時間周波数表現と深層学習を用いた病的音声の声質自動評価に関する研究

日髙, 駿介

時間周波数表現と深層学習を用いた病的音声の声質自動評価に関する研究

A Study on Automatic Pathological Voice Quality Evaluation Using Time-Frequency Representations and Deep Learning

日髙　駿介
HIDAKA Shunsuke

2023年6月

目次

第 1 章

序論

1.1 音声障害

社会性動物であるヒトは，日常の至る所でコミュニケーションを行う。特に，音声言語を母語とするヒトにとって，音声によるコミュニケーションは非常に重要な役割を果たす[*1]。音声コミュニケーションでは，文字に直接書き起こせる言語情報のみならず，感情や態度，意図，話者の個人性といった多様な情報が伝達される [1]。加えて，音声コミュニケーションは即時的であり，ごく短時間の間に，これらの多様な情報がやり取りされる。

このように，音声コミュニケーションは社会関係の円滑な構築に大きく寄与している。しかし，逆に言えば，音声コミュニケーションに強く依存していたヒトが，何らかの障害で音声コミュニケーションが自在に行えなくなると，quality of life（QOL）[2] の低下に繋がる場合がある [3–6]。American Speech-Language-Hearing Association（ASHA）による分類 [7] に基づくと，音声コミュニケーションに支障を来す障害は，発語障害（speech disorder），言語障害（language disorder），聴覚障害（hearing disorder），聴覚情報処理障害（central auditory processing disorders）に分けられる。また，発語障害はさらに，音声障害（voice disorder），構音障害（articulation disorder），流暢性障害（fluency disorder）に分けられる[*2]。本論文では，発語障害のうち，音声障害に関する内容を扱う。

[*1] ろう者や盲ろう者を含む場面等では，手話言語や触覚言語が非常に重要な役割を果たす場合もある。

[*2] 英名に対応する和訳は，文献によって異なる場合がある。例えば，“voice disorder” は，近年では「音声障害」と訳すことが定着しているが [8, 9]，かつては「声の障害」 [10, 11] 等と訳されたこともあった。その一方で，“speech disorder” に対する定着した訳はまだ存在しないが，本論文では，日本音声言語医学会によって出版された執筆時点で最新の文献で用いられた，「発語障害」 [12] という訳を用いた。なお，ここでは，1993 年の ASHA による分類 [7] を示したが，音声障害を発語障害に含めない考え方 [12] や，発語運動障害（dysarthria）は発語障害の中で独立した障害として扱うべきであるという指摘がある [12, 13] ことには注意が必要である。また，発語障害という訳は，音声障害を発語障害に含めない立場で使われたものである。いずれにせよ，障害の分類や和訳は，知見の蓄積によって時代と共に変遷しうるものであり，将来的に，ASHA による分類 [7] が修正される可能性もあると考えられる。

音声障害は,「音質，声の高さ，声の大きさ，発声努力などの変化により，コミュニケーションを損なう，あるいは声の QOL が低下すること」 [9,14] として定義される[*3]。音声障害は，病因に基づき疾患分類が行われる。ASHA および日本音声言語医学会のガイドラインでは，疾患分類のために以下の大分類を示している [9,15]。

- 喉頭の組織異常（structural pathologies of the larynx）
- 喉頭の炎症性疾患（inflammatory conditions of the larynx）
- 喉頭の外傷（trauma or injury of the larynx）
- 全身性疾患（systemic conditions affecting voice）
- 音声障害を来す呼吸器・消化器疾患（non-laryngeal aerogestive disorders affecting voice）
- 心理的疾患・精神疾患（psychiatric and psychological disorders affecting voice）
- 神経疾患（neurologic disorders affecting voice）
- その他の音声障害（機能性発声障害を含む）（other disorders affecting voice）
- 原因不明の音声障害（voice disorders: undiagnosed or not otherwise specified (NOS)）

上記の分類からもわかるように，音声障害を引き起こす病因は数多く存在し，決して少なくない人数が音声障害に罹患することが報告されている。例えば，アイオワ州とユタ州の成人 1,326 名を対象とした，アンケートに基づく音声障害の疫学調査では，音声障害の生涯有病率は 29.9%，調査時に音声障害を罹患していた割合は 6.6% で，就業中の回答者の 7.2% が過去 1 年間に一日以上，声が原因で仕事を休んだことが報告されている [16]。また，韓国の成人 19,039 名を対象とした，喉頭鏡検査に基づく器質的喉頭疾患（発声に関する構造や神経の異常）の疫学調査では，調査時に器質的喉頭疾患を罹患していた人の割合は 7.1% であったことが報告されている [17][*4]。

音声障害患者は，しばしば社会的孤立や，抑うつ，不安，欠勤，賃金の減少，ライフスタイルの変化等に悩まされ，QOL や労働生産性が低下することがある [14,16,18–21]。また，致死的疾患である喉頭癌や指定難病であるパーキンソン病等の疾患は，早期発見することが極めて重要な疾患であるが，これらの疾患では，初期症状として音声障害が現れる場合があることが知られている [14,22,23]。これらの事実からも裏付けられるように，音声障害の診断と治療は，QOL の改善やさらに深刻なリスクの回避のために重要である。

[*3] 原典は American Academy of Otolaryngology-Head and Neck Surgery（AAO-HNS）によるガイドライン [14] であり，和訳は日本音声言語医学会と日本喉頭科学会によるガイドライン [9] に従った。なお，AAO-HNS によるガイドライン [14] では，音声障害を表す単語として “dysphonia” を用いている。

[*4] 元の論文 [17] では，喉頭疾患ごとの罹患率が示されていたが，ここでは，すべての罹患率の和を器質的喉頭疾患の罹患率として表示した。

1.2 病的音声の声質評価

一般に，音声障害患者によって発せられた音声を病的音声（pathological voice）と呼び，対比的に，音声障害患者ではないヒトによる音声を健常音声（healthy voice）と呼ぶ。健常音声と比較した病的音声の特徴は，嗄声（hoarseness）と呼ばれる病的な声質異常を有することである。嗄声は「特に音色の異常のために，音声の聴取者の意識上に「嗄れている」「しわがれている」という聴覚印象を励起する性質，またはその性質を有する声」[24] として定義される。一口に嗄声と言っても，その性質や程度は，音声障害の病因の種類や性状に依存して大きく異なる。そのため，病的音声の声質，すなわち嗄声の評価は，音声医学の臨床における重要な診断項目になっている。しかし，一次元的である声の高さや大きさとは対照的に，多次元的である声質（音色）を包括的に評価することは容易ではなく，その評価法も一意に定まるものではない。病的音声の他覚的な声質評価法は，主観的手法である聴覚心理的評価（auditory-perceptual evaluation）と，客観的手法である音響分析（acoustic analysis）に二分される[*5]。本節では，これらの 2 つの手法を順に説明する。

1.2.1 聴覚心理的評価

聴覚心理的評価は，医師や言語聴覚士が患者の声を聴き，聴覚印象に基づき声質を主観評価する声質評価法である。音声はヒト同士のコミュニケーションにおいて使われるものであり，「声質とは本質的に知覚的なものである（voice quality is fundamentally perceptual in nature）」[25] ことから，ヒトが実際に聴いてどう感じるかという知覚的評価は重要である。そのため，聴覚心理的評価は，音声医学の臨床でゴールドスタンダードとして用いられている。また，聴覚心理的評価のその他の意義としては，音声障害の病因推定に役立つこと，音声障害の経過や治療効果の観察に役立つこと，非侵襲的であり患者へ心身上の苦痛を与えないこと等が挙げられている [26–28]。

聴覚心理的評価法としては，日本音声言語医学会によって提案された GRBAS 尺度（the GRBAS scale）[9, 26–29] が国内外で広く用いられている。GRBAS 尺度は，以下の 5 項目を定量評価する手法である [9, 28]。

- 項目 G（嗄声度; grade of hoarseness）：嗄声の総合的な程度
- 項目 R（粗糙性; roughness）：がらがら，ごろごろと表現できる聴覚印象

[*5] 他覚的な声質評価法以外の声質に関わる評価法としては，音声障害患者自身による，声質を含めた音声障害が生活を与える影響に対する自覚的評価法がある [3, 4]。

- 項目 B（気息性; breathiness）：息漏れがある聴覚印象
- 項目 A（無力性; asthenia）：声の弱々しさ
- 項目 S（努力性; strain）：気張っていかにも無理をして発声している聴覚印象

各項目は，その聴覚印象の程度に応じて，0，1，2，3 の 4 段階で評価される [9]。

- 0：正常・なし（normal or absence of deviance）
- 1：軽度（slight deviance）
- 2：中等度（moderate deviance）
- 3：重度（severe deviance）

評価結果は，例えばすべての項目が 0 点であった場合，「G0R0B0A0S0」のように記載する [9]。GRBAS 尺度の評価対象の発話内容としては，日本音声言語医学会では持続母音（sustained vowel）の使用が推奨されてきたが [9, 26–28]，近年では，持続母音に加えて，短文音読のような連続発話（continuous speech; running speech; connected speech）を用いることについて検討した研究も行われている [30]。

項目 G は嗄声の総合的な程度を表すのに対し，他の項目 R，B，A，S はより具体的な声質に対応する。項目 R，B，A，S は，意味微分法（semantic differential method; SD 法）を用いた嗄声の聴覚的因子の抽出結果 [31] を参考に，日本音声言語医学会発声機能検査法委員会の聴覚心理的検査小委員会で策定されたものである [27]。喉頭疾患と GRBAS 尺度の間には関連があり，典型的な例としては，ポリープ様声帯では粗糙性が強く（項目 R の評点が高く），片側声帯麻痺では気息性が強く（項目 B の評点が高く），低緊張性発声障害では無力性が強く（項目 A の評点が高く）過緊張性発声障害では努力性が強く（項目 S の評点が高く）なることが知られている [32]。但し，同じ疾患名であっても，疾患の性状の違いによって必ずしも同じ様な声質になるとは限らない。例えば，声帯ポリープに関して言えば，ポリープの部位や大きさ，形態，硬さの違いによって，粗糙性が強い（項目 R の評点が高い）場合もあれば，気息性が強い（項目 B の評点が高い）場合もある [32]。

上述のように，聴覚心理的評価と GRBAS 尺度には様々な意義があり，音声医学の臨床に則した声質評価法となっている。しかし，評価が評者の主観に基づくものである性質上，その評価が再現性に欠けてしまうという問題が必然的に生じる。特に，同じ音声に対する評価結果に評者間の差異が生じることを評者間変動（inter-rater variation），同じ評者が同じ音声を繰り返し評価した時に，評価結果に差異が生じることを評者内変動（intra-rater variation）と呼ぶ。評者間変動が大きくなる要因としては，評者の臨床経験や所属集団 [33–35]，職業 [36, 37]，熟練度 [33, 36] の違い等が存在することが示唆されている。また，評者内変動が大きくなる要因としては，臨床の経験年数の少なさ [36] や，記憶や注意力の低下等の内部要因や，音響的文脈や聴取課題等の外部要因があると考え

られている [25, 34, 38]。また，GRBAS 尺度の各項目に関する話題としては，項目 A と S が他の項目と比べて信頼性が低くなることが報告されている [36, 39]。このことから，GRBAS 尺度に影響を受けて考案された RBH 尺度（the RBH scale）[40] では項目 A と S が，consensus auditory-perceptual evaluation of voice（CAPE-V） [41] では項目 A が定量評価項目から除外されている。なお，少し話題が逸れるが，CAPE-V は，ASHA によって提案された聴覚心理的評価法であり，論文執筆時点で，日本では GRBAS 尺度の方が定着しているものの，日本語化した CAPE-V を普及するための取り組みが現在進行形で行われている [42]。

1.2.2 音響分析

先述したように，聴覚心理的評価では，評者の主観に起因する再現性の欠如が問題になる。この再現性の欠如の問題に対処するための声質評価法として，音響分析と呼ばれる手法を用いることがある。音響分析は，録音した患者の音声波形データに対して信号処理を施し，嗄声に関連する音響特徴量を算出することによって声質を定量評価する手法である。音響分析の長所は，客観的かつ再現性のある評価を行えることである。古典的であるが現在も広く使われている音響特徴量としては，周期揺らぎを表すジッタ（jitter），振幅揺らぎを表すシマ（shimmer），倍音対雑音比（harmonics-to-noise ratio）があり，これらの特徴量は音声の準周期性の仮定に基づいて計算される。これらの他にも，離散フーリエ変換比（discrete Fourier transform ratio） [43] 等のスペクトルに基づく特徴量，平滑化ケプストラル・ピーク卓越度（smoothed cepstral peak prominence; CPPS） [44] 等のケプストラムに基づく特徴量，変調スペクトル（modulation spectrum） [45] や，非線形力学のカオス理論 [46, 47] に基づく特徴量等，様々な音響特徴量が提案されている。

しかし，音響分析は客観的かつ再現性のある評価を行えることが長所であるものの，必ずしも臨床で広く普及しているとは言えない。その原因としては，「操作が煩わしい，コストがかかる，音響分析技術の達成度がまだ不足である，国際的にも国内でも標準化されていない」 [48] といったことに加えて，音響分析の結果の解釈がしばしば難しいことも挙げられている [49]。音響特徴量の値と聴覚心理的評価の評点の関係について言えば，両者の高い相関を報告した研究 [50] もあれば，両者の関係は単純ではないと報告した研究 [51] もあり，音響特徴量の分析結果と主観的印象としての声質を結びつけることは容易でない。また，ジッタやシマ，倍音対雑音比等の古典的な音響特徴量の計算は，音声波形の準周期性を仮定した基本周波数推定に依存しているため，重度の嗄声に対し正確な分析が難しいという問題がある [46, 52]。その他に，CPPS のような一部の音響特徴量を除いた多くの音響特徴量では，主として母音の定常部あるいは有声区間を解析対象としているが，重度の嗄声では有声区間と無声区間を区別することがしばしば容易ではない [44]。これら

