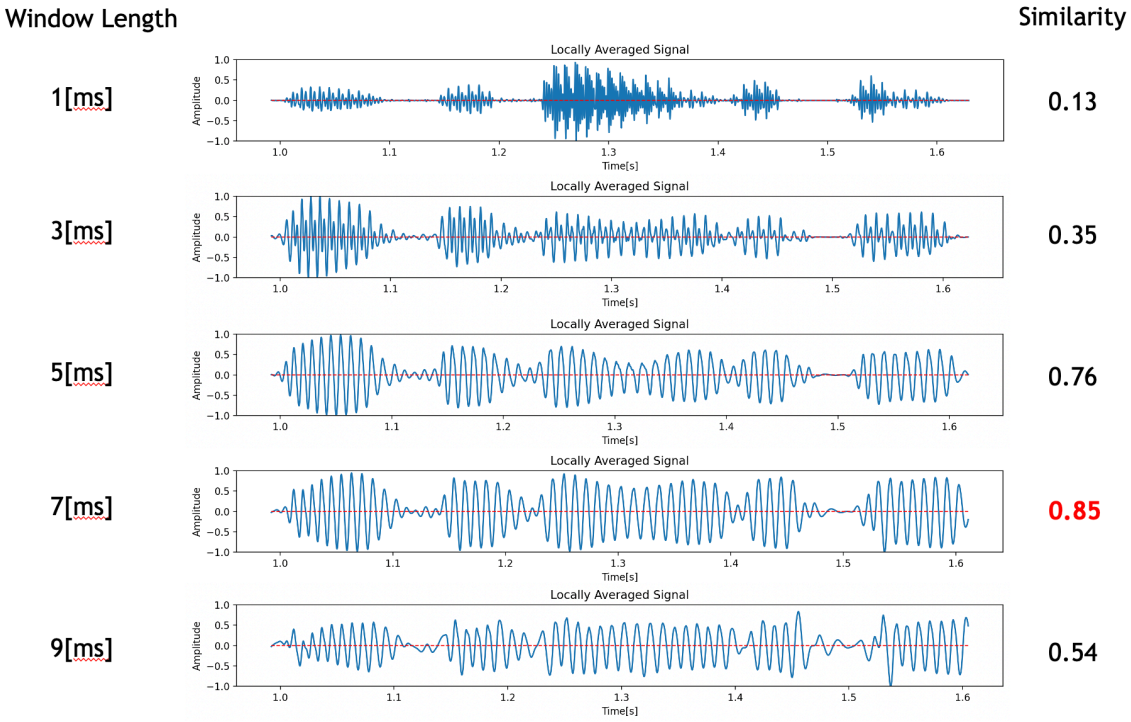

図 3.2: フィルタ処理された信号に対するトレンド除去の窓長の影響

図 3.2 の上からそれぞれ 1[ms]，3[ms]，5[ms]，7[ms]，9[ms] の窓長で局所平均によるトレンド除去を行なっており，各プロットの右側には評価の結果である相関係数が示されている．今回の図 3.2 の場合，GCIs の抽出に最も有効な窓長は相関係数が一番高い 0.85 の 7[ms] であったと言える．

このように変化を自動的に，かつ数値的に評価することで最適なサンプル数を選択することが可能となる．

3.2　声道特徴量

声道は，呼吸や発声などの機能を果たすための通路であり，声門から口までの経路を指す．この通路は，主に喉頭，咽頭，口腔，鼻腔といった部分から構成されている．

声道を通る振動する空気は，声門で発生した振動が伝わり，声帯によって発声される．この振動した空気は，喉，咽頭，口腔，鼻腔を通り，最終的に口や鼻から外部に排出され，我々が聞くことができる声となる．

以下の図 3.3 はヒトの声道，声門，声帯のそれぞれの位置を表している．

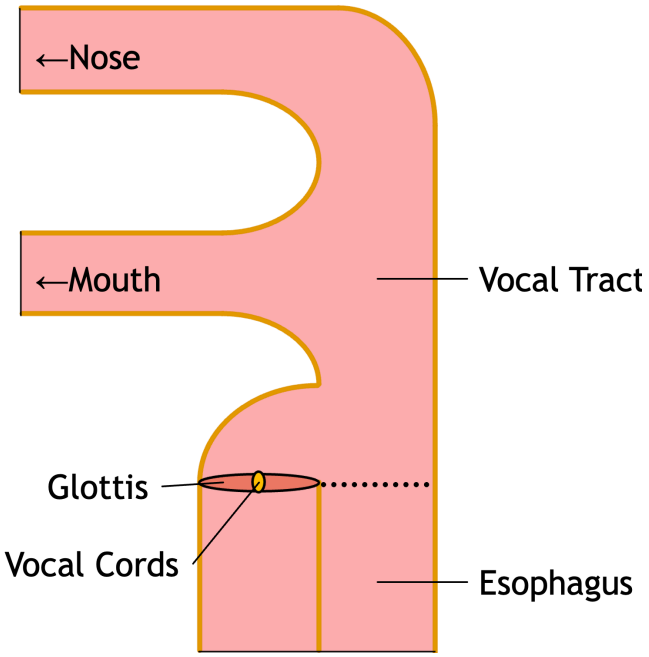

図 3.3: 声道，声門，声帯の位置関係

悲しみや他の感情によって音声が変化するという現象は，声道の長さや形状が影響を与える可能性がある．先行研究によれば，悲しみの感情を含む音声を発声する際には声道が短くなる傾向があり，逆に他の感情 (無感情，怒り，喜びなど) を含む音声を発声する際には声道が長くなる傾向があるとされている [12] [13] [14].

声道の長さが変化することで音声の基本周波数や音の高さが変わる．声道が長いときには，発せられる音の波長が長くなり，音が低く聞こえる．逆に，声道が短いときには，発せられる音の波長が短くなり，音が高く聞こえる．この現象は管楽器などでも見られ，図 3.4 に示されるような管内の空気の振動に基づいている．

悲しみの感情を含む音声を発声する際に声道が短くなることで，音の高さが増す傾向があるため，この特徴を利用して「無感情と悲しみ」の認識精度向上が期待できると考えられる．

このような声道の動きに基づく特徴量を抽出することで感情の違いをより効果的に捉え，音声認識の性能向上に寄与できる可能性がある [15] [16].

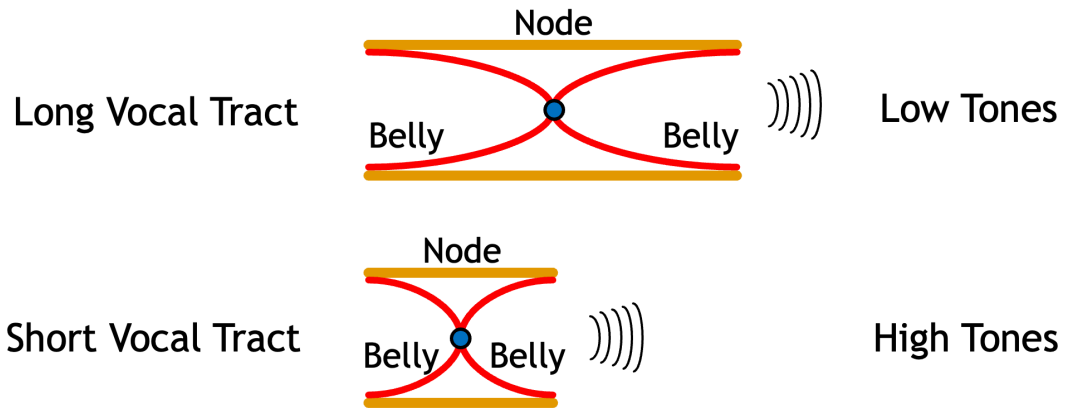

図 3.4: 声道中における空気の基本振動 (両端は開いている状態)

声道の長さが音声の高さに影響を与える概念は，音波の振動数と声道の長さの相関に基づいている．振動数は波の周期的な変動回数を示し，音の高さと密接に関連している．

声道が音波の振動に対してどのように反応するかが鍵となる．声道が短い場合音波の振動数が増加する．これは短い声道が波を迅速に伝播させてより多くの振動が 1 秒あたりに生じるからである．逆に声道が長い場合は振動数が減少する．これは長い声道は波をよりゆっくり伝播させて 1 秒あたりの振動数が低くなる．

この振動数の変動が声の高さと低さに影響を与える．振動数が高いほど音が高くなり振動数が低いほど音が低くなる．したがって声道が短い場合は高い音が，声道が長い場合は低い音が発生しやすくなる．

音声データにおいて，無感情と悲しみの発声における声道スペクトルが異なるという観察結果がある．図 3.5 は，無感情，怒り，喜び，悲しみの声道スペクトルを比較したものである．声道スペクトルは，音声信号からも取り出される声道の特徴を表す周波数成分の強度や分布を表すものであり，異なる感情状態における声道の特徴が視覚的に把握することが可能である．

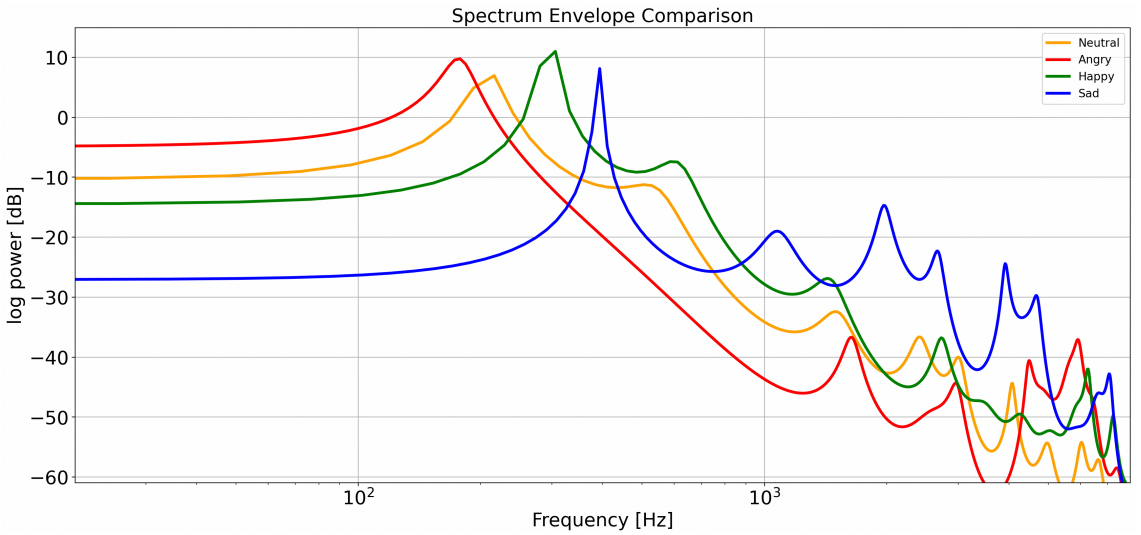

図 3.5: 声道スペクトル

図 3.5 の横軸は周波数 (対数スケール)，縦軸はその周波数における信号の対数パワーを示している．

声道スペクトルは音声信号が声道を通過する際にどの周波数が強調されるかを表しており，感情によって声道の特性が変わることが示唆されている．

無感情と悲しみの声道スペクトルを比較すると特定の周波数成分での強度や形状に違いが見られる．これは，悲しみの感情が発声時に声道の長さや形状に変化を引き起こし，それが声道スペクトルに反映されている可能性がある．

具体的には，悲しみの感情においては声道が短くなる傾向があるため，高い周波数成分がより強調される可能性がある．このような声道スペクトルの異なりは，感情状態の認識において有用な特徴量となり得る．

これらの声道スペクトルの情報を使用することで，無感情と悲しみの音声を効果的に区別し，感情認識の精度向上に寄与することが期待される．

3.3　声道と声帯の関係

図 3.5 から悲しみの感情を含む音声と無感情の感情を含む音声の声道スペクトルには明確な違いが観察される．この声道の違いを利用して感情の認識を行うことが可能だが，その前に音声信号から声道を取り出すために必要な処理が存在する．

図 3.6 から明らかなように，声帯特性と声道特性の周波数成分は異なる構成を持っている．声帯特性は細かく変動する特性を持つため，高い周波数成分で構成されており，逆に，声道特性は滑らかに変動するため，低い周波数成分で構成されている．

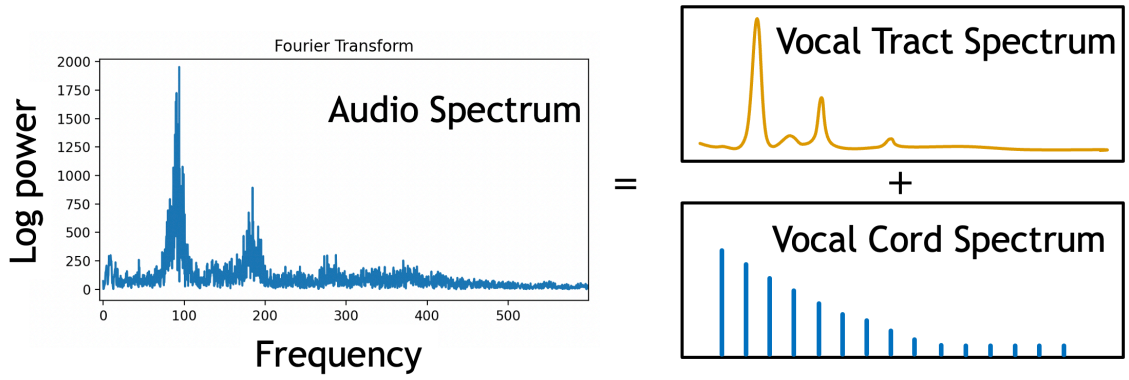

図 3.6: 音声スペクトルと声道スペクトルと声帯スペクトルの関係

ここで音声信号は，声帯特性と声道特性の畳み込みで表現されると考えることが可能である．時間領域における音声信号 $s(n)$ は，声帯の音源信号 $h(n)$ と声道のインパルス応答 $u(n)$ を用いて次の式 3.1 で表される．

$$s(n) = \sum_{m=-\infty}^{\infty} h(m)u(n-m) \tag{3.1}$$

畳込みは周波数領域において積の形として表されるため，式 3.1 に対してフーリエ変換を行うことで周波数領域に変換する．周波数領域における音声信号 $S(k)$ は，声帯特性 $H(k)$ と声道特性 $U(k)$ を用いて次の式 3.2 で表される．

$$S(k) = H(k) \times U(k) \tag{3.2}$$

音声信号 $S(k)$ のパワースペクトルは，次の式 3.3 で表される．

$$|S(k)|^2 = |H(k)|^2 \times |U(k)|^2 \tag{3.3}$$

対数パワースペクトルは，次の式 3.4 のように表される．

$$\log|S(k)|^2 = \log|H(k)|^2 + \log|U(k)|^2 \tag{3.4}$$

対数パワースペクトルでは，声帯特性と声道特性が和の形で表されることが式 3.4 から分かる．声帯特性は対数パワースペクトルの調波構造 (基本周波数とその倍音が生起する構造) として表現されるため，高周波成分に現れる．声道特性は周波数スペクトルの包絡として表されるため，低周波成分として現れる．

式 3.1 から式 3.4 までの処理を行うことで，声帯の音源信号と声道のインパルス応答に係る畳み込みが和の形に変換され，声帯特性と声道特性に分離が可能となる．これによって，音声信号から声道情報を抽出しやすくなり，感情の認識などが可能となる．

3.4　メル周波数ケプストラム係数

メル周波数ケプストラム係数 (以下，Mel Frequency Cepstrum Coefficient，MFCC) とは，音声信号処理や音声認識の分野で広く使用される特徴抽出手法の一つである [17]．MFCC は，声道の長さの違いによって生じる声道特性のパワーの違いを捉えるのに役立ち，音声信号をメル周波数スケールに変換し，その後ケプストラム解析を行う．

メル周波数スケールは，人間の聴覚特性に基づいているため，声道特性や声帯特性が影響を与える周波数帯域を効果的に表現できる．ケプストラム解析は，時間波形のスペクトルの対数のスペクトルを取ったものである [18]．

MFCC は音声信号の特徴抽出に広く使用され，異なる感情状態や発話者間の差異を捉えるのに有効である．これにより，声道の長さや形状の違いに基づく音声特性の変化を効果的に利用して感情の認識や音声処理のタスクに取り組むことが可能となる．具体的に音声信号から MFCC を抽出するプロセスについては，図 3.7 に記載している．

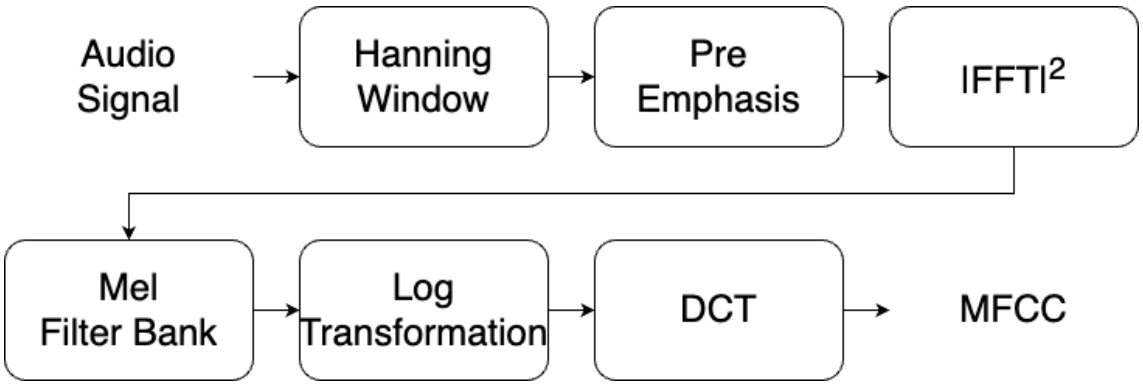

図 3.7: 音声信号から MFCC を抽出するまでのプロセス

次項から図 3.7 に使用されている各処理について説明を行う．

3.4.1　プリエンファシス

MFCC を取り出す準備として，まずは音声信号にプリエンファシスをかけていく．プリエンファシスは，音声信号の高域成分を強調するためのフィルタリング手法である [17] [19]．この手法は，高周波成分が低周波成分に比べて減衰することによって，音声信号の高域成分を強調し，後続の処理での特徴抽出を助けることができる．プリエンファシスを適用することにより，MFCC の取得前に音声信号を事前処理することが一般的である．プリエンファシスは以下の式 3.5 で表される．

$$pre[n] = s[n] - \alpha \cdot s[n-1] \tag{3.5}$$

ここで，pre$[n]$ はプリエンファシスが適用された音声信号，$s[n]$ は元の音声信号，α はプリエンファシスフィルタの強度係数である．一般的には，α は 0.9 から 1.0 の範囲の値が使われる (今回 α は，0.97 に設定) [17] [19]．