の理由から，音響分析は，聴覚心理的評価に対して補助的な声質評価法となっている。

1.3　聴覚心理的評価の評点の自動推定

前節で述べたように，聴覚心理的評価と音響分析には，どちらも問題があるが，両者の長所を組み合わせることによって，これらの問題は解決しうる。具体的には，聴覚心理的評価の評点を，音響分析のように計算機を使って自動推定すれば，直観的かつ再現性のある声質評価を実現することができる。

これまでの多くの研究では，音響分析に関する知見を援用した，音響特徴量と機械学習（machine learning）を組み合わせたアプローチが用いられてきた。このアプローチは，単一の音響特徴量だけでは聴覚心理的評価の評点を表現するのは難しい [51] ため，複数の音響特徴量を組み合わせることによってこの問題を解決するという考えに基づいている [53]。機械学習手法としては，線形回帰（linear regression） [53, 54]，学習ベクトル量子化（learning vector quantization） [55]，サポートベクターマシン（support vector machine）[56,57], エクストリームラーニングマシン（extreme learning machine）[56,58], 比例オッズモデル（proportional odds model）[58], 混合ガウスモデル（Gaussian mixture model） [59, 60]，ディープビリーフネットワーク（deep belief network） [60]，多層パーセプトロン（multilayer perceptron） [61] 等が使用されてきた。しかし，音響分析で使われる音響特徴量には，第 1.2.2 節で述べた様々な問題があることに加えて，次元を大幅に削減する処理である特徴量抽出の過程で，音声信号の原波形（raw waveform）に含まれている重要な情報が失われてしまう可能性がある。そのため，音響分析で使われる音響特徴量に基づくアプローチでは，推定精度に性能限界が存在すると考えられる。

このアプローチに対して，近年では，原波形 [49] や対数振幅メル周波数スペクトログラム（log-amplitude mel-frequency spectrogram）[62]，ケプストログラム（cepstrogram）[63] のような，高度に人の手で設計されていない（not highly hand-crafted）特徴量と，深層学習（deep learning）を組み合わせたアプローチを使った研究も行われている。深層学習は，非線形数理モデルである Deep Neural Network（DNN）を最適化する機械学習手法であり，特徴表現学習（feature learning）[64] を実現可能であることが大きな特徴の一つである。特徴表現学習とは，高度に人の手で設計されていない特徴量（究極的には原波形）から，目的とする課題のために必要な特徴表現を，数理モデル内で自動的に獲得することを意味する。高度に人の手で設計された特徴量では，その特徴量の抽出過程で，目的とする課題のために重要な情報が失われる可能性がある。その一方で，高度に人の手で設計されていない特徴量を入力として用いる深層学習では，DNN へ入力する以前の段階での情報の損失が少ないため，従来の機械学習手法よりも高い性能を実現できる可能性がある。実際，音声認識（speech recognition） [65, 66] や音声合成（speech synthesis） [67–69]，

音声強調（speech enhancement） [70–72]，音源分離（source separation） [73, 74]，話者認識（speaker recognition） [75–77]，話者認証（speaker verification） [78–80] 等の様々な課題に対して，深層学習の高い有効性が確かめられている。

このように，深層学習を用いた聴覚心理的評価の評点の自動推定に関する研究を行う動機づけは十分であり，先行研究もいくつか存在する [49, 62, 63]。しかし，その一方で，深層学習を用いたこれまでのアプローチに関しては、以下の課題が存在する。

まず，音響信号を対象にした深層学習を用いた研究では，対数振幅メル周波数スペクトログラムやメル周波数ケプストラム係数（mel-frequency cepstral coefficients; MFCCs）に代表されるような，スペクトルの振幅情報から計算される音響特徴量がよく使用される [66, 68, 69, 75, 77–79]。深層学習を用いた，聴覚心理的評価の評点の自動推定 [62, 63] や音声障害検出 [81–85] の研究でも，同様に，スペクトルの振幅情報に基づく音響特徴量が頻用されている。その一方で，深層学習以外の機械学習を用いた病的音声の声質自動評価や音声障害検出 [86–88]，深層学習を用いた発語運動障害患者の音声認識 [89] では，スペクトルの振幅情報だけでなく，位相情報も有効であることが示唆されている。しかし，深層学習を用いた聴覚心理的評価の評点の自動推定に関する研究では，スペクトルの位相情報の有効性に関する検討が行われていない。また，聴覚心理的評価の評点の自動推定のために原波形を入力として用いた研究もあるが [49]，やはり位相情報の有効性に関しては検討されていない。

さらに，聴覚心理的評価の評点の自動推定を機械学習によって実現するためには，音声データと，専門家（医師や言語聴覚士）によって与えられた聴覚心理的評価の評点のペアを集めた，病的音声データセットを構築する必要がある。自動推定の性能を上げるためには，データセットのサンプル数を大きくすることが有効である。また，臨床的な有用性という観点から自動推定の性能を評価する場合，正解率等の指標を単に示すだけでは不十分であり，専門家の評者内信頼性や評者間信頼性と比較することが必要であると考えられる。そのため，評者内信頼性や評者間信頼性を正確に調べるためにも，聴覚心理的評価を行う専門家の人数は多いことが望ましい。しかし，病的音声データを大量に収集すること自体が難しい上に，聴覚心理的評価の評点を与える作業には多大な人的資源を要する。Fujimura らの研究 [49] と Kojima らの研究 [62] では，3 名の医師によって GRBAS 尺度の全項目が付与された，1,377 サンプルのデータセットが構築されていたが，これだけのサンプル数に対する聴覚心理的評価はかなりの時間を要する。病的音声データセット構築のコストを削減するためには，サンプル数をさらに少なくすることが望まれる。また，García らの研究 [63] では，データセットのサンプル数は 290 と，1,377 よりかなり少なかったが，聴覚心理的評価を行った評者が専門家 1 名，評価されたのが GRBAS 尺度の項目 G のみであり，研究としては，専門家の人数や評価項目の数について改善の余地がある。

1.4　本論文の目的と構成

本研究では，先行研究の課題を踏まえて，以下のことを目的として，深層学習を用いた聴覚心理的評価の評点の自動推定実験に取り組んだ。

- 聴覚心理的評価の評点の自動推定に対する，スペクトルの位相情報の有効性に関する検討
- 病的音声データセットが小規模な場合における，推定精度向上のための手段に関する検討

本論文は，本章を含め全 8 章から構成される。第 2 章では，時間周波数表現について説明し，特に，短時間フーリエ変換の位相に関する話題を詳述する。第 3 章では，本研究の実験に関連する深層学習の技術について説明する。第 4 章では，GRBAS 尺度の評点が付与された病的音声のデータベース構築について述べる。2,545 サンプルの病的音声データに対しては，1 名の耳鼻咽喉科医によって，その一部である 300 サンプルの音声データに対しては，複数名の評者によって GRBAS 尺度の評点を付与した。第 5 章では，1 名の耳鼻咽喉科医によって評価された 2,545 サンプルの病的音声データを元に，瞬時周波数を応用した位相に基づく時間周波数表現と Recurrent Neural Network を用いて，GRBAS 尺度の評点の自動推定実験を行った内容について述べる。この実験では，位相に基づく時間周波数表現が GRBAS 尺度の評点の自動推定に効果的であることを示唆し，瞬時周波数を含む様々な位相情報の有用性を調査する動機づけを与える。第 6 章では，GRBAS 尺度の評点の自動推定という課題から一度離れ，様々な音分類課題に対して，Convolutional Neural Network を用いて，位相のフェーザ表示，瞬時周波数，群遅延の有用性を調査した結果を報告する。様々な音分類課題で調査を行ったのは，GRBAS 尺度の自動推定という特定の課題以外においても，位相情報は有用であるかどうかを検証することを意図したためである。この調査では，位相情報は他の音分類課題においても有用であり，位相の値それ自体よりも，瞬時周波数や群遅延として表現されるような情報が有用であることを示す。第 7 章では，複数名の専門家によって評価された 300 サンプルの病的音声データを元に，瞬時周波数と群遅延，および Convolutional Neural Network と Recurrent Neural Network を用いて，GRBAS 尺度の評点の自動推定実験を行った内容について述べる。この実験では，データの少なさを補うデータ拡張に関しても調査し，時間周波数表現やデータ拡張に関する設計を適切に行うことによって，専門家に匹敵する推定精度を実現しうることを示す。また，ここでは，深層学習に基づくアプローチと，慣習的な音響特徴量と他の機械学習手法に基づくアプローチの性能比較についても述べる。最後に，第 8 章では，本論文で述べた研究の内容とその成果を総括した後，今後の展望について述べる。

第 2 章

音声信号の時間周波数表現

2.1 はじめに

ヒトの発声において，声帯（vocal folds）は極めて重要な役割を果たす。声帯は，呼気あるいは吸気によって励振され，音源として機能する。一般に，声帯の振動は準周期的であり，豊かな調波構造を持つ。そのため，音声を分析する際に，周波数の概念を利用することは多いに役立つ。しかし，音声信号は，たとえ持続母音であっても本質的に非定常であるため，信号の定常性を仮定するフーリエ変換（Fourier transform）のような古典的な周波数解析（frequency analysis）は解析手法として不適切である。持続母音の合成音声に関する知覚実験では，完全に周期的な合成音声は人工的に知覚される一方で，ピッチ揺らぎを表すジッタ（jitter）や振幅揺らぎを表すシマ（shimmer）の値を大きくすることによって知覚的な自然性が向上することが示されている [90, 91]。また，音声障害を罹患している患者の音声は，音声障害の種類や程度に依存した非定常性が顕れ [92, 93]，時にカオス的になることもある [94, 95]。

古典的な周波数解析は信号の定常性を仮定しているために，非定常な音声信号の解析手法としては不適切である。一方，時間周波数解析（time-frequency analysis）は定常性を仮定しない手法であり，音声信号のような非定常信号を解析するために使われる。時間周波数解析では，1 次元の時間信号に対して何らかの変換を施すことによって，時間軸（time axis）と周波数軸（frequency axis）を持つ平面である 2 次元信号を生成する。この 2 次元信号のことを時間周波数表現（time-frequency representation）あるいは時間周波数平面（time-frequency plane）と呼ぶ。理想的な時間周波数表現は，任意の時刻に発生する任意の周波数に関する直接的な情報を与える [96]。しかし，信号エネルギーを時間周波数表現の局所領域に集中させることはできないという不確定性原理（the uncertainty principle）のため，理想的な時間周波数表現は数理的に実現不可能であることが証明されている [96]。不確定性原理の制約のため，時間周波数表現の時間分解能（time resolution）

を上げると周波数分解能（frequency resolution）が下がり，逆もまた然りである。

時間分解能と周波数分解能がトレードオフの関係にあるため，万能な時間周波数表現は存在しない。また，時間周波数表現を求めるための手法も一つではなく，時間周波数分解能の特性が異なる様々な手法が提案されている。具体的には，短時間フーリエ変換（short-time Fourier transform; STFT）や定 Q 変換（constant-Q transform）[97]，連続ウェーブレット変換（continuous wavelet transform）[98] などの手法が提案されており，考えている問題に応じて使い分けられている。どの手法も一長一短はあるが，音響信号処理に関する深層学習では，対数振幅メル周波数スペクトログラム（log-amplitude mel-frequency spectrogram）と呼ばれる時間周波数表現が広く使用されている。対数振幅メル周波数スペクトログラムの計算では，ヒトの音高の心理量と，物理量である周波数とを対応づけるメル尺度（mel scale）[99, 100] の知見が応用される。メル尺度は，音高の心理量を周波数の対数関数として表現する精神物理学的尺度である。対数振幅メル周波数スペクトログラムは，短時間フーリエ変換の絶対値の対数を取った後，そうして得られた時間周波数表現の周波数軸を線形尺度からメル尺度に変換することで得られる。線形尺度からメル尺度への周波数軸の変換は，精神物理学的知見に基づいた効率的な情報圧縮を意図して提案された [101]。しかし，当初の意図には含まれていなかったが，時間歪み（time-warping deformation）に対する頑健性が望まれる場合，周波数軸がメル尺度である時間周波数表現は，時間歪みに対する変動が小さいという望ましい性質を持つことも指摘されている [102]。実際，対数振幅メル周波数スペクトログラムの有効性は，複数種類の時間周波数表現の比較実験で確かめられている [103–105]。さらに，深層学習によってフィルタバンク自体を最適化する実験では，最適化によってメル尺度に類似するパターンが自然と選択されることが観察されている [49, 106, 107]。なお，精神物理学的尺度としては，メル尺度以外にも，臨界帯域幅（critical bandwidth）と周波数とを対応づけるバーク（Bark）尺度 [108] や ERB（equivalent rectangular bandwidth; 等価矩形帯域幅）尺度 [109] が存在する。メル尺度に基づく時間周波数表現と ERB 尺度に基づく時間周波数表現を入力特徴量として比較した実験では，両者の違いが性能に与える影響は大きくないことが示唆されている [105]。このように，これらの尺度の違いが与える影響は限定的であると考えられるため，本研究では尺度間の比較に関しては検討の対象外とし，メル尺度についてのみ取り扱う。

短時間フーリエ変換や定 Q 変換，複素関数のウェーブレットを用いる連続ウェーブレット変換では，実領域の時間信号は複素領域の時間周波数表現に変換される。複素領域の時間周波数表現の絶対値は振幅（amplitude），偏角は位相（phase）と呼ばれる。振幅は信号の強さに対応するため解釈が容易である他，時間領域での信号の平行移動（translation）に対する不変性 [110] 等の望ましい性質を持つ。その一方で，位相は振幅と比べて解釈が直観的ではない上に，時間領域での信号の平行移動によって値が大きく変化したり [110],