上記の式では，現在のサンプル $s[n]$ から前のサンプル $s[n-1]$ の 0.97 倍を引くことで，高域成分の強調が行わる．

このプリエンファシス処理によって音声信号中の不要な低周波成分が減衰され，高周波成分が強調されるため，後続のフレームごとに異なる周波数成分を持つ音声信号の特徴をより効果的に抽出することが期待される．

3.4.2　ハニング窓とフーリエ変換

次に，プリエンファシスが適用された信号 $pre[n]$ に窓関数 (今回はハミング窓を使用) を適用し，その結果を用いて高速フーリエ変換 (以下，Fast Fourier Transform，FFT) を行う．この処理は，フーリエ変換を用いて信号を周波数成分に変換するための重要な処理である [17] [19]．

窓関数を適用することで，信号を一定のフレームに区切り，そのフレームごとに異なる周波数成分を抽出することができる．ハミング窓は，信号の両端が滑らかになるような形状を持ち，フーリエ変換においてスペクトル漏れを抑えるためによく使用される．

適用された窓関数を持つ信号を $X[k]$ とし，その計算は以下の式 3.6 で表される．

$$X[k] = \sum_{n=0}^{N-1} pre[n] e^{-i\frac{2\pi kn}{N}} \quad k = 0, 1, 2, \cdots, N-1 \tag{3.6}$$

ここで，N は音声フレームのサンプル数であり，これはサンプリング周波数と窓の長さに依存する．k は周波数のインデックスを示している．この式は，離散フーリエ変換 (以下，Discrete Fourier Transform，DFT) を表し，FFT はこの DFT を高速かつ効率的に計算するアルゴリズムである．

この FFT によって得られる $X[k]$ は，周波数領域での信号の表現となる．これを用いて MFCC の計算が進められ，音声信号から特徴量を抽出する準備が整う．

3.4.3　メルフィルタバンク

次に，メルフィルタバンクについての説明を行う．メルフィルタバンクは，音声信号処理や音響信号処理において頻繁に使用されるフィルタバンクの一種であり，ヒトの聴覚に基づいたメル周波数スケールに線形な周波数スケールを変換する．

ヒトの聴覚は，すべての周波数に対して均等な感受性を持たず，非線形な性質を持っている．そのため，聴覚情報を効果的に表現するためには，この非線形性を考慮する必要がある．メル周波数は以下のような式 3.7 が一般的に利用される [19].

$$F = 1125 \ln\left(1 + \frac{f}{700}\right) \tag{3.7}$$

ここで，F はメル周波数，f はサンプリング周波数を窓長で割り，それに窓の数をかけたものである．この変換によって，下限周波数が 0，上限周波数がサンプリング周波数の半分になり，帯域間が 50%の重なりを持つようになる．

具体的には，メル尺度上で均等に分割された M 個の帯域があり，これを線形周波数のスケールに戻すためには以下の式 3.8 が利用される [19].

$$f = 700\left(\mathrm{e}^{\left(\frac{\mathrm{F}}{1125}\right)} - 1\right) \tag{3.8}$$

これにより，メルスケール上で均等に分割された周波数帯域が元の線形周波数のスケールに戻される．メルフィルタバンクは，このようなメル周波数スケールに基づいたフィルタバンクを構築し，音声信号をこの尺度に合わせた特徴量に変換する．この手法は音声の高次元特徴を取り扱いやすい形に変換し，後続の解析において有益な情報を抽出するのに効果的である．

図 3.8 に本研究で使用するメルフィルタバンクを示す．図 3.8 のメルフィルタバンクは 24 個の等間隔なメルフィルタが 16,000[Hz] のサンプリングレートで設計され，各フィルタが特定のメル周波数帯域に対応している．フィルタの帯域はナイキスト周波数と同値の 8,000[Hz] までとし，FFT サイズは 512 である．これにより，音声信号の重要な周波数成分を抽出しやすくなっており，人間の聴覚特性に基づいた効果的な特徴抽出が可能である．

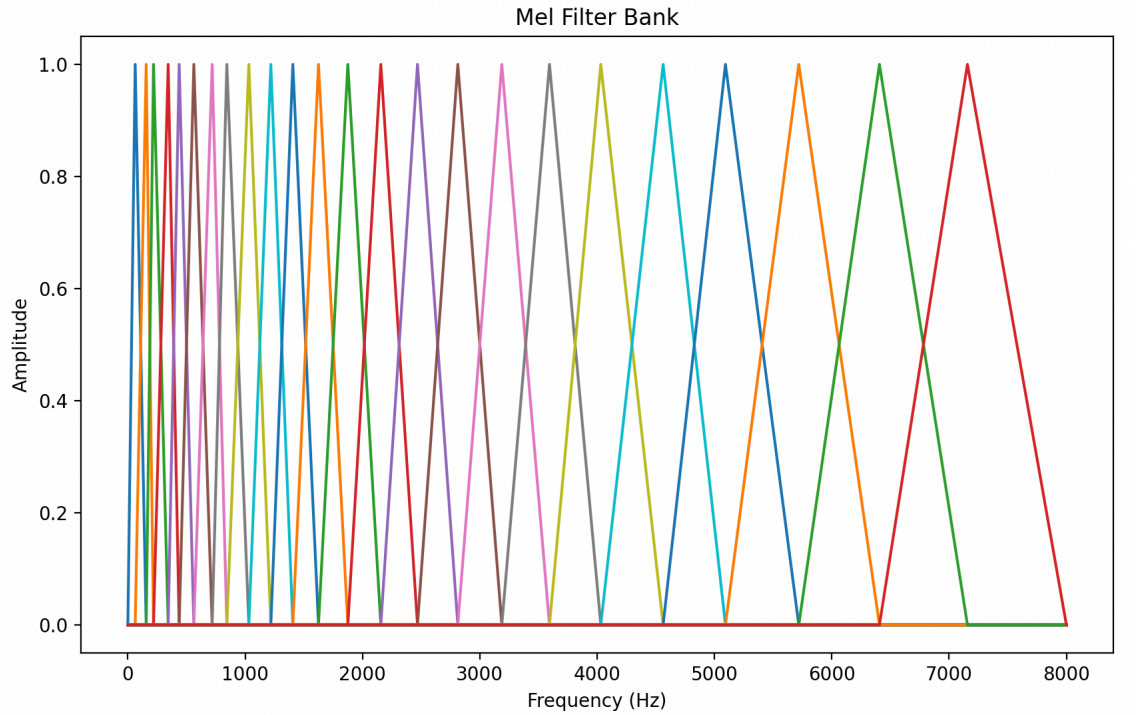

図 3.8: メルフィルタバンク

このメルフィルタバンクと $X[k]$ を内積することで，各メルフィルタバンクにおける周波数成分の強度を計算し，その情報を抽出する．

3.4.4　MFCC の抽出

式 3.9 は，MFCC を計算するための式であり，音声信号から得られたメルフィルタバンクの情報を基にして音声の特徴を抽出する [19]．

$$\mathrm{MFCC}[m] \;=\; DCT\left(\log_{10}\left(\sum_{k=0}^{P_m-1}|X_m[k]|^2\, f_m[k]\right)\right), \qquad 1 \;\leq\; m \;\leq\; M \quad (3.9)$$

ここで，$X_m[k]$ は m 番目の帯域における k 番目のフーリエ変換係数，$f_m[k]$ は m 番目の帯域における k 番目のメルフィルタ，P_m は m 番目の帯域における周波数成分の数である．

具体的には，m 番目の帯域における k 番目のフーリエ変換係数 $X_m[k]$ と，対応するメルフィルタ $f_m[k]$ を用いて，その帯域内の周波数成分の強度を表す値を計算する．

これらの情報は，帯域内の周波数成分のエネルギーを示しており，それを対数スケールで表現するために $\log_{10}$ が適用される．そして，この対数変換されたエネルギーに離散コサイン変換 (以下，Discrete Cosine Transform，DCT) が適用され，最終的に MFCC が得られる [19]．DCT はフーリエ変換を使う場合と比べて，低次元に圧縮でき，かつ係数間の相関を小さくできる性質を持っているメリットがある．

この DCT を適用することで，対数変換されたエネルギーをケプストラム (信号のフーリエ変換の対数をフーリエ変換したもの) にする．ケプストラムはスペクトルの滑らかな変動と細かな変動を分離することができる．