2π で値が循環したりするため，数値計算上の取り扱いにも注意を要する。このような理由から，第 1 章でも言及したが，音響信号処理では，位相よりも振幅が積極的に利用されてきた。しかし，位相情報は音声品質に関わることが，主観評価実験および客観指標で示されている [111–113]。このことから，位相を利用することによって，本研究で取り組む声質自動評価を含め，音響信号処理の性能が向上することが期待される。実際，近年の深層学習を用いた研究で，振幅だけでなく位相も考慮することによって，音声強調 [72, 114, 115] や話者分離 [116] の性能が向上することが確認されている。また，位相の時間方向や周波数方向の偏微分は位相そのものよりも解釈しやすい上に，信号の平行移動に対する頑健性 [110] や 2π で値が循環しないことから数値計算上も比較的扱いやすい。実際，位相の偏微分は，基本周波数推定 [117, 118]，フォルマント推定 [119–122]，自動音声認識 [122–124]，調波打撃音分離（harmonic/percussive source separation）[125, 126]，オンセット検出 [127–130]，音声信号の非周期性指標推定 [131–133]，音声障害の検出や病的音声の声質評価 [86–88] 等，様々な用途で応用されている。

本章では，位相に注目しながら，短時間フーリエ変換とメル尺度による周波数軸のスケール変換に関する説明を行う。なお，本章では，特に断らない限り，ベクトル $\boldsymbol{u} \in \mathbb{R}^N$ の i 番目の要素を $\boldsymbol{u}[i]$，行列 $\boldsymbol{A} \in \mathbb{R}^{M \times N}$ の (j, k) 成分を $\boldsymbol{A}[j, k]$ と表記し，添字は 0 から始める。ベクトル $\boldsymbol{u} \in \mathbb{R}^N$ は，$\boldsymbol{u} : \{0, 1, \ldots, N-1\} \to \mathbb{R}$ のように関数としても解釈され，離散信号や窓関数等を表現するために使用される。また，行列 $\boldsymbol{A} \in \mathbb{R}^{M \times N}$ に関して，$\boldsymbol{A}[m, \cdot] \in \mathbb{R}^N$ は $\boldsymbol{A}$ の m 行目を取り出したベクトル，$\boldsymbol{A}[\cdot, n] \in \mathbb{R}^M$ は $\boldsymbol{A}$ の n 列目を取り出したベクトルを表す。

2.2 短時間フーリエ変換

短時間フーリエ変換には，位相の取り扱いが異なる 2 種類の定義が存在する[*1]。本節では，短時間フーリエ変換の 2 種類の定義を紹介した後，短時間フーリエ変換の位相の偏微分について説明する。なお，短時間フーリエ変換には連続信号に対する定義と離散信号に対する定義が存在するが，特に断らない限り，離散信号に関する話題を取り扱う。

2.2.1 定義①：窓によって切り出された信号のフーリエ変換

離散フーリエ変換（discrete Fourier transform; DFT）は，与えられた信号全体に対する周波数を求めようとする処理であり，音声信号のような非定常信号の解析手法としては不適切である。短時間フーリエ変換は，窓掛けによって信号を短い時間区間に分割し，各時間区間で離散フーリエ変換を行うことによって，非定常信号の時々刻々と変化する性状

[*1] 振幅の正規化定数も考慮すると，さらに定義の種類は増える。

を捉える。

$\boldsymbol{w} \in \mathbb{R}^W (W \in \mathbb{N})$ を窓関数（window function）とする離散信号 $\boldsymbol{x} \in \mathbb{R}^X (X \in \mathbb{N})$ の短時間フーリエ変換 $\hat{\mathcal{F}}^{(\boldsymbol{w},a,D)}\boldsymbol{x} \in \mathbb{C}^{M\times N} (M, N \in \mathbb{N})$ は，以下のように定義される。

$$\left(\hat{\mathcal{F}}^{(\boldsymbol{w},a,D)}\boldsymbol{x}\right)[m,n] := \sum_{l=0}^{W-1} \boldsymbol{x}[an+l]\boldsymbol{w}[l] \exp\left(-\frac{2\pi i m l}{D}\right) \tag{2.1}$$

ここで，$i = \sqrt{-1}$，$m \in \{0, 1, \ldots, M-1\}$ は周波数の添字，$n \in \{0, 1, \ldots, N-1\}$ は時間の添字，$a \in \mathbb{N}$ はシフト長（shift length），$D \in \mathbb{N}$ $(D \geq W)$ は DFT 点数を表わす。

$D > W$ の場合，通常，窓関数 $\boldsymbol{w}$ の次元を D に修正し，$\boldsymbol{w}[W] = \boldsymbol{w}[W+1] = \cdots = \boldsymbol{w}[D-1] = 0$ とするゼロ埋め（zero padding）を行う。しかし，式 (2.1) では，窓関数自体の次元を $\mathbb{R}^D$ にする代わりに，総和の範囲を 0 から $W-1$ までとすることによって，窓関数のゼロ埋めを代替的に表現した。そのため，本論文では，窓関数 $\boldsymbol{w}$ はゼロ埋めが行われていない表現であると考える。$\hat{\mathcal{F}}^{(\boldsymbol{w},a,D)}\boldsymbol{x}$ の周波数軸の次元 M は，$\boldsymbol{x}$ と $\boldsymbol{w}$ が実数値関数であり，$\hat{\mathcal{F}}^{(\boldsymbol{w},a,D)}\boldsymbol{x}$ がナイキスト周波数（Nyquist frequency）に関して対称であるため，冗長な折り返し部分を省いて $M = \lfloor D/2 \rfloor + 1$ とする（$\lfloor\cdot\rfloor$ は床関数）[*2]。

窓掛けによって切り出された時間領域の信号，あるいはその周波数表現をフレーム（frame），$\left(\hat{\mathcal{F}}^{(\boldsymbol{w},a,D)}\boldsymbol{x}\right)[\cdot, n]$ を第 n フレーム（n-th frame）と呼ぶ。このことから，窓長（window length）のことをフレーム長（frame length）と呼ぶこともある。総フレーム数 N は，信号長 X と窓長 W，およびシフト長 a に依存し，$N = 1 + \lfloor (X-W)/a \rfloor$ となる（但し，$X \geq W$ とする）。実装上は，離散信号 $\hat{\boldsymbol{x}} \in \mathbb{R}^{\hat{X}} (\hat{X} \in \mathbb{N})$ の短時間フーリエ変換を考える場合，$\hat{\boldsymbol{x}}[0]$ が第 0 フレームの中心となるように，$\hat{\boldsymbol{x}}$ の両端のそれぞれに対し，窓長の半分の長さでゼロ埋めを行って得られた $\boldsymbol{x} \in \mathbb{R}^{(\hat{X}+2\lfloor W/2 \rfloor)}$ を用いることが多い。

「短時間フーリエ変換」という用語は，演算それ自体 $\hat{\mathcal{F}}^{(\boldsymbol{w},a,D)}$ と演算結果 $\hat{\mathcal{F}}^{(\boldsymbol{w},a,D)}\boldsymbol{x}$ の両方を表すために使われるが，特に $\hat{\mathcal{F}}^{(\boldsymbol{w},a,D)}\boldsymbol{x}$ を指す場合は，複素スペクトログラム（complex spectrogram） [134–136] という用語を用いる。複素スペクトログラムの振幅は振幅スペクトログラム（amplitude spectrogram），パワー（二乗振幅）はパワースペクトログラム（power spectrogram），位相は位相スペクトログラム（phase spectrogram）と呼ばれる[*3]。本論文では，絶対値 $|\cdot|$ や偏角 $\angle(\cdot)$ 等の関数を各要素に作用させることを意味して，振幅スペクトログラム，パワースペクトログラム，位相スペクトログラムをそれぞれ $|\hat{\mathcal{F}}^{(\boldsymbol{w},a,D)}\boldsymbol{x}|$，$|\hat{\mathcal{F}}^{(\boldsymbol{w},a,D)}\boldsymbol{x}|^2$，$\angle(\hat{\mathcal{F}}^{(\boldsymbol{w},a,D)}\boldsymbol{x})$ のように記すことにする。なお，位相

[*2] $\boldsymbol{x}$ と $\boldsymbol{w}$ の少なくともどちらか一方が複素数値関数である場合には，ナイキスト周波数に関する対称性はなくなることに注意。

[*3] スペクトログラム（spectrogram）という用語は，元来「演算結果 $\hat{\mathcal{F}}^{(\boldsymbol{w},a,D)}\boldsymbol{x}$ の二乗振幅（パワースペクトログラム）」を表すために使われていたが [96]，近年では，「短時間フーリエ変換を経由して得られる時間周波数表現」という緩やかな意味で使われる場面が多い。本論文では，後者の緩やかな意味でスペクトログラムという用語を使用する。

は 2π で循環する多義性により一意に値が定まらず，主値による表現と，隣接要素間の値が滑らかに接続するようにアンラップ（unwrapping）された表現がある。そのため，論文によって，位相の取り扱いの違いから，位相スペクトログラムの定義の与え方が異なる場合がある [126, 137]。本論文では，誤解を与えない限りは，主値による表現かアンラップした表現のどちらであるかは明記せず，必要に応じて適宜補って説明する。

2.2.2 変調に基づく解釈

第 2.2.1 節では，「窓関数によって切り出した信号の離散フーリエ変換」として短時間フーリエ変換を定義した。その一方で，このように定義された短時間フーリエ変換は，「変調された窓関数で構成されるフィルタバンクの畳み込み」として解釈することもできる [138, 139]。

変調（modulation）は，次のように定義される演算 $\mathcal{M}_f : \mathbb{C}^K \to \mathbb{C}^K (K \in \mathbb{N})$ である。

$$\left(\mathcal{M}_f \boldsymbol{x}\right)[n] := \boldsymbol{x}[n] \exp(2\pi i f n) \tag{2.2}$$

ここで，$\boldsymbol{x} \in \mathbb{C}^K$ は変調される信号，$f \in \mathbb{R}$ は正規化周波数 [Hz] である。f は変調周波数（modulation frequency）と呼ばれ，$\mathcal{M}_f$ は $\boldsymbol{x}$ のスペクトルを f だけ並進（translation）させる効果を持つ。変調の概念を用いることによって，シフト長が 1 である短時間フーリエ変換は以下のように変形される。

$$\left(\hat{\mathcal{F}}^{(\boldsymbol{w},1,D)}\boldsymbol{x}\right)[m,n] = \sum_{l=0}^{W-1} \boldsymbol{x}[n+l]\boldsymbol{w}[l] \exp\left(-\frac{2\pi i m l}{D}\right) \tag{2.3}$$

$$= \sum_{l=0}^{W-1} \boldsymbol{x}[n+l]\big(\mathcal{M}_{(-m/D)}\boldsymbol{w}\big)[l] \tag{2.4}$$

$$= \sum_{l=n}^{n+W-1} \boldsymbol{x}[l]\big(\mathcal{M}_{(-m/D)}\boldsymbol{w}\big)[l-n] \tag{2.5}$$

ここでさらに，添字を反転させる処理 $\big(\mathrm{flip}(\boldsymbol{x})\big)[n] := \boldsymbol{x}[-n]$ を導入することで，以下の式が導かれる。

$$\left(\hat{\mathcal{F}}^{(\boldsymbol{w},1,D)}\boldsymbol{x}\right)[m,n] = \sum_{l=n}^{n+W-1} \boldsymbol{x}[l]\big(\mathrm{flip}(\mathcal{M}_{(-m/D)}\boldsymbol{w})\big)[n-l] \tag{2.6}$$

$$= \boldsymbol{x} * \big(\mathrm{flip}(\mathcal{M}_{(-m/D)}\boldsymbol{w})\big) \tag{2.7}$$

ここで，$*$ は畳み込み演算を表わす。Hann 窓等のようにローパスフィルタ（正規化中心周波数が 0 Hz のバンドパスフィルタ）として機能する窓関数 $\boldsymbol{w}$ は，$\mathcal{M}_{(-m/D)}\boldsymbol{w}$ のように変調されることで正規化中心周波数が $-m/D$ Hz のバンドパスフィルタになる。

加えて，flip 関数によって周波数の符号が反転するため，$\hat{\mathcal{F}}^{(\boldsymbol{w},1,D)}[m,\cdot]$ は，$\boldsymbol{x}$ に正規化中心周波数が m/D Hz のバンドパスフィルタ $\mathrm{flip}(\mathcal{M}_{(-m/D)}\boldsymbol{w})$ を畳み込んだものとして解釈できる。なお，シフト長が 1 以上の場合（$a \geq 1$）は，ダウンサンプリング $\left(\mathrm{downsample}^a(\boldsymbol{x})\right)[n] := \boldsymbol{x}[an]$ を導入することで，以下のように表現できる。

$$\left(\hat{\mathcal{F}}^{(\boldsymbol{w},a,D)}\boldsymbol{x}\right)[m,n] = \mathrm{downsample}^a\left(\boldsymbol{x} * \left(\mathrm{flip}(\mathcal{M}_{(-m/D)}\boldsymbol{w})\right)\right) \tag{2.8}$$

2.2.3　窓関数の位相特性

短時間フーリエ変換では，通常，Hann 窓等のように左右対称な窓関数が使用される。添字集合が $\{0,1,\ldots,W-1\}$ で与えられる窓関数 $\boldsymbol{w} \in \mathbb{C}^W (W \in \mathbb{N})$ が対称であるとは，具体的には，窓長 W が偶数の場合，$\boldsymbol{w}[0] = \overline{\boldsymbol{w}[W-1]}, \boldsymbol{w}[1] = \overline{\boldsymbol{w}[W-2]}, \ldots, \boldsymbol{w}[\lfloor W/2 \rfloor - 1] = \overline{\boldsymbol{w}[\lfloor W/2 \rfloor]}$，窓長 W が奇数の場合，$\boldsymbol{w}[0] = \overline{\boldsymbol{w}[W-1]}, \boldsymbol{w}[1] = \overline{\boldsymbol{w}[W-2]}, \ldots, \boldsymbol{w}[\lfloor W/2 \rfloor - 1] = \overline{\boldsymbol{w}[\lfloor W/2 \rfloor + 1]}$ であることを指す。このような対称性を持つ窓は，周波数に対して位相が線形に変化する直線位相特性（linear phase characteristic）と呼ばれる周波数特性を持つ。窓関数の時刻を表わす添字集合が $\{0,1,\ldots,W-1\}$ で与えられた場合，直線位相成分（周波数に対する位相の線形変化）を 0 にすることはできない。しかし，窓長 W が奇数である場合に限り，$\{-\lfloor W/2 \rfloor, -\lfloor W/2 \rfloor + 1, \ldots, -1, 0, 1, \ldots, \lfloor (W-1)/2 \rfloor\}$ のように窓関数の中心の添字（時刻）を 0 とすることによって，直線位相成分を 0 にすることができる [140][*4]。このような位相特性を零位相特性（zero phase characteristic）と呼ぶ[*5]。

式 (2.1) で定義される短時間フーリエ変換は，式 (2.8) のように，変調された窓関数を畳み込んだものとして解釈することができる。時間領域での畳み込みは，周波数領域での乗算に対応し，特に位相に関しては加算に対応する[*6]。よって，窓関数が零位相特性を持つ場合，窓関数は位相に影響を及ぼさない [140]。そこで，窓関数が零位相特性を持つように，窓関数の添字（時刻）集合を $\{-\lfloor W/2 \rfloor, -\lfloor W/2 \rfloor + 1, \ldots, -1, 0, 1, \ldots, \lfloor (W-1)/2 \rfloor\}$ とし，短時間フーリエ変換を以下のように再定義する。

$$\left(\mathcal{F}_1^{(\boldsymbol{w},a,D)}\boldsymbol{x}\right)[m,n] := \sum_{l=-\lfloor W/2 \rfloor}^{\lfloor (W-1)/2 \rfloor} \boldsymbol{x}[an+l]\boldsymbol{w}[l] \exp\left(-\frac{2\pi i m l}{D}\right) \tag{2.9}$$

[*4] 窓長 W が偶数の場合でも，窓関数の片方の端点のみが 0 になるように設定して実質的な窓長を奇数にすることで，零位相特性を実現することができる [140]。本論文の実験で使用された窓長が偶数の窓関数は，左端の値のみ 0 に設定された。左端の値のみを 0 にする場合，添字集合を $\{-\lfloor W/2 \rfloor, -\lfloor W/2 \rfloor + 1, \ldots, -1, 0, 1, \ldots, \lfloor (W-1)/2 \rfloor\}$ とすることによって，窓長が奇数の場合も偶数の場合も零位相特性になる。

[*5] 窓関数が時刻 0 に関して対称である場合，スペクトルは実数になるが，負になることもある。スペクトルが負の領域では，位相は 0 でなく，π（$-\pi$ 等とも表現しうる）になる。

[*6] 偏角（位相）がそれぞれ a, b である 2 つの複素数 $Ae^{ia}, Be^{ib} (A, B, a, b \in \mathbb{R})$ に関して，それらの積は $ABe^{i(a+b)}$ であり，積の偏角（位相）は $a+b$ である。

計算機で離散フーリエ変換を実行する際は，通常，時間添字は 0 から始まるため工夫が必要である。実装上は，窓掛けによって得られた $\Big[\boldsymbol{x}[an]\boldsymbol{w}[0], \ldots, \boldsymbol{x}[an+W-1]\boldsymbol{w}[W-1]\Big]$ を，$\lfloor W/2 \rfloor$ だけ巡回シフトさせて $\Big[\boldsymbol{x}[an+\lfloor W/2 \rfloor]\boldsymbol{w}[\lfloor W/2 \rfloor], \ldots, \boldsymbol{x}[an+\lfloor W/2 \rfloor - 1]\boldsymbol{w}[\lfloor W/2 \rfloor - 1]\Big]$ とし，窓関数の中心の添字が 0 に，すなわち離散フーリエ変換の時間原点に来るように調整することで対処する [140]。なお，窓関数の添字集合を変更しても，式 (2.8) のような変調フィルタバンクの解釈はそのまま成り立つ。

$$\left(\mathcal{F}_1^{(\boldsymbol{w},a,D)}\boldsymbol{x}\right)[m,n] = \mathrm{downsample}^a\left(\boldsymbol{x} * \left(\mathrm{flip}(\mathcal{M}_{(-m/D)}\boldsymbol{w})\right)\right) \tag{2.10}$$

図 2.1 (A) は，正弦波，周波数変調波，パルス，微弱なホワイトノイズ等から構成される混合信号に対する振幅スペクトログラムのヒートマップを表わす。図 2.1 (B), (C) は，同じ信号に対する位相スペクトログラム $\angle(\hat{\mathcal{F}}\boldsymbol{x}), \angle(\mathcal{F}_1\boldsymbol{x})$（パラメータの記載は省略）のヒートマップを表わす。一見するだけでも，$\angle(\hat{\mathcal{F}}\boldsymbol{x})$ と $\angle(\mathcal{F}_1\boldsymbol{x})$ のヒートマップが異なる見た目をしていることがわかる。例えば，5 Hz の正弦波が存在する領域では，$\angle(\hat{\mathcal{F}}\boldsymbol{x})$ に関しては斜め方向のストライプが見られるのに対し，$\angle(\mathcal{F}_1\boldsymbol{x})$ に関しては周波数方向に揃ったストライプが見られる。また，3 s に位置するパルスが存在する領域では，$\angle(\hat{\mathcal{F}}\boldsymbol{x})$ に関してはパターンが見えづらいが，$\angle(\mathcal{F}_1\boldsymbol{x})$ に関しては 3 s で周波数方向に位相が揃っている様子が見て取れる。このように，窓関数の直線位相成分を除去することによって，正弦波やパルスに対して表われる位相のパターンを顕在化させることができる。

2.2.4 定義②：窓と変調された信号の畳み込み

ここまで用いてきた短時間フーリエ変換の定義では，窓関数と複素正弦波が時間添字を共有していた。これとは別に，短時間フーリエ変換には，次のように，解析対象の離散信号と複素正弦波が時間添字を共有する定義が存在する。

$$\left(\mathcal{F}_2^{(\boldsymbol{w},a,D)}\boldsymbol{x}\right)[m,n] := \sum_{l=an-\lfloor W/2 \rfloor}^{an+\lfloor (W-1)/2 \rfloor} \boldsymbol{x}[l]\boldsymbol{w}[l-an]\exp\left(-\frac{2\pi i m l}{D}\right) \tag{2.11}$$

この定義は，窓関数 $\boldsymbol{w}$ と変調された離散信号 $\boldsymbol{x}$ の畳み込みを通じて解釈することができる。

$$\left(\mathcal{F}_2^{(\boldsymbol{w},a,D)}\boldsymbol{x}\right)[m,n] = \mathrm{downsample}^a\left(\mathcal{M}_{(-m/D)}\boldsymbol{x} * \left(\mathrm{flip}(\boldsymbol{w})\right)\right) \tag{2.12}$$

変調されるのが窓関数であるか，解析対象の離散信号であるかの違いは，結果として表れる位相に影響を及ぼす。$\left(\mathcal{F}_1^{(\boldsymbol{w},a,D)}\boldsymbol{x}\right)[m,\cdot]$ は，正規化中心周波数が m/D Hz であるバンドパスフィルタ $\mathrm{flip}(\mathcal{M}_{(-m/D)}\boldsymbol{w})$ の出力であるため，位相の時間変化の速さは概ね m/D Hz 付近になる。その一方で，$\left(\mathcal{F}_2^{(\boldsymbol{w},a,D)}\boldsymbol{x}\right)[m,\cdot]$ は，$-m/D$ Hz だけ変調された

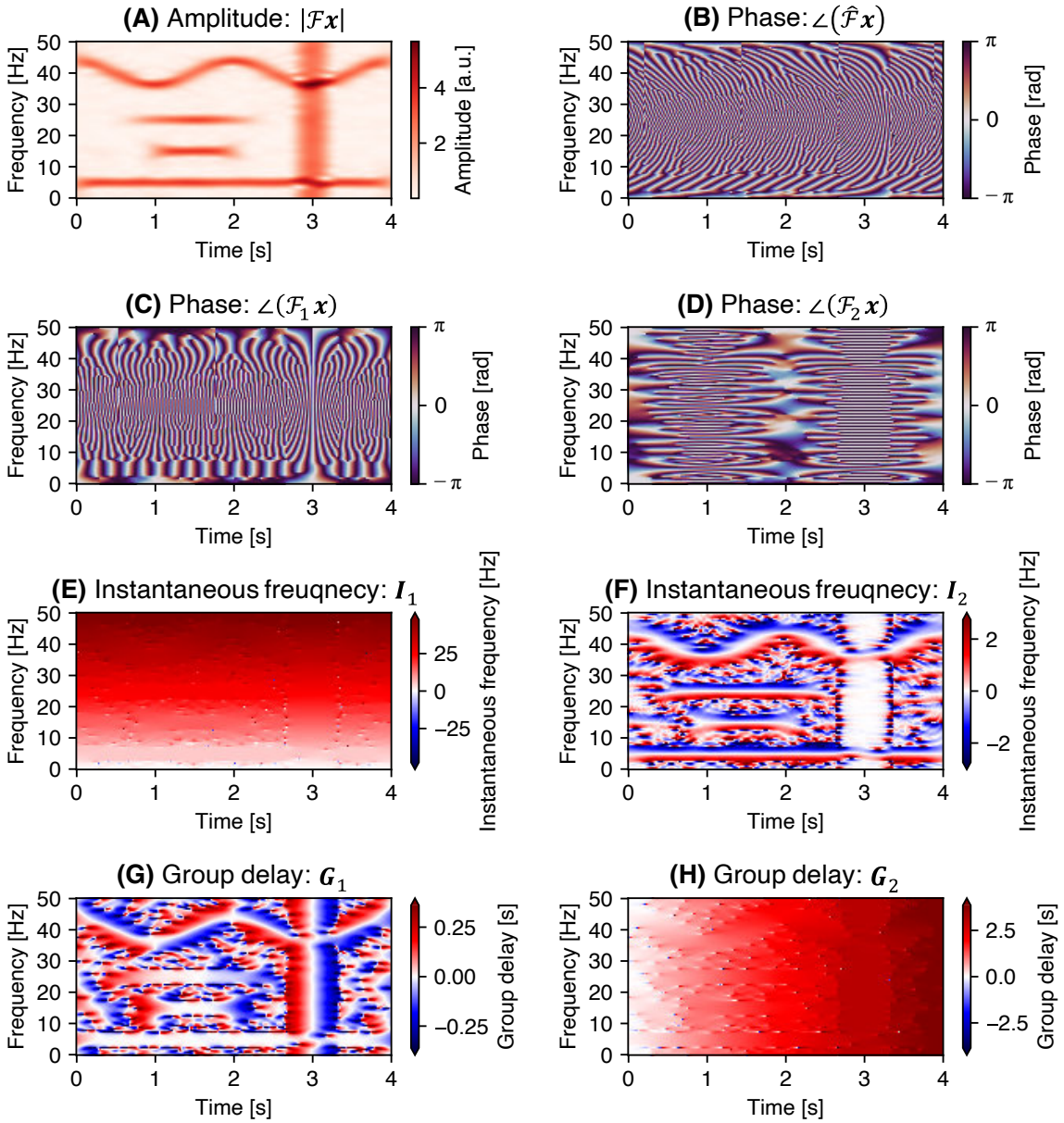


図 2.1 正弦波，周波数変調波，パルス，微弱なホワイトノイズ等から構成される混合信号に対する様々な時間周波数表現のヒートマップ。標本化周波数は 100 Hz，窓長は 750 ms，DFT 点数は 400 (4 s)，シフト長は 1 sample (10 ms) とした。(A) は振幅 $|\mathcal{F}\boldsymbol{x}|$（どの短時間フーリエ変化の定義を採用しても同じ値であるため，定義に関する記号は省略して単に $|\mathcal{F}\boldsymbol{x}|$ と表記）を，(B), (C), (D) はそれぞれ位相 $\angle(\hat{\mathcal{F}}\boldsymbol{x}), \angle(\mathcal{F}_1\boldsymbol{x}), \angle(\mathcal{F}_2\boldsymbol{x})$ を，(E) と (F) はそれぞれ瞬時周波数 $\boldsymbol{I}_1, \boldsymbol{I}_2$ を，(G) と (H) はそれぞれ群遅延 $\boldsymbol{G}_1, \boldsymbol{G}_2$ を表わす。(E)，(F)，(G)，(H) の各図に関して，絶対値の 95 パーセンタイル値を $\theta_{95\%}$ と記すとき，色の表示範囲を $[-\theta_{95\%}, \theta_{95\%}]$ に制限した。瞬時周波数や群遅延の値の絶対値は，パワーが非常に小さい要素の近辺で極端に大きくなるため，表示範囲の制限がない場合，ヒートマップは非常に見にくくなる。