ここで，DCT の低次元部分は声道情報を表しており，逆に高次元部分は声帯情報を表しており，このことから，係数の値で声道を捉えることができる．

3.5　声道特徴量の抽出

提案手法では，音声信号を処理するために 50%のオーバーラップを持つ 20[ms] ごとの窓を音声信号に適用し，各窓ごとに MFCC を計算する．このとき，窓の数を L，MFCC の次元数を M として得られる MFCC は $L \times M$ の次元を持つ．

図 3.9 では，指定サイズの窓が音声信号から切り出されている様子が示されている．窓は 20[ms] ごとに配置され，隣り合う窓は 50%のオーバーラップを持っている．この窓ごとに音声信号を分割し，各窓に対して MFCC を計算する．

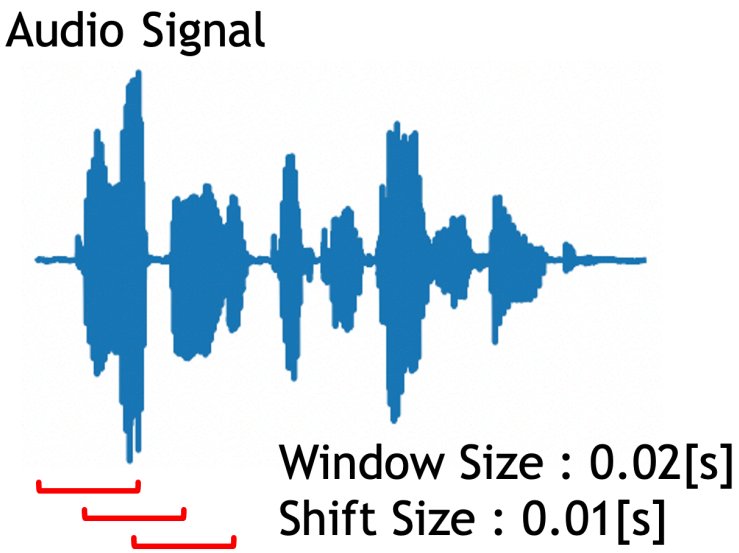

図 3.9: 指定サイズの窓による音声信号の切り出し

ここから，計算された MFCC の情報が SER に有効かどうかを調査する．各係数 m において，L からランダムに 10 個ずつのサンプルを取り出す．これらのサンプルが各係数で，MFCC の値がどのように分布しているかをプロットする．

ここまでの手法を「無感情ラベル付きの音声ファイル 70 個」と「悲しみのラベル付き音声ファイル 70 個」に対して同様のプロセスを繰り返す．これにより，無感情と悲しみの音声において，異なる感情状態が MFCC の値にどのように反映されるかを定量的に評価し，特定の係数が感情の特定に寄与する可能性を検討する．

図 3.10 および図 3.11 は，それぞれ提案手法で取得した MFCC のうち，1 から 12 番目までの係数 (低次元部分) と 13 から 24 番目までの係数 (高次元部分) における無感情データと悲しみデータの分布を示している．MFCC の低次元部分は声道情報を，高次元部分は声帯情報を含んでいるとされている．各図では，横軸が MFCC の値，縦軸がサンプルの数を表しており，無感情と悲しみの各音声データからランダムに選ばれたサンプルにおける MFCC の値の分布を可視化している．

これにより，各係数において感情ごとの MFCC の傾向や特徴がどのように異なるかを把握することが可能となる．声道の音響特性は通常，MFCC の低次成分に反映されやすいとされている．従って，係数 1-24 すべての MFCC の分布を使用せず，提案手法で使用するものとして必要な係数を選択する必要がある．適切な係数の選択方法については後述する．

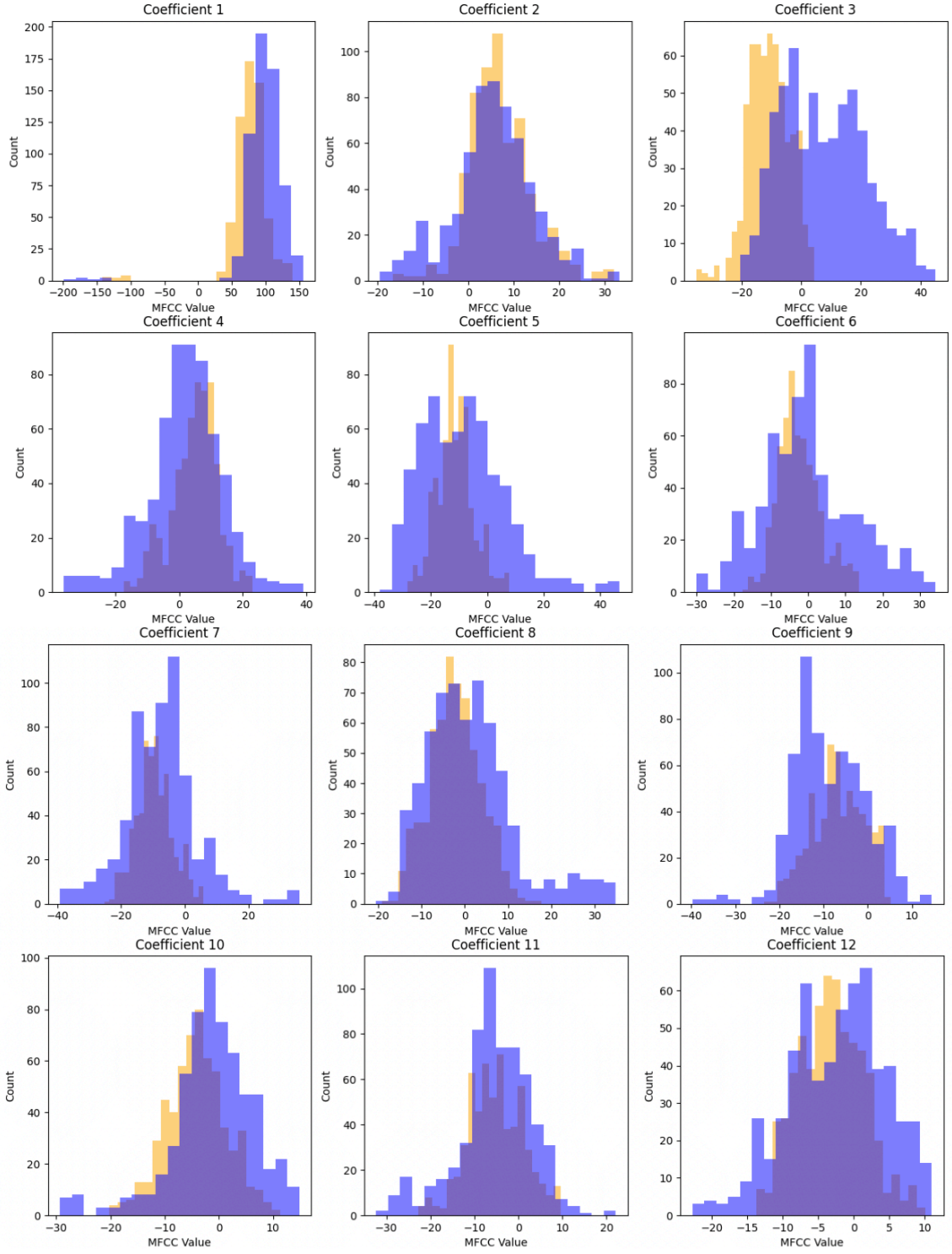

図 3.10: 24 次元分の無感情 (橙色) と悲しみ (青色) の MFCC の分布 (係数 1-12)