離散信号 $\mathcal{M}_{(-m/D)}\boldsymbol{x}$ をローパスフィルタ（正規化中心周波数が 0 Hz であるバンドパスフィルタ）$\mathrm{flip}(\boldsymbol{w})$ に通した出力であるため，位相の時間変化の速さは 0 Hz 付近になる。つまり，$\left(\mathcal{F}_1^{(\boldsymbol{w},a,D)}\boldsymbol{x}\right)[m,\cdot]$ の位相の時間変化の速さは m に対して比例的に増加するのに対し，$\left(\mathcal{F}_2^{(\boldsymbol{w},a,D)}\boldsymbol{x}\right)[m,\cdot]$ の位相の時間変化の速さは m に比例しない。振幅スペクトログラムのヒートマップが図 2.1 (A) のようになる信号に関して，図 2.1 (C), (D) は，位相スペクトログラム $\angle(\mathcal{F}_1\boldsymbol{x}), \angle(\mathcal{F}_2\boldsymbol{x})$ のヒートマップを表わす。$\angle(\mathcal{F}_1\boldsymbol{x})$ と比較して，$\angle(\mathcal{F}_2\boldsymbol{x})$ では位相の時間変化が小さくなっている様子が見て取れる。また，正弦波やパルスが存在する領域において，$\angle(\mathcal{F}_1\boldsymbol{x})$ に関しては位相が周波数方向に揃うパターンが表れるが，$\angle(\mathcal{F}_2\boldsymbol{x})$ に関してはそのようなパターンは表れない。

2.2.5 連続信号に対する短時間フーリエ変換の位相の偏微分

第 2.2.4 節では，短時間フーリエ変換の定義の違いが位相に及ぼす影響を，位相の時間方向や周波数方向の変化に着目して説明した。定義の違いが位相に及ぼす影響は，位相の時間微分である瞬時周波数や周波数微分である群遅延を使うことで，より直接的に表現される。また，瞬時周波数や群遅延を用いることによって，2π で値が循環する問題が解消され，位相に関する考察が行いやすくなる。但し，これまで離散信号に対する短時間フーリエ変換の説明を行ってきたが，微分という概念自体は連続関数に対して定義されるものであるため，本節では，連続信号に対する短時間フーリエ変換を使って位相の偏微分の説明を行う。

離散信号に対する短時間フーリエ変換 $\mathcal{F}^1, \mathcal{F}^2$ に対応する連続信号に対する短時間フーリエ変換をそれぞれ $\mathscr{F}^1, \mathscr{F}^2$ と表わすことにする。$w : \mathbb{R} \to \mathbb{R}$ を窓関数とする連続信号 $x : \mathbb{R} \to \mathbb{R}$ の短時間フーリエ変換 $\mathscr{F}_w^1, \mathscr{F}_w^2$ は，それぞれ次のように定義される。

$$\left(\mathscr{F}_1^w x\right)(f,t) := \int_{\mathbb{R}} x(\tau)w(\tau - t)e^{-2\pi i f(\tau - t)}\,\mathrm{d}\tau \tag{2.13}$$

$$\left(\mathscr{F}_2^w x\right)(f,t) := \int_{\mathbb{R}} x(\tau)w(\tau - t)e^{-2\pi i f\tau}\,\mathrm{d}\tau \tag{2.14}$$

これらの二つの定義の間にある違いは指数関数の冪指数（位相）のみである。ここで，短時間フーリエ変換を $\left(\mathscr{F}_1^w x\right)(f,t) = A_1^w(f,t)e^{i\phi_1^w(f,t)}, \left(\mathscr{F}_2^w x\right)(f,t) = A_2^w(f,t)e^{i\phi_2^w(f,t)}$ のように極形式で表すと，二つの定義は次の等式で結ばれる。

$$A_1^w(f,t)e^{i\phi_1^w(f,t)} = e^{2\pi i f t}\left(A_2^w(f,t)e^{i\phi_2^w(f,t)}\right) \tag{2.15}$$

$$= A_2^w(f,t)e^{i(\phi_2^w(f,t) + 2\pi f t)} \tag{2.16}$$

この等式より，振幅と位相，位相の偏微分に関して以下の関係が成り立つ。

$$A_1^w(f,t) = A_2^w(f,t) \tag{2.17}$$
$$\phi_1^w(f,t) = \phi_2^w(f,t) + 2\pi f t \tag{2.18}$$
$$\frac{\partial \phi_1^w}{\partial t}(f,t) = \frac{\partial \phi_2^w}{\partial t}(f,t) + 2\pi f \tag{2.19}$$
$$\frac{\partial \phi_1^w}{\partial f}(f,t) = \frac{\partial \phi_2^w}{\partial f}(f,t) + 2\pi t \tag{2.20}$$

つまり，ある信号 x に関して，片方の定義に対する値がわかると，もう一方の定義に対する値も直ちに求まる。

さて，位相の偏微分は，勿論位相を直接使って求めることもできるが，位相を直接使わずに求めることもできる [141]。まず，短時間フーリエ変換の極形式の偏微分を考える。

$$\frac{\partial A_1^w e^{i\phi_1^w}}{\partial t} = \frac{\partial A_1^w}{\partial t} e^{i\phi_1^w} + i\frac{\partial \phi_1^w}{\partial t} A_1^w e^{i\phi_1^w}, \quad \frac{\partial A_2^w e^{i\phi_2^w}}{\partial t} = \frac{\partial A_2^w}{\partial t} e^{i\phi_2^w} + i\frac{\partial \phi_2^w}{\partial t} A_2^w e^{i\phi_1^w} \tag{2.21}$$

$$\frac{\partial A_1^w e^{i\phi_1^w}}{\partial f} = \frac{\partial A_1^w}{\partial f} e^{i\phi_1^w} + i\frac{\partial \phi_1^w}{\partial f} A_1^w e^{i\phi_1^w}, \quad \frac{\partial A_2^w e^{i\phi_2^w}}{\partial f} = \frac{\partial A_2^w}{\partial f} e^{i\phi_2^w} + i\frac{\partial \phi_2^w}{\partial f} A_2^w e^{i\phi_1^w} \tag{2.22}$$

これらの両辺を $A_1^w e^{i\phi_1^w}$ または $A_2^w e^{i\phi_2^w}$ で割って，虚部を取ることによって，次の関係が導かれる。

$$\frac{\partial \phi_1^w}{\partial t} = \mathrm{Im}\left(\frac{1}{A_1^w e^{i\phi_1^w}} \frac{\partial A_1^w e^{i\phi_1^w}}{\partial t}\right), \qquad \frac{\partial \phi_2^w}{\partial t} = \mathrm{Im}\left(\frac{1}{A_2^w e^{i\phi_2^w}} \frac{\partial A_2^w e^{i\phi_2^w}}{\partial t}\right) \tag{2.23}$$

$$\frac{\partial \phi_1^w}{\partial f} = \mathrm{Im}\left(\frac{1}{A_1^w e^{i\phi_1^w}} \frac{\partial A_1^w e^{i\phi_1^w}}{\partial f}\right), \qquad \frac{\partial \phi_2^w}{\partial f} = \mathrm{Im}\left(\frac{1}{A_2^w e^{i\phi_2^w}} \frac{\partial A_2^w e^{i\phi_2^w}}{\partial f}\right) \tag{2.24}$$

$\partial(\mathscr{F}_1^w x)/\partial t = \partial(A_1^w e^{i\phi_1^w})/\partial t$, $\partial(\mathscr{F}_2^w x)/\partial t = \partial(A_2^w e^{i\phi_2^w})/\partial t$, $\partial(\mathscr{F}_1^w x)/\partial f = \partial(A_1^w e^{i\phi_1^w})/\partial f$, $\partial(\mathscr{F}_2^w x)/\partial f = \partial(A_2^w e^{i\phi_2^w})/\partial f$ を計算すると，次のようになる。

$$\frac{\partial \mathscr{F}_1^w x}{\partial t}(f,t) = -\int_{\mathbb{R}} x(\tau)\frac{\partial w}{\partial t}(\tau - t)e^{-2\pi i f(\tau - t)}\,\mathrm{d}\tau + 2\pi i f\Big(\mathscr{F}_1^w x\Big)(f,t) \tag{2.25}$$

$$\frac{\partial \mathscr{F}_2^w x}{\partial t}(f,t) = -\int_{\mathbb{R}} x(\tau)\frac{\partial w}{\partial t}(\tau - t)e^{-2\pi i f\tau}\,\mathrm{d}\tau \tag{2.26}$$

$$\frac{\partial \mathscr{F}_1^w x}{\partial f}(f,t) = -2\pi i\int_{\mathbb{R}} x(\tau)(\tau - t)w(\tau - t)e^{-2\pi i f(\tau - t)}\,\mathrm{d}\tau \tag{2.27}$$

$$\frac{\partial \mathscr{F}_2^w x}{\partial f}(f,t) = -2\pi i\int_{\mathbb{R}} x(\tau)(\tau - t)w(\tau - t)e^{-2\pi i f\tau}\,\mathrm{d}\tau - 2\pi i t\Big(\mathscr{F}_2^w x\Big)(f,t) \tag{2.28}$$

ここで，$\big(\mathscr{D}w\big)(t) := \big(\partial w/\partial t\big)(t)$, $\big(\mathscr{T}w\big)(t) := tw(t)$ という演算子を導入することによっ

て，上記の式は $\mathscr{D}w$ や $\mathscr{T}w$ を用いた表現で書き直すことができる。

$$\frac{\partial \mathscr{F}_1^w x}{\partial t}(f,t) = -\Big(\mathscr{F}_1^{\mathscr{D}w} x\Big)(f,t) + 2\pi i f \Big(\mathscr{F}_1^w x\Big)(f,t) \tag{2.29}$$

$$\frac{\partial \mathscr{F}_2^w x}{\partial t}(f,t) = -\Big(\mathscr{F}_2^{\mathscr{D}w} x\Big)(f,t) \tag{2.30}$$

$$\frac{\partial \mathscr{F}_1^w x}{\partial f}(f,t) = -2\pi i \Big(\mathscr{F}_1^{\mathscr{T}w} x\Big)(f,t) \tag{2.31}$$

$$\frac{\partial \mathscr{F}_2^w x}{\partial f}(f,t) = -2\pi i \Big(\mathscr{F}_2^{\mathscr{T}w} x\Big)(f,t) - 2\pi i t \Big(\mathscr{F}_2^w x\Big)(f,t) \tag{2.32}$$

式 (2.29) から式 (2.32) までを式 (2.23) と式 (2.24) に代入することによって，以下の関係式が導かれる。

$$\frac{\partial \phi_2^w x}{\partial t} = -\operatorname{Im}\left(\frac{(\mathscr{F}_1^{\mathscr{D}w} x)\overline{(\mathscr{F}_1^w x)}}{\|\mathscr{F}_1^w x\|^2}\right) = -\operatorname{Im}\left(\frac{(\mathscr{F}_2^{\mathscr{D}w} x)\overline{(\mathscr{F}_2^w x)}}{\|\mathscr{F}_2^w x\|^2}\right) \tag{2.33}$$

$$\frac{\partial \phi_1^w x}{\partial f} = -2\pi \operatorname{Re}\left(\frac{(\mathscr{F}_1^{\mathscr{T}w} x)\overline{(\mathscr{F}_1^w x)}}{\|\mathscr{F}_1^w x\|^2}\right) = -2\pi \operatorname{Re}\left(\frac{(\mathscr{F}_2^{\mathscr{T}w} x)\overline{(\mathscr{F}_2^w x)}}{\|\mathscr{F}_2^w x\|^2}\right) \tag{2.34}$$

実は，どちらの短時間フーリエ変換の定義を採用しても，このように全く同じ形式で $\partial\phi_2^w x/\partial t$ と $\partial\phi_1^w x/\partial f$ が得られる。$\partial\phi_1^w x/\partial t$ と $\partial\phi_2^w x/\partial f$ に関しては，式 (2.33) と式 (2.34) を，位相の偏微分間の関係を表わす式 (2.19) と式 (2.20) を組み合わせることによって求まる。

$\partial\phi_1^w x/\partial t$ と $\partial\phi_2^w x/\partial t$ はどちらも瞬時周波数と呼ばれ，解析信号（analytic signal）の瞬時周波数と区別したい場合は，特にチャンネル化された瞬時周波数（channelized instantaneous frequency）と呼ばれることもある [120, 142]。$\partial\phi_1^w x/\partial t$ と $\partial\phi_2^w x/\partial t$ は式 (2.19) の関係式で結ばれ，互いに異なる値を取る。離散信号に関して式 (2.10) と式 (2.12) で示したように，時間の単位が s である場合，$\big(\mathscr{F}_1 x\big)(f,\cdot)$ は中心周波数が f Hz であるバンドパスフィルタを経た出力，$\big(\mathscr{F}_2 x\big)(f,\cdot)$ は $-f$ Hz 変調された入力信号がローパスフィルタ（中心周波数が 0 Hz のバンドパスフィルタ）を経た出力である。そのため，$\big(\partial\phi_1^w x/\partial t\big)(f,\cdot)$ が表わすのは入力信号の実際の瞬時周波数の大きさであり，$\big(\partial\phi_2^w x/\partial t\big)(f,\cdot)$ が表わすのは f Hz を中心に相対化された瞬時周波数の大きさである。例えば，入力信号 $x(t)$ が f_0 Hz の正弦波であるとき，$\big(\partial\phi_1^w x/\partial t\big)(f,\cdot) = f_0$，$\big(\partial\phi_2^w x/\partial t\big)(f,\cdot) = f_0 - f$ となる。このことから，$\partial\phi_1^w x/\partial t$ は絶対瞬時周波数（absolute instantaneous frequency），$\partial\phi_2^w x/\partial t$ は相対瞬時周波数（relative instantaneous frequency）と呼ばれる [135, 136]。

$(-1/2\pi)(\,\partial\phi_1^w x/\partial f\,)$ は群遅延と呼ばれ，大局的な情報を表わす古典的な意味での群遅延と区別したい場合は，特に局所群遅延（local group delay）と呼ばれることもある [120, 142]。式 (2.19) より $(-1/2\pi)(\,\partial\phi_2^w x/\partial f\,) = (-1/2\pi)(\,\partial\phi_1^w x/\partial f\,) + t$ となるこ

とからわかるが，$(-1/2\pi)(\partial\phi_2^w x/\partial f)$ の値は，バイアス t の存在によって時刻の絶対位置に依存してしまう。そのため，通常，群遅延といえば $(-1/2\pi)(\partial\phi_1^w x/\partial f)$ のことを指す。

2.2.6　離散信号に対する短時間フーリエ変換の位相の偏微分

第 2.2.5 節では，連続信号に対する短時間フーリエ変換の位相の偏微分について述べた。本節では，離散信号に対する短時間フーリエ変換の位相の偏微分について説明する。

離散データで微分を計算する際の素朴なアプローチは，微分を差分で近似することである。実際，離散信号の短時間フーリエ変換の位相の偏微分は，アンラップ処理と差分を組み合わせることによって，近似的に求められる場合もある。しかし，位相が 2π で循環する性質のため，隣接要素間の位相変化量の絶対値が実際には π より大きな場合には，アンラップ処理を適切に行えず，差分は微分の近似ではなくなってしまうという問題が生じる。例えば，時間差分に関しては，シフト長 a が 2 以上の場合，ある周波数ビンにおいて，隣接フレーム間での位相変化量の絶対値は π より大きい可能性がある。また，周波数差分に関しては，DFT 点数 D が不十分な場合，隣接周波数ビン間での位相変化量の絶対値が π より大きくなる可能性がある。また，第 2.2.5 節の最後の段落でも述べたように，$(\partial\mathscr{F}_2^w x/\partial f)$ は時間に比例するバイアス項の存在によって値の増減が激しいため，アンラップ処理に関する問題が（短時間フーリエ変換の時間長が極端に短くない限り）ほぼ確実に生じる。

差分計算で微分を近似できない問題を回避するには，式 (2.33) と式 (2.34) の式を用いて位相の偏微分を計算したら良い。式 (2.33) と式 (2.34) の式は連続信号に対するものであるが，式中の $\mathscr{F}_1, \mathscr{F}_2, x, w$ を $\mathcal{F}_1, \mathcal{F}_2, \boldsymbol{x}, \boldsymbol{w}$ に置換することによって，そのまま離散信号に適用することができる。ここで，窓関数 $\boldsymbol{w} \in \mathbb{R}^W$ を用いた離散信号 $\boldsymbol{x} \in \mathbb{R}^X$ の短時間フーリエ変換 $\mathcal{F}^{\boldsymbol{w}}\boldsymbol{x} \in \mathbb{R}^{M\times N}(W, X, M, N \in \mathbb{N})$ に関して，$\partial\mathscr{F}_1^w x/\partial t$ と $\partial\mathscr{F}_2^w x/\partial t$ に対応する瞬時周波数をそれぞれ $\boldsymbol{I}_1, \boldsymbol{I}_2 \in \mathbb{R}^{M\times N}$，$(-1/2\pi)(\partial\mathscr{F}_1^w x/\partial f)$ と $(-1/2\pi)(\partial\mathscr{F}_2^w x/\partial f)$ に対応する群遅延をそれぞれ $\boldsymbol{G}_1, \boldsymbol{G}_2 \in \mathbb{R}^{M\times N}$ と表わすことにする。但し，短時間フーリエ変換の記号に関して，定義の種類とシフト長，DFT 点数に関する表記は省略した。このとき，$\boldsymbol{I}_2$ と $\boldsymbol{G}_1$ は次のように計算できる。

$$\boldsymbol{I}_2 = -\operatorname{Im}\left(\frac{(\mathcal{F}^{\mathcal{D}\boldsymbol{w}}\boldsymbol{x})\overline{(\mathcal{F}^{\boldsymbol{w}}\boldsymbol{x})}}{|\mathcal{F}^{\boldsymbol{w}}\boldsymbol{x}|^2}\right) \tag{2.35}$$

$$\boldsymbol{G}_1 = \operatorname{Re}\left(\frac{(\mathcal{F}^{\mathcal{T}\boldsymbol{w}}\boldsymbol{x})\overline{(\mathcal{F}^{\boldsymbol{w}}\boldsymbol{x})}}{|\mathcal{F}^{\boldsymbol{w}}\boldsymbol{x}|^2}\right) \tag{2.36}$$

ここで，$\mathcal{D}\boldsymbol{w}$ は $\boldsymbol{w}$ を微分したもの，$\big(\mathcal{T}\boldsymbol{w}\big)[n] := n\boldsymbol{w}[n]$ である。離散関数である $\boldsymbol{w}$ の微

分の計算の仕方としては，離散フーリエ変換を経由する方法の他に，Hann 窓等の元々連続関数で定義された窓関数ならば，連続関数として定義された窓関数の導関数をサンプリングする方法等があり [142]，本論文の実験では連続関数の導関数をサンプリングする方法を採用する。また，$\big(\mathcal{T}\boldsymbol{w}\big)[n] := n\boldsymbol{w}[n]$ という定義で計算される群遅延の値の単位は sample になる。群遅延の値の単位を s（秒）に変更したい場合は，サンプリング周波数が f_s Hz ならば，$\big(\mathcal{T}\boldsymbol{w}\big)[n] := (n/f_s)\boldsymbol{w}[n]$ とすれば良い。なお，サンプリング周波数が f_s Hz，シフト長が a，瞬時周波数の値の単位が Hz，群遅延の値の単位が s である場合，$\boldsymbol{I}_1[m,n]$ と $\boldsymbol{G}_2[m,n]$ は次のように求められる。

$$\boldsymbol{I}_1[m,n] = \boldsymbol{I}_2[m,n] + \frac{mf_s}{D} \tag{2.37}$$

$$\boldsymbol{G}_2[m,n] = \boldsymbol{G}_1[m,n] + \frac{an}{f_s} \tag{2.38}$$

さて，瞬時周波数と群遅延はそれぞれ式 (2.35) と式 (2.36) によって計算できるが，式中の分母がパワースペクトログラムであるため，パワースペクトログラムが 0 となる要素では値が発散する [143]。また，パワースペクトログラムの値が 0 でなくとも，その値がかなり小さい要素では，瞬時周波数や群遅延の絶対値は極端に大きな値を取ってしまう。この問題を避けるための手段としては，分母に微小値を足す方法 [136] や，パワースペクトログラムを重み付けとして使うことによって，パワースペクトログラムの値が小さな要素の瞬時周波数や群遅延の影響力を弱める手法 [132, 133] が提案されている。

振幅スペクトログラムのヒートマップが図 2.1 (A) のようになる信号に関して，図 2.1 (E), (F) は瞬時周波数 $\boldsymbol{I}_1, \boldsymbol{I}_2$，図 2.1 (G), (H) は群遅延 $\boldsymbol{G}_1, \boldsymbol{G}_2$ のヒートマップを表わす。これらの図では，パワーが非常に小さい要素の近辺で瞬時周波数や群遅延の値の絶対値が大きくなる問題に関して，前段落で言及した数値的手法は用いず，ヒートマップの色の表示範囲を制限することで対処した。絶対瞬時周波数 $\boldsymbol{I}_1$ や（通常は使用されない）群遅延 $\boldsymbol{G}_2$ では，それぞれ周波数軸と時間軸に比例するバイアスが存在している。その一方で，相対瞬時周波数 $\boldsymbol{I}_2$ や群遅延 $\boldsymbol{G}_1$ ではそのようなバイアスが存在せず，正弦波や周波数変調波，パルスに対して特徴的なパターンが表れている様子が見て取れる。$\boldsymbol{I}_2$ に関して，正弦波や周波数変調波が支配的な領域では縞模様のパターンが表れ，パルスが支配的な領域では 0 に近い値となっていることが観察できる。$\boldsymbol{G}_1$ に関して，正弦波が支配的な領域では 0 に近い値を示し，周波数ビンごとに考えた場合のオンセットとオフセットに対して，それぞれ大きな値，小さな値を示している様子が見られる。これらの性質から，基本周波数推定 [117, 118]，群遅延はオンセット検出 [127–130] 等に応用されている。

ここで，持続母音/a/の音声波形に対する瞬時周波数や群遅延の観察も行う。図 2.2 は持続母音/a/の音声波形に対して狭帯域分析を，図 2.3 は広帯域分析を行ったときのヒートマップを表わす。図 2.1 (F), (G) で正弦波成分に表れていたような特徴的なパターン

が，狭帯域分析実行時は調波成分に対して，広帯域分析実行時はフォルマントに対して表れていることが観察できる。このため，瞬時周波数に関して，狭帯域分析を用いて基本周波数推定に応用した研究 [117,118] もあれば，広帯域分析を用いてフォルマント推定に応用した研究 [119] もある。

2.3　メル尺度

2.3.1　メル尺度の実装

メル尺度 [99,100] は，精神物理学的な聴覚心理実験の結果に基づいて，ヒトの音高の心理量と物理量である周波数とを対応づける尺度である。メル（mel）は，音高の心理量に比例する単位であり，1000 Hz の純音から知覚される音高の心理量は 1000 mel と定義される。メル尺度は，大雑把には，ある周波数（多くの場合 1000 Hz）までは線形で，それ以上の周波数に対しては対数的であると説明されることがある [144]。

数式によるメル尺度の表現はいくつか存在するが [144]，近年広く利用されている Python の音響信号処理ライブラリである librosa [145] では[*7]，Slaney [146] による定義と HTK [147] による定義の 2 つが実装されている。Slaney による定義では，1000 Hz 未満が線形関数，1000 Hz 以上が対数関数で表現される[*8]。

$$m(f) = \begin{cases} f & (f < 1000) \\ 1000 + \dfrac{1800}{\ln 6.4} \ln\left(\dfrac{f}{1000}\right) & (f \geq 1000) \end{cases} \tag{2.39}$$

ここで，$f \geq 0$ は周波数 [Hz] であり，$m \geq 0$ は音高の心理量 [mel] を表わす。その一方で，音声信号処理に関する研究で使用されることの多いソフトウェアである HTK の定義では，場合分けを行わない以下の表現を用いる。

$$m(f) = 2595 \log_{10}\left(1 + \frac{f}{700}\right) \tag{2.40}$$

HTK の定義では場合分けは行われない一方で，対数関数の中にバイアス 1 があるため，f が小さい場合は線形関数に近い振る舞いを，f が大きい場合は対数関数に近い振る舞いを見せ，その振る舞いは f の増加に伴い滑らかに変化する。なお，HTK で採用されている定義は，元々 O'Shaughnessy による著書 [148] で示されたものである。

[*7] 2022 年 12 月 25 日現在，librosa の GitHub ページ（https://github.com/librosa/librosa）では約 5,600（画面上「5.6K」としか表示されないため，正確な数値は不明）のスターが付与されている。スターの数の多さは，ライブラリのユーザ数を直接的に表わすものではないが，人気の高さを示す指標となる。

[*8] librosa の実装では，式 (2.39) の右辺の大きさは 15/1000 倍となるように定義されているが，ここでは，理解しやすさのため，1000 Hz が 1000 mel に対応するように修正した。なお，右辺をスカラー倍しても，対数振幅メル周波数スペクトログラム等を求める際には影響がない。

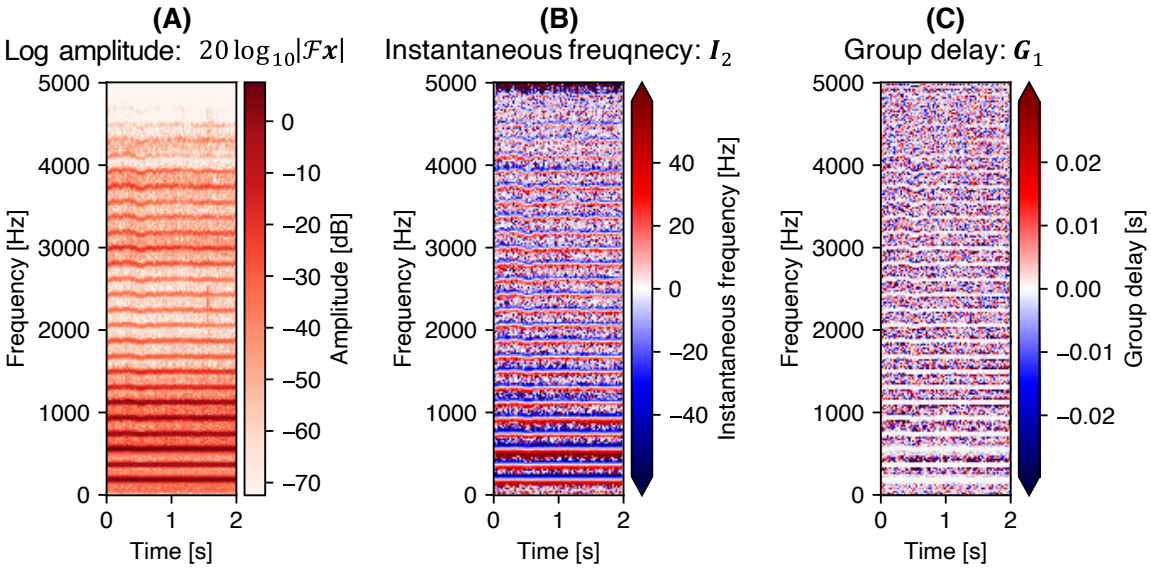


図 2.2 狭帯域分析実行時に得られる，持続母音/a/の音声波形に対する (A) 対数振幅スペクトログラム $20\log_{10}|\mathcal{F}\boldsymbol{x}|$，(B) 瞬時周波数 $\boldsymbol{I}_2$，(C) 群遅延 $\boldsymbol{G}_1$ のヒートマップ。標本化周波数は 10 kHz，窓長は 50 ms，DFT 点数は 1000（100 ms），シフト長は 1 ms とした。(B)，(C) の各図に関して，絶対値の 95 パーセンタイル値を $\theta_{95\%}$ と記すとき，色の表示範囲を $[-\theta_{95\%}, \theta_{95\%}]$ に制限した。