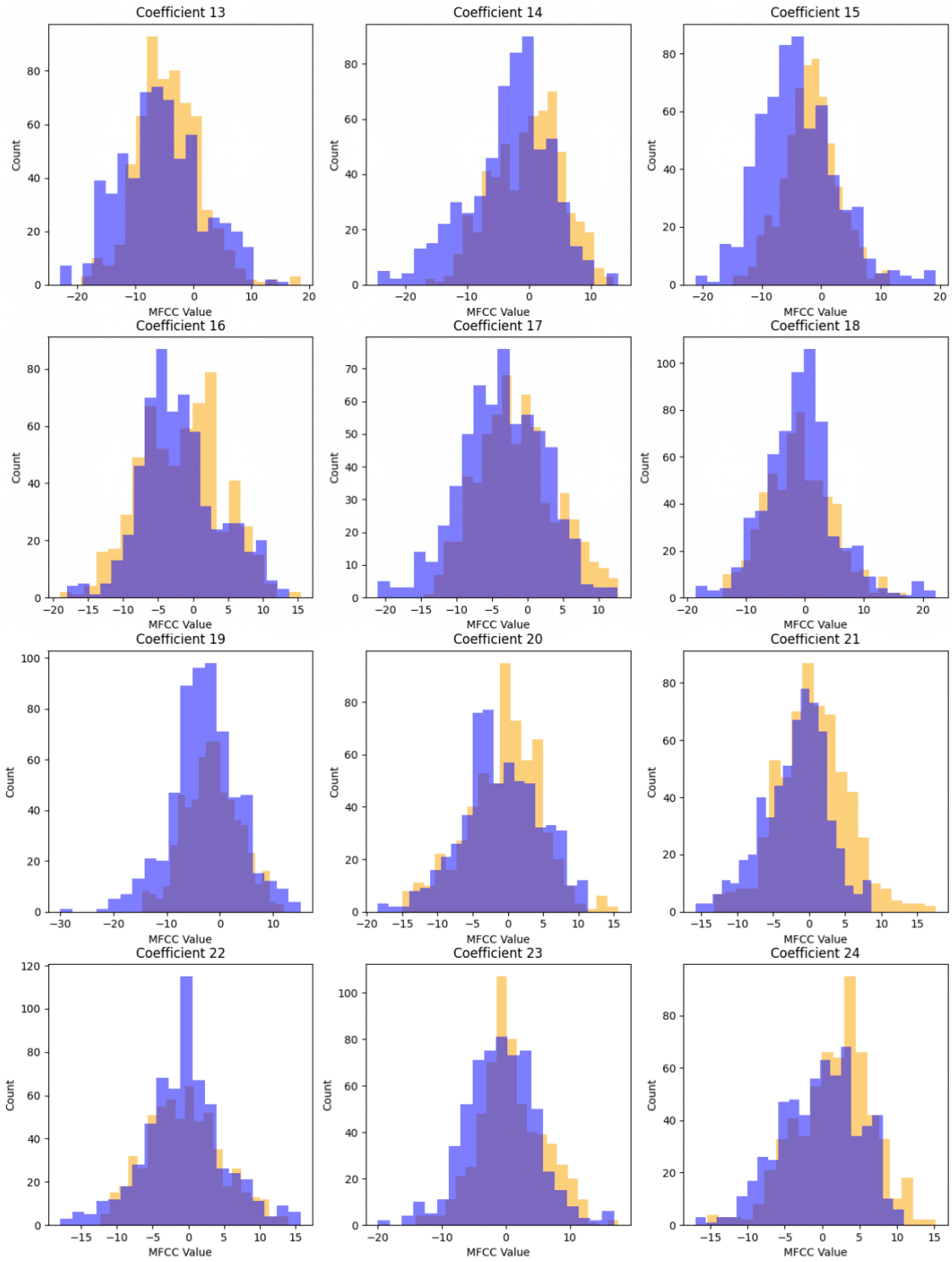

図 3.11: 24 次元分の無感情 (橙色) と悲しみ (青色) の MFCC の分布 (係数 13-24)

第4章　実験

本章では，評価方法に使用する尺度や，各特徴量の扱い方についてを示す．また，実際に「励起特徴量のみを用いた手法」,「声道特徴量のみを用いた手法」,「励起特徴量と声道特徴量を組み合わせた提案手法」の実験結果を提示し，各手法の認識精度の比較を行う．

4.1　評価方法

文献 [5] では，カルバック・ライブラー (以下，Kullback Leibler，KL) 情報量を使用して，特徴量の確率分布の差異を評価している．KL 情報量は，二つの確率分布の差を測る尺度であり，常に非負の値を取り得る．KL 情報量の値が小さいほど，二つの確率分布の類似性が高く，値が大きいほど類似性が低いと考えられる．

KL 情報量の具体的な計算を以下の式 4.1 に示す.

$$D = \frac{1}{2}\left(\mathrm{tr}\left(\Sigma_1^{-1}\Sigma_0\right) + (\mu_1-\mu_0)^T \Sigma_1^{-1}(\mu_1-\mu_0) - k - \ln\left(\frac{\det\Sigma_0}{\det\Sigma_1}\right)\right) \tag{4.1}$$

ここで，D は KL 情報量の値，k は分布の次元数，Σ_0 はデータ群 A の特徴対の共分散行列，Σ_1 データ群 B の特徴対の共分散行列，μ_0 はデータ群 A の特徴対の平均ベクトル，μ_1 はデータ群 B の特徴対の平均ベクトル，tr は正方行列における対角要素の総和，det は行列の行列式を表す.

具体的には，$\mathrm{tr}(\Sigma_1^{-1}\Sigma_0)$ は共分散行列の逆行列の積のトレースであり，共分散行列の差異に関連している．$(\mu_1-\mu_0)^T\Sigma_1^{-1}(\mu_1-\mu_0)$ は平均ベクトルの差を共分散行列の逆行列で重み付けた項であり，平均の差異に関連している．k は確率分布の次元数であり，分布の複雑さを表す．$\ln\left(\frac{\det\Sigma_0}{\det\Sigma_1}\right)$ は共分散行列の行列式の比の対数であり，分布の形状に関する情報を提供している．

4.2　声道特徴量の選定

「無感情と悲しみ」の認識に貢献できる係数を選出する手法として，KL 情報量を用いる．KL 情報量は二点間の距離を計測する際に，n 次元の次元ごとに距離 (二点間の差) の絶対値を求め，最後に全次元の値を合計する方法である．

これを無感情と悲しみの音声ファイル群における MFCC の分布の類似性を評価するために適用した．以下の図 4.1 は「無感情と悲しみ」の各係数における MFCC の分布の類似度を KL 情報量で表したものである．

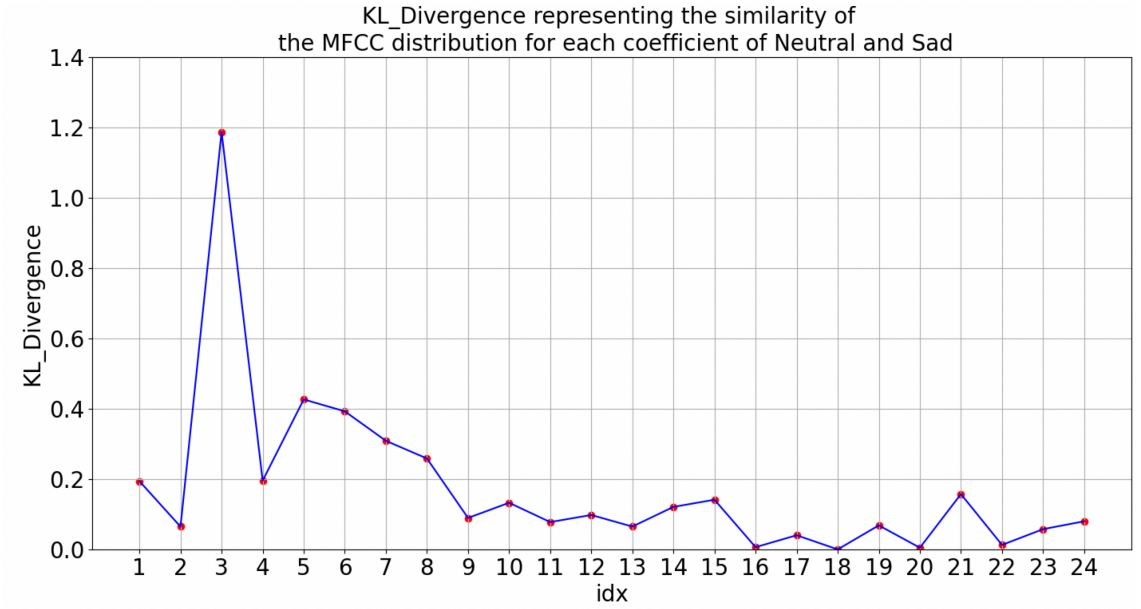

図 4.1:「無感情と悲しみ」の各係数における
MFCC の分布の類似度を表す KL 情報量

結果として，各係数において顕著な差異が観測され，特に MFCC の 3 番目，5 番目，6 番目の係数においてその差異が顕著であることが確認された (1 番目の係数は直流成分であるため除外)．3 番目，5 番目，6 番目の係数は MFCC の低次元情報に位置することから十分に声道情報が含まれている．また，声道特徴量として三つの係数を選択したのは，励起特徴量で使用する F0，SoE，EoE の三次元情報と合わせるためである．

したがって,「無感情と悲しみ」の認識に用いる声道特徴量として MFCC の 3 番目，5 番目，6 番目の係数を選択することは妥当であり，これを活用することで SER の性能向上が期待できる．この選択は，声道の特徴を効果的に捉え，異なる感情状態の音声を区別する上で有益な情報を提供するものと考えられる．

4.3　実験方法

4.3.1　二感情の認識実験

この実験では，主に「無感情と悲しみ」の認識精度が向上したかどうかを検証するために二感情を認識する実験を行う．二感情の認識実験についての手順を以下に示す．

1. 基準となる参照データを選定する．このデータは特定の感情を表すものであり，その感情の特徴を捉えるための基準となる．
2. 参照データと全く同じ感情のデータ群と比較を行い，KL 情報量の値をいくつか取り出す．
3. 同じ感情のデータ同士が比較された KL 情報量が複数抽出されるが，その中の最大値を閾値として採用する．この閾値は参照データにおける感情を反映し，他の感情との差異を示す指標となる．
4. 参照データと，認識対象となる試験データに対して KL 情報量を取り出す．
5. 参照データと試験データの KL 情報量が閾値を超えた場合，それは参照データの感情とは異なる感情であると評価される．

つまり，KL 情報量を利用し，閾値を超えるかどうかで感情の違いを検知する．

4.3.2　四感情の認識実験

四感情の認識実験についての手順を以下に示す．

1. 試験データと各感情データ群 (無感情，怒り，喜び，悲しみ) の各特徴量 (励起特徴量，声道特徴量) を取り出す．
2. 各特徴量について KL 情報量を取り出す．励起特徴量同士で求めた KL 情報量を $C1$，声道特徴量同士で求めた KL 情報量を $C2$ とする．
3. $C1$ と $C2$ の二次元情報を一次元情報にまとめるため，式 4.2 でノルムを求める．
4. 各感情から取り出されたノルムを計算し，それらの大きさを比較する．
5. ノルムが最小となる感情 i が試験データの感情の結果として出力され，認識される．

$$\|v_i\| = \sqrt{C1(i)^2 + C2(i)^2} \tag{4.2}$$

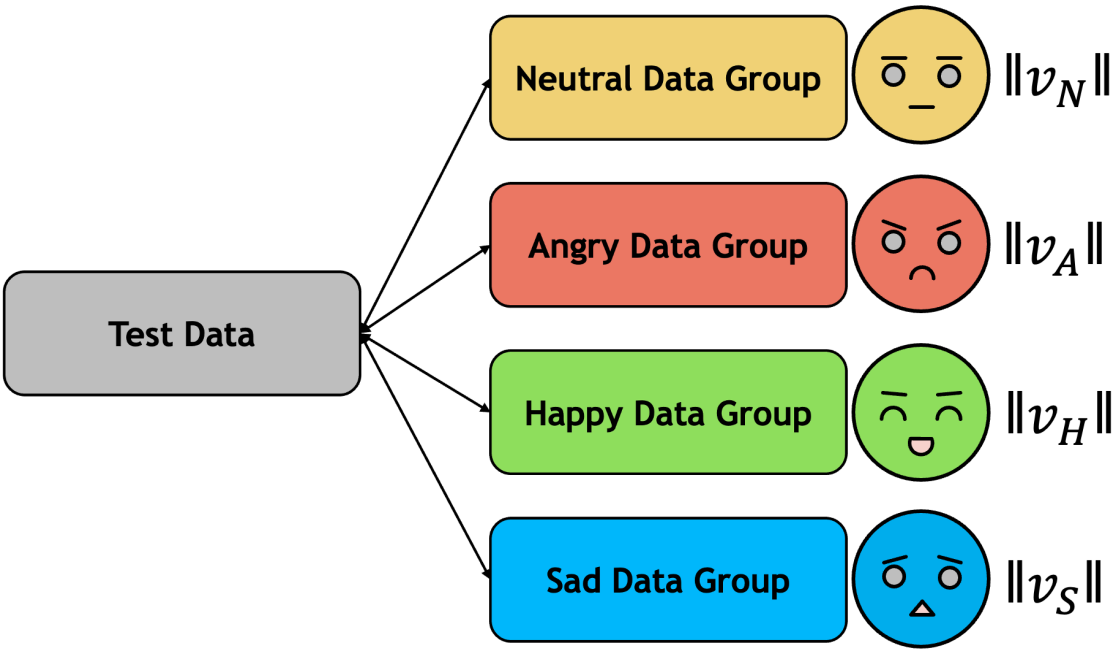

図 4.2: 感情の判定方法

ここで，$C1$ は励起特徴量同士で求めた KL 情報量を表し，$C2$ は声道特徴量同士で求めた KL 情報量を表す．$|v_i|$ は感情 i(N: 無感情，A: 怒り，H: 喜び，S: 悲しみ) に対応するノルムを表している．

以下の図 4.2 では，各感情のデータ群と試験データの比較について表されている．

図 4.2 のように取り出された四つのノルムは以下のように比較され，$\|v_N\|$，$\|v_A\|$，$\|v_H\|$，$\|v_S\|$ の中で一番値が小さいものをその感情の解答として，出力する．

- $|v_N| < |v_A|, |v_H|, |v_S|$ の場合，　試験データは「無感情」と判定．
- $|v_A| < |v_N|, |v_H|, |v_S|$ の場合，　試験データは「怒り」と判定．
- $|v_H| < |v_N|, |v_A|, |v_S|$ の場合，　試験データは「喜び」と判定．
- $|v_S| < |v_N|, |v_A|, |v_H|$ の場合，　試験データは「悲しみ」と判定．

4.3.3　データセット

本研究で使用するデータセットはドイツ語の分類問題用の感情音声データベース (EMO-DB) である [20].

EMO-DB は男女 5 人ずつ計 10 人の演者が 10 種類のドイツ語の短い文章を読み上げたものを録音したもので，無感情，怒り，喜び，悲しみの計四感情がある．

このデータセットの詳細については，以下の表 4.1 に記載する．また，各感情のデータ数については，以下の表 4.2 にまとめる．

表 4.1: 使用データセット

Number of People	10 Actors
Gender	Male 5 Actors, Female 5 Actors
Language	German
Kinds	Air Conduction Voice
Emotions	Neutral, Angry, Happy, Sad

表 4.2: データセット内のデータ数

Name	Gender	Neutral Data	Angry Data	Happy Data	Sad data
03	Male	11	14	7	7
08	Female	10	12	11	9
09	Female	8	13	4	4
10	Male	4	8	4	3
11	Male	8	8	8	7
12	Male	4	7	2	4
13	Female	9	6	10	5
14	Female	6	12	8	10
15	Male	10	7	6	4
16	Female	3	11	11	5

4.4　実験結果

従来手法 (励起特徴量のみ，声道特徴量のみ) と提案手法 (励起特徴量と声道特徴量) における二感情の認識結果を示す表 4.3 と表 4.4 と表 4.5 を比較した．これらの表では，各感情に対する認識結果が示されており，正答率 (CAR) がパーセンテージで示されている．

表 4.3: 従来手法 (励起特徴量のみ) における二感情の認識正答率 [%]

Reference data	Input Data	CAR	Reference data	Input Data	CAR
Neutral	Angry	99.2	Happy	Neutral	87.4
	Happy	95.8		Angry	77.8
	Sad	87.3		Sad	93.2
Angry	Neutral	93.3	Sad	Neutral	86.8
	Happy	47.2		Angry	97.6
	Sad	100.0		Happy	97.2

表 4.4: 従来手法 (声道特徴量のみ) における二感情の認識正答率 [%]

Reference data	Input Data	CAR	Reference data	Input Data	CAR
Neutral	Angry	84.6	Happy	Neutral	89.7
	Happy	86.1		Angry	80.5
	Sad	91.7		Sad	92.9
Angry	Neutral	85.9	Sad	Neutral	90.1
	Happy	56.1		Angry	90.4
	Sad	94.3		Happy	88.5

表 4.5: 提案手法 (励起特徴量と声道特徴量) における二感情の認識正答率 [%]

Reference data	Input Data	CAR	Reference data	Input Data	CAR
Neutral	Angry	98.4	Happy	Neutral	85.1
	Happy	96.4		Angry	74.2
	Sad	93.2		Sad	95.6
Angry	Neutral	96.1	Sad	Neutral	96.3
	Happy	51.3		Angry	92.1
	Sad	99.2		Happy	98.6

まず，従来手法 (励起特徴量) においては,「無感情と悲しみ」の感情認識正答率がそれぞれ 87.3%と 86.8%，従来手法 (声道特徴量) においては,「無感情と悲しみ」の感情認識正答率がそれぞれ 91.7%と 90.1%だった.

一方，提案手法では，これらの感情認識正答率がそれぞれ 93.2%と 96.3%に向上していることが確認された．特に「無感情と悲しみ」の感情認識において，提案手法が従来手法に比べて優れた性能を示している.

表 4.6: 従来手法 (励起特徴量のみ) における四感情の認識正答率 [%]

Input Data \ Result	Neutral	Angry	Happy	Sad
Neutral	63.38	0.0	6.22	30.40
Angry	1.02	83.67	15.31	0.0
Happy	9.86	36.62	53.52	0.0
Sad	25.86	0.0	0.0	74.14

表 4.7: 従来手法 (声道特徴量のみ) における四感情の認識正答率 [%]

Input Data \ Result	Neutral	Angry	Happy	Sad
Neutral	70.25	2.4	6.77	20.58
Angry	3.75	79.38	13.47	3.4
Happy	4.28	33.1	58.91	3.71
Sad	16.38	0.0	4.46	79.16

表 4.8: 提案手法 (励起特徴量と声道特徴量) における二感情の認識正答率 [%]

Input Data \ Result	Neutral	Angry	Happy	Sad
Neutral	90.14	0.0	5.63	4.23
Angry	1.02	77.55	21.43	0.0
Happy	1.41	25.49	70.28	2.82
Sad	3.45	0.0	0.0	96.55

従来手法 (励起特徴量のみ，声道特徴量のみ) と提案手法 (励起特徴量と声道特徴量) における四感情の認識結果を示す表 4.6 と表 4.7 と表 4.8 を比較した．

まず，従来手法 (励起特徴量) においては,「無感情と悲しみ」の感情認識正答率がそれぞれ 63.38%と 74.14%，従来手法 (声道特徴量) においては,「無感情と悲しみ」の感情認識正答率がそれぞれ 70.25%と 79.16%だった．

一方で提案手法では，これらの感情認識正答率がそれぞれ 90.14%と 96.55%に向上していることが確認された．特に「無感情と悲しみ」の感情認識において，提案手法が従来手法に比べて優れた性能を示している．