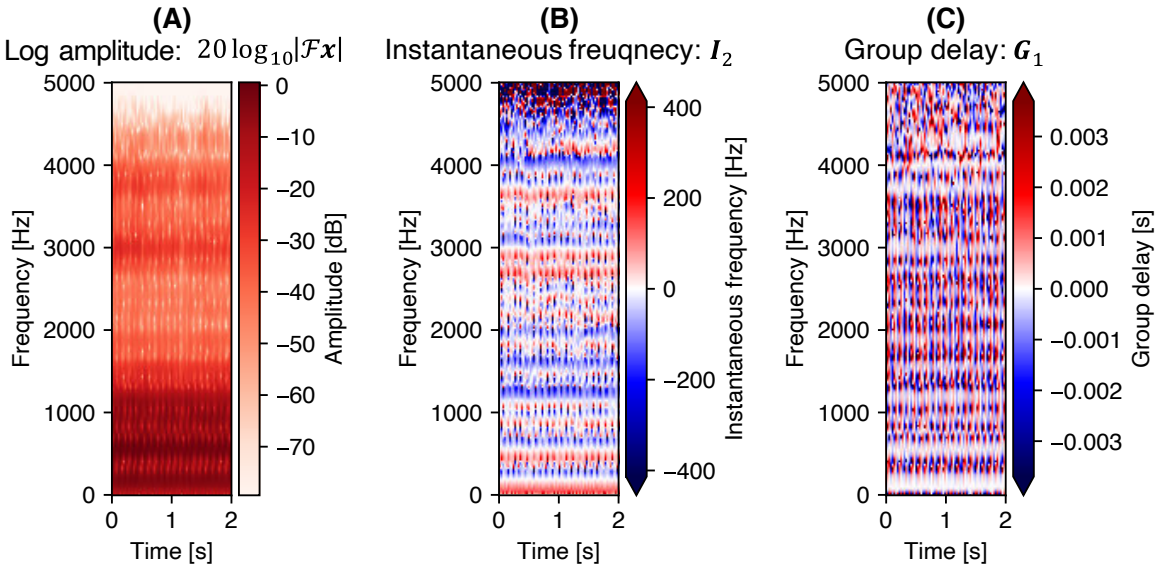


図 2.3 広帯域分析実行時に得られる，持続母音/a/の音声波形の (A) 対数振幅スペクトログラム $20\log_{10}|\mathcal{F}\boldsymbol{x}|$，(B) 瞬時周波数 $\boldsymbol{I}_2$，(C) 群遅延 $\boldsymbol{G}_1$ のヒートマップ。標本化周波数は 10 kHz，窓長は 8 ms，DFT 点数は 1000（100 ms），シフト長は 1 ms とした。(B)，(C) の各図に関して，絶対値の 95 パーセンタイル値を $\theta_{95\%}$ と記すとき，色の表示範囲を $[-\theta_{95\%}, \theta_{95\%}]$ に制限した。

本論文の最初の実験（第 5 章）では，librosa ではデフォルトで Slaney の定義が使われることから，Slaney の定義を採用した。一方，後続の実験（第 6 章，第 7 章）では，深層学習の研究では HTK の定義が主に使われること [105, 149] を鑑みて，HTK の定義を採用した。また，実際に聴覚心理実験で得られたデータ [100] に対するフィッティングでは，Slaney の定義のように定義域を線形領域と対数領域で分割するよりも，HTK の定義のような単一の対数関数によって表現する方が，フィッティング誤差が小さくなることが示されている [144]。このことからも，HTK の定義を採用することはより妥当な選択であると考えられる。

2.3.2　振幅情報のメル尺度表現

従来の音響信号処理で使用されてきた音響特徴量である Mel Frequency Cepstral Coefficients (MFCCs) [101, 146, 147] では，その計算過程でメル尺度が使用される。MFCCs は，パワー（あるいは振幅）スペクトログラムの周波数軸をメル尺度に変換した後にパワー（あるいは振幅）の対数を取り，周波数軸の方向に関して離散コサイン変換を行うことで算出され，典型的にはスペクトル包絡を表す低次成分のみが抽出される。深層学習で多用される対数振幅メル周波数スペクトログラムは，MFCCs の計算過程の離散コサイン変換以降の操作を省略することで算出される。この省略は，離散コサイン変換以降の処理が情報損失や空間情報の破壊を引き起こしてしまうのを防ぐことを意図している [150]。実際，深層学習では，MFCCs よりも対数振幅メル周波数スペクトログラムを用いた方が，様々な課題でより高い性能を実現できることが実験的に示されている [151–154]。また，周波数軸は線形尺度よりメル尺度である方が高性能になりやすいことも，実験的に示唆されている [103, 104]。

メル尺度への周波数軸の変換には，メルフィルタバンクと呼ばれる，周波数領域上で定義されるフィルタバンクが使用される。メルフィルタバンクは，図 2.4 (A) のように，交互に重なりあう三角フィルタの集合として構成される。単位が Hz で表される周波数範囲 $[f_{\min}, f_{\max}]$ を，$K \in \mathbb{N}$ チャンネルのメルフィルタバンクで変換する場合，第 $k \in \{0, 1, \ldots, K-1\}$ 番目の三角フィルタの頂点の周波数は，低い方から順に次のようになる。

$$f\left(m(f_{\min}) + \frac{k}{K+1}\Big(m(f_{\max}) - m(f_{\min})\Big)\right), \tag{2.41}$$

$$f\left(m(f_{\min}) + \frac{k+1}{K+1}\Big(m(f_{\max}) - m(f_{\min})\Big)\right), \tag{2.42}$$

$$f\left(m(f_{\min}) + \frac{k+2}{K+1}\Big(m(f_{\max}) - m(f_{\min})\Big)\right) \tag{2.43}$$

ここで，$m : \mathbb{R}_{\geq 0} \to \mathbb{R}_{\geq 0}$ は Hz を mel に変換する式 (2.39) や式 (2.40) の関数を，

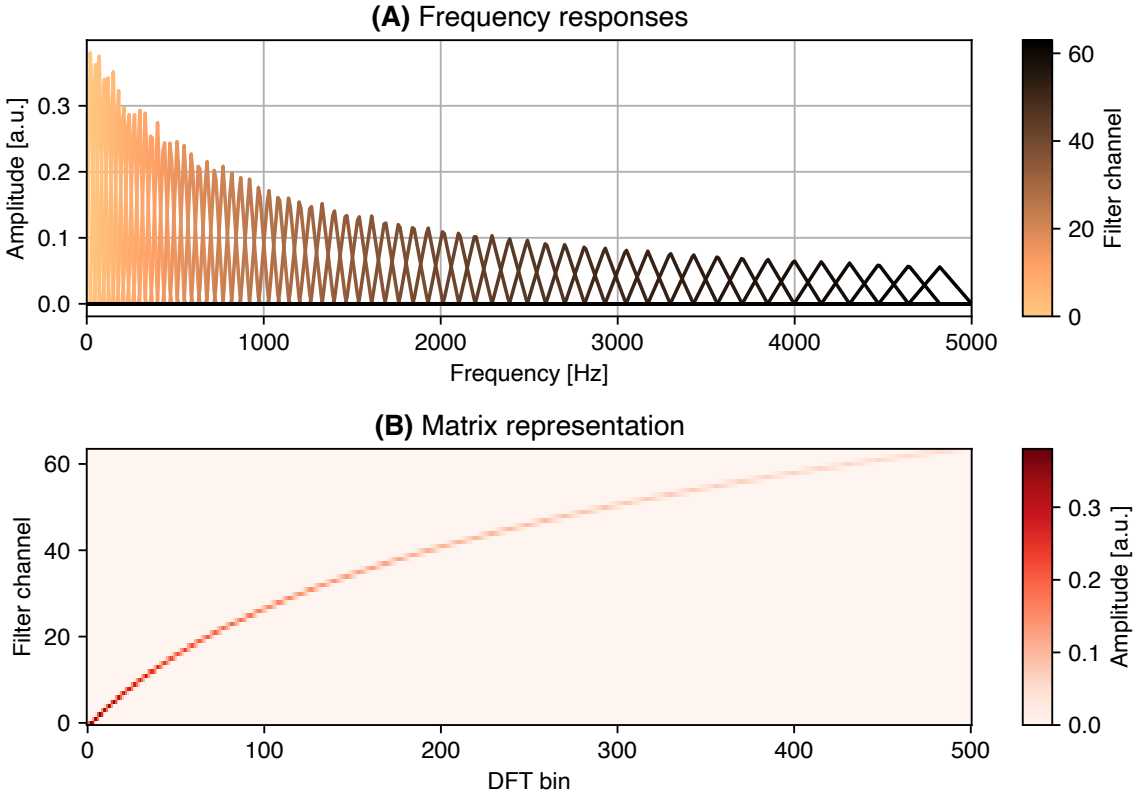


図 2.4 あるメルフィルタバンクに関する，(A) 周波数応答の折れ線グラフと (B) 周波数軸の変換に利用する行列表現 $\boldsymbol{M}$ のヒートマップ。この図のメルフィルタバンクでは，最低周波数は 0 Hz，最高周波数は 5000 Hz，チャンネル数は 64 とし，HTK によるメル尺度の定義を用いた。

$f : \mathbb{R}_{\geq 0} \to \mathbb{R}_{\geq 0}$ は mel を Hz に変換する m の逆関数を表す。典型的には，f_{min} には 0 Hz，f_{max} にはナイキスト周波数が設定される。上式からわかるように，フィルタバンクの各フィルタは，メル尺度上で等間隔に配置される。そのため，図 2.4 (A) に示すように，フィルタ次数が上がる（低域から高域に向かう）につれてフィルタ幅は広くなる。なお，フィルタの大きさに関して，HTK の実装 [147] では最大振幅が 1 になるように，Slaney の実装 [146] では各フィルタの面積が等しくなるように正規化されている。本論文の実験では，Slaney の実装に準じ，各フィルタの l^1 ノルム（すなわち離散面積）が 1 になるように正規化を施した。こうすることによって，メルフィルタバンクの各フィルタを重み付け平均フィルタと見做すことができる。

対数振幅メル周波数スペクトログラムは，図 2.4 (B) のようにメルフィルタバンクを行列で表現し，パワー（あるいは振幅）スペクトログラムと行列の意味で積を取ることによって算出される。パワー（あるいは振幅）スペクトログラム $\boldsymbol{P} \in \mathbb{R}_{\geq 0}^{M \times N} (M, N \in \mathbb{N})$ と，K チャンネルのメルフィルタバンクの行列表現 $\boldsymbol{M} \in \mathbb{R}_{\geq 0}^{K \times M} (K \in \mathbb{N})$ が与えられた

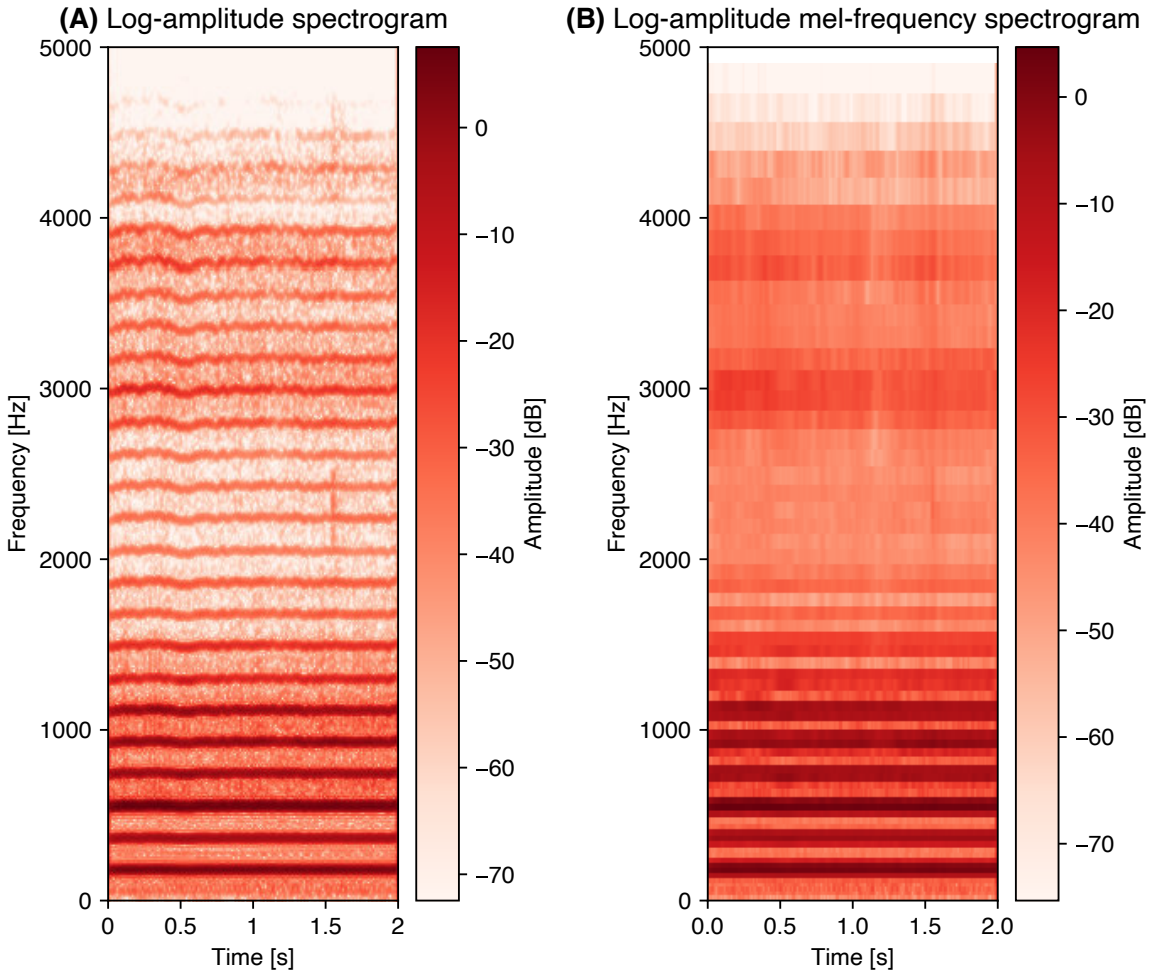


図 2.5 持続母音/a/の音声波形の (A) 対数振幅スペクトログラムと (B) 対数振幅メル周波数スペクトログラムのヒートマップ。短時間フーリエ変換に関して，標本化周波数は 10 kHz，窓長は 50 ms，DFT 点数は 1000（100 ms），シフト長は 1 ms とした。メルフィルタバンクに関して，チャンネル数は 64 とし，HTK によるメル尺度の定義を用いた。