これらの結果により，声道特徴量を用いることで「無感情と悲しみ」の認識精度が向上することが分かった．

第5章　結論

5.1　まとめ

本研究では，励起特徴量を用いたSERにおける「無感情と悲しみ」の認識精度を向上させることを目的とし，これを達成するために励起特徴量と声道特徴量の組み合わせを提案した．

第1章では，近年におけるSERの需要，背景，問題点について言及した．第2章では，「励起特徴量を用いたSER」における利点と問題点について抜き出し，励起特徴量が感情認識にどのような効果をもたらすかについて述べた．第3章では，「励起特徴量を用いたSER」に声道特徴量を加えた提案手法について示し，声道から得られる特徴量の有用性について示した．第4章では，音声コーパス「EMO-DB」を用いて，従来手法と提案手法との比較実験を行った．

総じて，励起特徴量と声道特徴量の双方を利用することで「無感情と悲しみ」における感情認識の性能が向上し，本研究の目的を達成した．

しかし，感情認識の手法において，励起特徴量のみ，声道特徴量のみ，および励起特徴量と声道特徴量の組み合わせを用いた場合，いずれにおいても「怒りと喜び」の認識精度が大きく改善しなかった結果が得られた．この結果は，励起特徴量が「怒りと喜び」の感情を区別するために十分な情報を提供できない可能性があることを示唆している．つまり，従来の特徴量だけではこれらの感情の微細な違いを捉えることが難しい可能性がある．

5.2　今後の展望

この課題を解決する今後の展望として，新たな特徴量の検討する必要があると考える．他の特徴量や深層学習に基づく特徴量抽出手法を検討し，この手法が感情認識においてどのような効果をもたらすかを評価することが重要である．

「怒りと喜び」の認識問題を解決し，四感情すべての効果的な感情認識手法の構築を目指すことが今後の研究の課題となる．

謝辞

本研究を進めるにあたり，杉田泰則准教授からは貴重なご指導をいただきました．その温かいサポートと専門的なアドバイスに感謝申し上げます．

さらに，本研究の審査に携わっていただいた岩橋政宏教授，圓道知博教授，原川良介准教授には，熱心なご指摘とご提案をいただきました．そのおかげで論文の質が向上し，深化することができました．心より感謝申し上げます．

最後に，研究室のメンバーや友人たちへも感謝の意を表します．ゼミでの活発な議論や意見交換は，私の研究に新しい視点をもたらしました．皆の協力と励ましのおかげで，研究がより一層充実したものとなりました．ありがとうございました．

付 録A　MFCCの分布の類似度

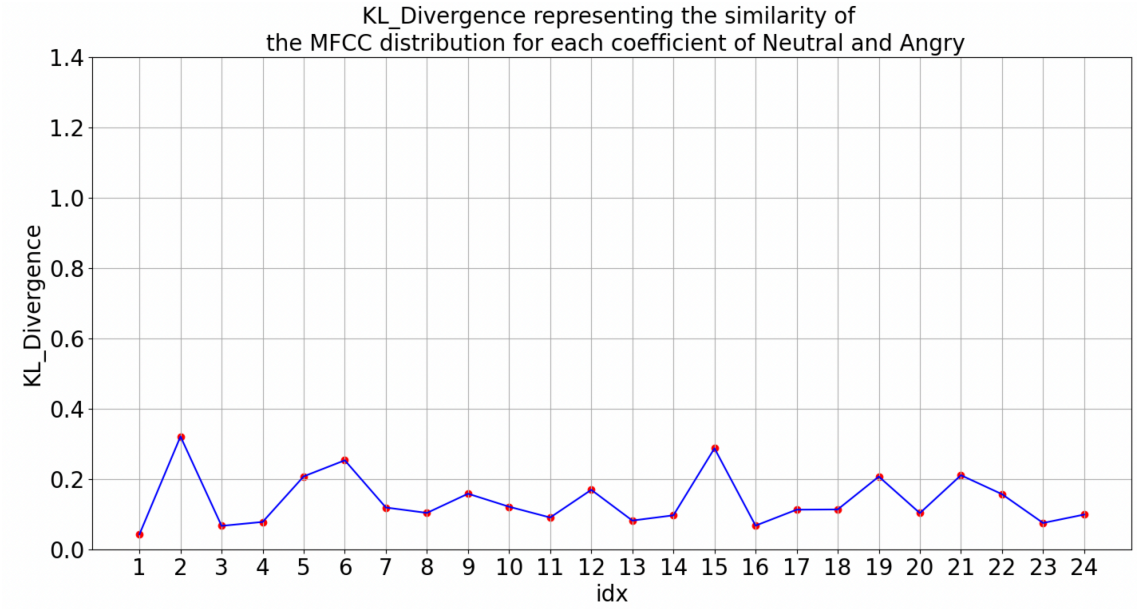

図 A.1: 「無感情と怒り」の各係数における
MFCC の分布の類似度を表す KL 情報量

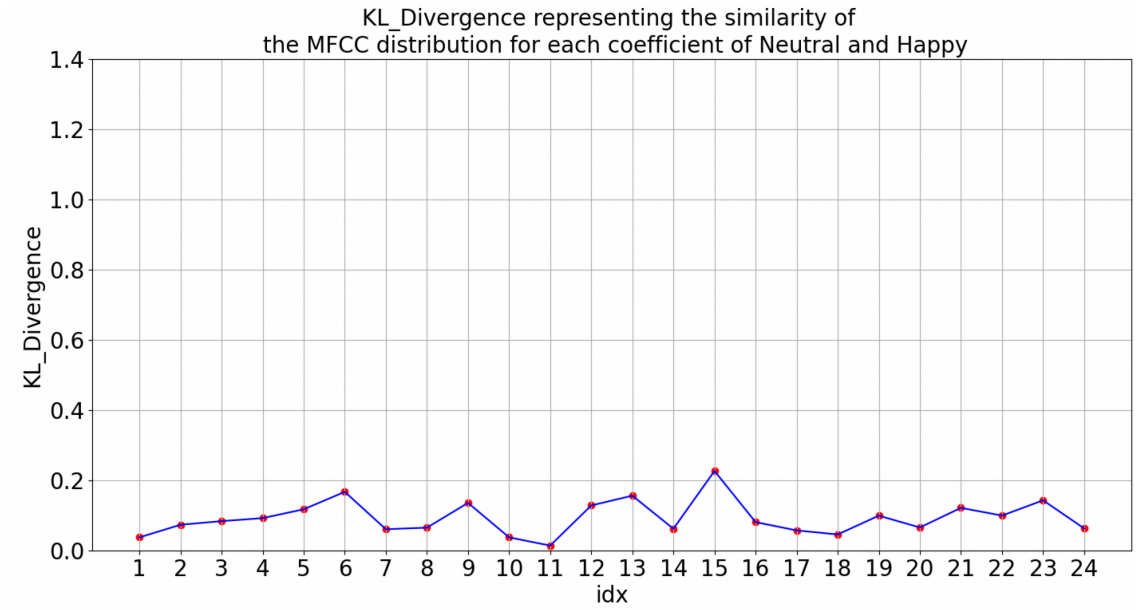

図 A.2: 「無感情と喜び」の各係数における
MFCC の分布の類似度を表す KL 情報量

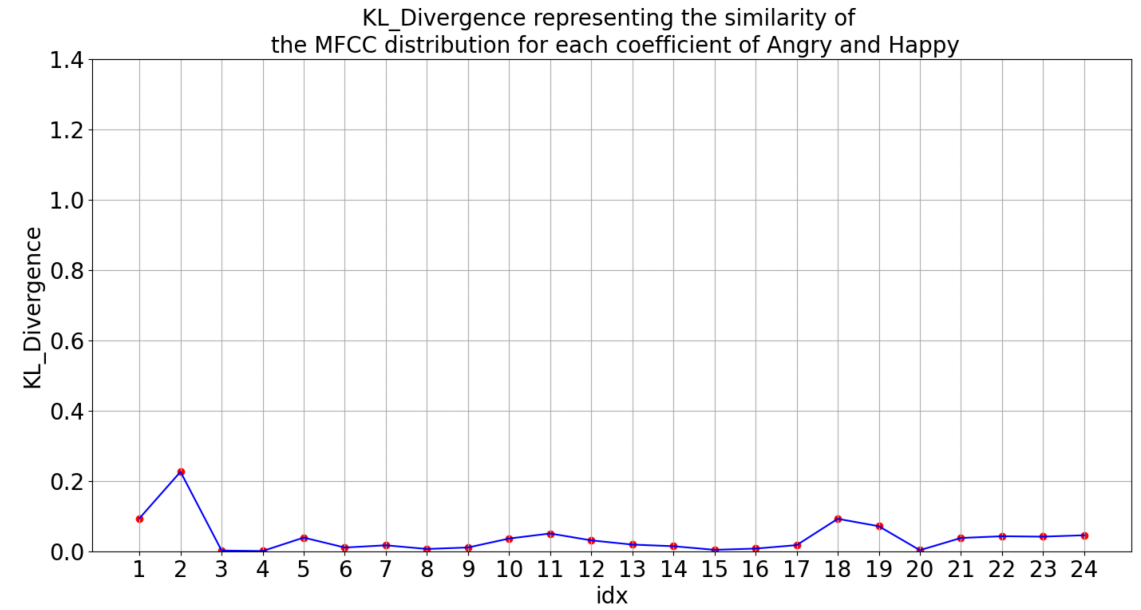

図 A.3: 「怒りと喜び」の各係数における
MFCC の分布の類似度を表す KL 情報量

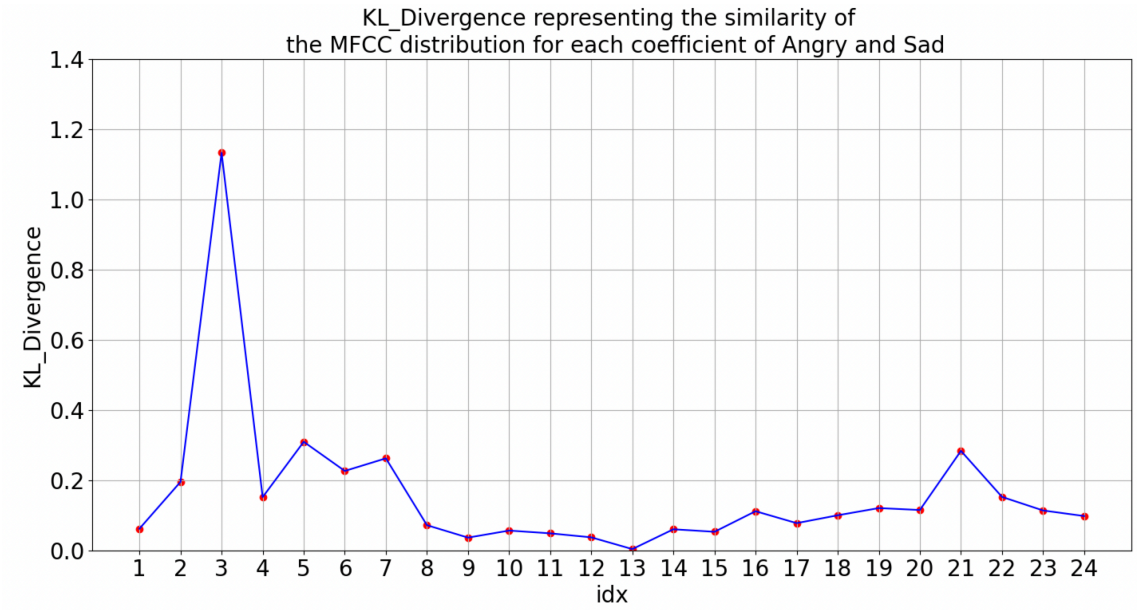

図 A.4: 「怒りと悲しみ」の各係数における
MFCC の分布の類似度を表す KL 情報量

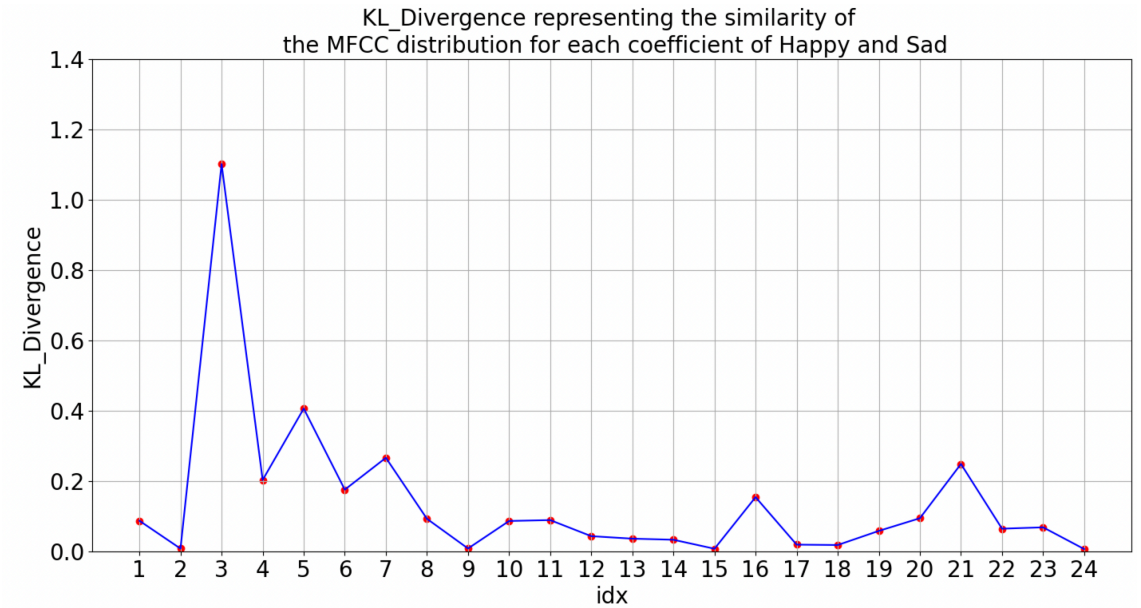

図 A.5: 「喜びと悲しみ」の各係数における
MFCC の分布の類似度を表す KL 情報量

参考文献

[1] Rohit Raj Sehgal, Shubham Agarwal, Gaurav Raj, "Interactive Voice Response using Sentiment Analysis in Automatic Speech Recognition Systems", p.662, ICACCE, june, 2018

[2] Promod Y, Abhay K, Suraj T, Chirag S, Sibsambhu K, Jithendra V, "Speech Emotion Recognition Using Spectrogram and Phoneme Embedding", pp.3688-3690, Interspeech, september, 2018

[3] Swarna K, H.D.Vankayalapati, R.S.Vaddi, K.R.Anne, "A comparative analysis of classifiers in emotion recognition through acoustic features", pp.402-405, Int J Speech Technol, December, 2014

[4] Jahangir R, Wah T, Hanif F, Mujtaba G, "Deep learning approaches for speech emotion recognition: state of the art and research challenges", p.23746-23748, Multimedia Tools and Applications, july, 2021

[5] Sudarsana Reddy Kadiri, and Paavo Alku, "Excitation Features of Speech for Speaker Specific Emotion Detection", pp.60382-60389, IEEE Access, March 5, 2020

[6] R.Surender Reddy, B.Kiran Kumar, B.Eshwar, "Excite Comparison of Emotional Speech Detection Algorithm for Glottal Closure Instants", p.2, IEEE International, May, 2020

[7] Sri Rama Murty Kodukula, "Sigbficance of Excitation source information for speech analysis", pp.49-56, Interspeech, March, 2009

[8] R.Surender Reddy, B.Kiran Kumar, B. Eshwar "Comparison of Glottal Closure Instants Detection Algorithms for Emotional Speech", p.1, IEEE, January 5, 2020

[9] 高島遼一, "Python で学ぶ音声認識", pp.74-75, 株式会社インプレス, May 21, 2021

[10] K. T. Deepak, S. R. M. Prasanna "Epoch Extraction Using Zero Band Filtering from Speech Signal", pp.2311-2312, Springer Science+Business Media New York 2014,December 25, 2014

[11] Gregor Pirker, Michael Wohlmayr, Stefan Petrik, Franz Pernkopf ”A Pitch Tracking Corpuswith Evaluationon Multipitch Tracking Scenario”, p.1, Interspeech, 2011

[12] 森大毅, 中村真, ”音声研究における感情の位置付け”, p.465, 日本音響学会誌 71 巻 9 号, 2015

[13] Yongwei Li, Ken-Ichi Sakakibara, Daisuke Morikawa, Masato Akagi, ”Commonalities of Glottal Sources and Vocal Tract Shapes Among Speakers in Emotional Speech”, p.1, ISSP, September 11, 2018

[14] JangwonKim, AsteriosToutios, SungbokLee, Shrikanth S.Narayanan, ”Vocal tract shaping of emotional speech”, pp.6-11, Computer Speech Language, p.1, April 16, 2020

[15] Surabhi Vaishnav, Saurabh Mitra, ”Emotion Recognition Using Vocal Tract Parameters and Artificial Neutral Networks”, pp.56-57, IJESIRD, July, 2016

[16] Adam C.Lammert, Shrikanth S.Narayanan, ”On Short-Time Estimationof Vocal Tract Length from Formant Frequencies”, p.12, PLOS ONE, July 15,2015

[17] ZRAR KH. ABDUL, ABDULBASIT K. AL-TALABANI, ”Mel Frequency Cepstral Coefficient and Its Applications”, pp.122136-122139, IEEE Access, November 23, 2022

[18] Alan V. Oppenheim Ronald W. Schafer, ”From Frequency to Quefrency: A History of the Cepstrum”, IEEE SIGNAL PROCESSING MAGAZINE, p.95, August 30, 2004

[19] J. Ancilin, A. Milton, ”Improved speech emotion recognition with Mel frequency magnitude coefficient”, pp.2-4, ScienceDirect, August, 2021

[20] PIYUSH AGNIHOTRI, EMO-DB, https://www.kaggle.com/datasets/piyushagni5/berlin-database-of-emotional-speech-emodb