とき，対数振幅メル周波数スペクトログラム $\boldsymbol{P}_{\text{log-mel}} \in \mathbb{R}_{\geq 0}^{K \times N}$ は次のように求まる。

$$\boldsymbol{P}_{\text{log-mel}}[k, n] = \ln\Big((\boldsymbol{M}\boldsymbol{P})[k, n]\Big) \tag{2.44}$$

パワー（あるいは振幅）の単位を dB にしたい場合は，$\ln(\cdot)$ の代わりに $10\log_{10}(\cdot)$（あるいは $20\log_{10}(\cdot)$）を用いる。

図 2.5 は，持続母音/a/の音声波形の対数振幅スペクトログラムと対数振幅メル周波数スペクトログラムのヒートマップを表わす。対数振幅スペクトログラムでは，全ての周波数帯域で周波数分解能が一定なのに対し，対数振幅メル周波数スペクトログラムラムでは，低域ほど周波数分解能が高く，高域ほど周波数分解能が低い様子が観察できる。

2.3.3 位相情報のメル尺度表現

第 2.3.2 節で説明したように，パワーや振幅は式 (2.44) によって周波数軸をメル尺度に変換したら良い。しかし，瞬時周波数や群遅延，位相に対して，パワーや振幅と同じアプローチでメル尺度への変換を行って良いかは議論の余地がある。パワーや振幅と同じアプローチを群遅延に適用した研究もあるが [123, 124]，これらの研究では純粋な意味での群遅延でなく，全極モデルとして表現される声道特性のみに着目し，特異値を巧みに回避して計算される特殊な群遅延が使用されている。瞬時周波数や群遅延に対してパワーや振幅と同じメル周波数化アプローチを用いると，パワーが非常に小さい要素での特異値が，周囲の要素にまで大きな影響を及ぼす可能性がある。また，位相は値が 2π で循環する性質があるため，パワーや振幅と同じメル周波数化アプローチを用いて得られる結果の解釈が自明ではない。本論文では，実験ごとに異なるアプローチを用いて，位相情報のメル尺度表現の計算を試みる。各アプローチの詳細に関しては，各実験の手法の節で説明を行う。

2.4 まとめ

本章では，位相に注意を払いながら，短時間フーリエ変換と，メル尺度による周波数軸のスケール変換に関する説明を行った。後述する本論文の実験では，本章で説明した内容を土台にして手法の構築や考察を行う。

第 3 章

深層学習

3.1 はじめに

機械学習は，与えられたデータに基づいて，何らかの目的に対する性能を自動で改善するコンピュータアルゴリズムである。機械学習の手法は，パラメータ θ を持つ関数 $f(x;\theta)$ として定義される数理モデルと，θ を最適化するアルゴリズムによって特徴づけられる。数理モデルのパラメータの最適化のことを「数理モデルの学習（learning）」と呼び，数理モデルのパラメータを最適化することを「数理モデルが学習する（learn）」あるいは「数理モデルを訓練する（train）」という。

深層学習は機械学習の手法の一つであり，微分可能なパラメータ付き非線形関数である Deep Neural Network（DNN）を数理モデルとし，勾配降下法に基づいて DNN のパラメータ最適化を行う。DNN の起源は，1958 年に Rosenblatt によって発表された機械学習手法である Perceptron [155] にある[*1]。Perceptron は，1943 年に McCulloch と Pitts によって提案された神経細胞の数理表現 [156] に着想を得ていた。このように，DNN の端緒が生物学的な意味での “Neural Network” の模倣であるのは間違いない。

しかし，近年，様々な領域で成果を上げている DNN の多くは，生物学的な意味での “Neural Network” の模倣からは逸脱した複雑な構造を持つ。その一方で，一見複雑な構造も，その構造は数種類の基本的な構成要素に還元される。本論文の実験では，Multilayer Perceptron（MLP），Convolutional Neural Network（CNN），Recurrent Neural Network（RNN）という 3 種類の数理モデルの特徴を組み合わせて DNN を構築した。

本章では，本研究での実験に関連する深層学習の技術に関して説明する。初めに，学習アルゴリズムの根幹である教師あり学習と確率的勾配降下法を紹介する。次に，DNN の最も基本的な数理モデルである MLP を導入した後，DNN の学習上の困難さを緩和する

[*1] Perceptron の数理モデルは微分不可能であり，学習アルゴリズムも勾配降下法には基づかない。

方法について述べ，最後に，MLP 以外の数理モデルである CNN と RNN について説明する。本章の記述にあたり，岡谷による深層学習の成書 [157] を参考にした。なお，本章では，特に断らない限り，ベクトル $\boldsymbol{u} \in \mathbb{R}^N$ の i 番目の要素を u_i，行列 $\boldsymbol{A} \in \mathbb{R}^{N \times M}$ の (j,k) 成分を $A_{j,k}$ と表記する（$N, M \in \mathbb{N}$，$i \in \{1, \dots, N\}$，$j \in \{1, \dots, M\}$）。

3.2　学習アルゴリズム

3.2.1　教師あり学習と損失

教師あり学習（supervised learning）は，入力データとそれに対応する目標の出力データのペアの集合を用いた最適化手法であり，入出力データのペアの集合のことを訓練データセット（training dataset）と呼ぶ。例えば，画像分類では，画像データとその分類ラベルから構成される訓練データセット，本研究で取り扱う GRBAS 尺度の自動評価では，音声データと GRBAS 尺度の評点から構成される訓練データセットを用いる。

ここで，$N \in \mathbb{N}$ サンプルの入力データ $\{\boldsymbol{x}^{(i)}\}_{i=1}^N$ と目標の出力データ $\{\boldsymbol{t}^{(i)}\}_{i=1}^N$ から構成される訓練データセットと，数理モデル $f(\boldsymbol{x}; \theta)$ が与えられたとする。このとき，教師あり学習では，入力データに対する数理モデルの出力 $\{\boldsymbol{y}^{(i)}\}_{i=1}^N$ $(= \{f(\boldsymbol{x}^{(i)}; \theta)\}_{i=1}^N)$ と $\{\boldsymbol{t}^{(i)}\}_{i=1}^N$ の差異が小さくなるように数理モデルを最適化することを目指す。より正式には，実際の出力と目標の出力の間の差異は損失関数（loss function）によって非負の実数である損失（loss）として定量化され，損失の最小化が最適化の指針となる。

目標の出力データが $\{\boldsymbol{t}^{(i)} \in \mathbb{R}^M\}_{i=1}^N$ $(M, N \in \mathbb{N})$ である回帰問題の場合は，損失関数として平均二乗誤差（mean squared error; MSE）が用いられることが多い。

$$\mathrm{MSE}(\{\boldsymbol{y}^{(i)}\}_{i=1}^N, \{\boldsymbol{t}^{(i)}\}_{i=1}^N) := \frac{1}{N} \sum_{i=1}^N \left\| \boldsymbol{y}^{(i)} - \boldsymbol{t}^{(i)} \right\|_{l^2}^2 \tag{3.1}$$

ここで，$\| \cdot \|_{l^2}$ は l^2 ノルムを表す。

目標の出力データが $\{\boldsymbol{t}^{(i)} \in [0,1]^M\}_{i=1}^N$ $(M, N \in \mathbb{N})$ である M クラス分類問題の場合は，損失関数として交差エントロピー誤差（cross entropy loss）が用いられることが多い。

$$\text{cross-entropy}(\{\boldsymbol{y}^{(i)}\}_{i=1}^N, \{\boldsymbol{t}^{(i)}\}_{i=1}^N) := \frac{1}{N} \sum_{i=1}^N \sum_{j=1}^M t_j^{(i)} \ln y_j^{(i)} \tag{3.2}$$

上式より，$\boldsymbol{y}^{(i)}$ に非正の要素が含まれる場合，損失 L は定義されないか複素数になる可能性がある。しかし，通常，出力には softmax 関数が適用されて $\boldsymbol{y} \in (0,1)^M$ であるため，この問題は回避される。なお，$\boldsymbol{t}^{(i)}$ が，正解に対応する要素のみが 1 で，その他の要素が 0 で表されている場合，正解が one-hot 表現（one-hot encoding）で表現されているという。そのほかに，$\boldsymbol{t}^{(i)}$ はラベルの離散確率分布等を表すこともある。

3.2.2 確率的勾配降下法

深層学習では，損失を最小化するために，勾配降下法（gradient descent）に基づく手法を用いる。勾配降下法は，パラメータに関する損失の勾配計算と，その勾配に基づくパラメータ更新を反復することによって損失の最小化を目指す手法である。

標準的な勾配降下法は, バッチ勾配降下法（batch gradient descent）とも呼ばれ, 毎回のパラメータ更新で訓練データセットのすべてのデータが使用される。$N \in \mathbb{N}$ サンプルの入力データ $\{\boldsymbol{x}^{(i)}\}_{i=1}^{N}$ と目標の出力データ $\{\boldsymbol{t}^{(i)}\}_{i=1}^{N}$ から構成される訓練データセットと，数理モデル $f(\boldsymbol{x};\theta)$ が与えられたとする。また, 実際の出力 $\{\boldsymbol{y}^{(i)}\}_{i=1}^{N}$ と目標の出力 $\{\boldsymbol{t}^{(i)}\}_{i=1}^{N}$ の部分集合に対する損失関数を $g : \{\boldsymbol{y}^{(i)}\}_{i\in\Lambda} \times \{\boldsymbol{t}^{(i)}\}_{i\in\Lambda} \to \mathbb{R}$ $(\Lambda \subset \{1,\dots,N\})$ とする。値が $t \in \mathbb{Z}_{\geq 0}$ 回更新されたパラメータを $\theta^{(t)}$ と記すとき，パラメータの更新式は次式のように表される。

$$\theta^{(t)} = \theta^{(t-1)} - \eta \left(\left. \frac{\partial L_{\text{batch}}}{\partial \theta} \right|_{\theta=\theta^{(t)}} \right)^{\top} \tag{3.3}$$

$$L_{\text{batch}}(\theta) = g\left(\{f(\boldsymbol{x}^{(i)};\theta)\}_{i=1}^{N}, \{\boldsymbol{t}^{(i)}\}_{i=1}^{N} \right) \tag{3.4}$$

ここで，$L_{\text{batch}}(\theta)$ は訓練データセット全体に対する損失関数であり，θ の関数として表現される。$\partial/\partial\theta$ はパラメータ θ に関するヤコビ行列であり，その転置は勾配に対応する。$\eta > 0$ は学習率（learning rate）と呼ばれるハイパーパラメータであり，一回あたりのパラメータ更新量を制御する。学習率は，小さすぎるとパラメータの収束までの反復回数が過度に増え，大きすぎるとうまく収束が起こらないため，適切に設定する必要がある。

バッチ勾配降下法では，パラメータ更新の度に同じ損失関数 L_{batch} を使う。しかし，DNN では一般に $\partial L_{\text{batch}}/\partial\theta$ が凸関数ではないため，パラメータがどの極小値に収束するかは初期パラメータ $\theta^{(0)}$ の値に大きく依存してしまう（図 3.1）。また，計算上の問題として，訓練データセットのサンプル数が多い場合，計算時間が膨大になったり，メモリ不足になったりすることがある。

これらの勾配降下法の問題は，確率的勾配降下法（stochastic gradient descent）によって解決される。確率的勾配降下法では，訓練データセットの真部分集合であるミニバッチ（minibatch）を使ってパラメータ更新を行う。ミニバッチは，訓練データセットの添字集合 $\mathcal{I} = \{1,\dots,N\}$ の真部分集合 $\mathcal{I}^{(t)} \subsetneq \mathcal{I}$ $(t \in \mathbb{N})$ を使って，$\{\boldsymbol{x}^{(i)}\}_{i\in\mathcal{I}^{(t)}}, \{\boldsymbol{t}^{(i)}\}_{i\in\mathcal{I}^{(t)}}$ と表される。1 つのミニバッチに含まれるサンプル数のことをバッチサイズ（batch size）と呼ぶ[*2]。標準的には，ミニバッチの系列は，添字集合 $\mathcal{I}$ を互いに素な真部分集合に分割

[*2] ミニバッチサイズ（minibatch size）という用語を使う場合もあるが，バッチサイズという言葉の方が頻用される。深層学習の分野では，ミニバッチのことを単にバッチと呼ぶことが多い。

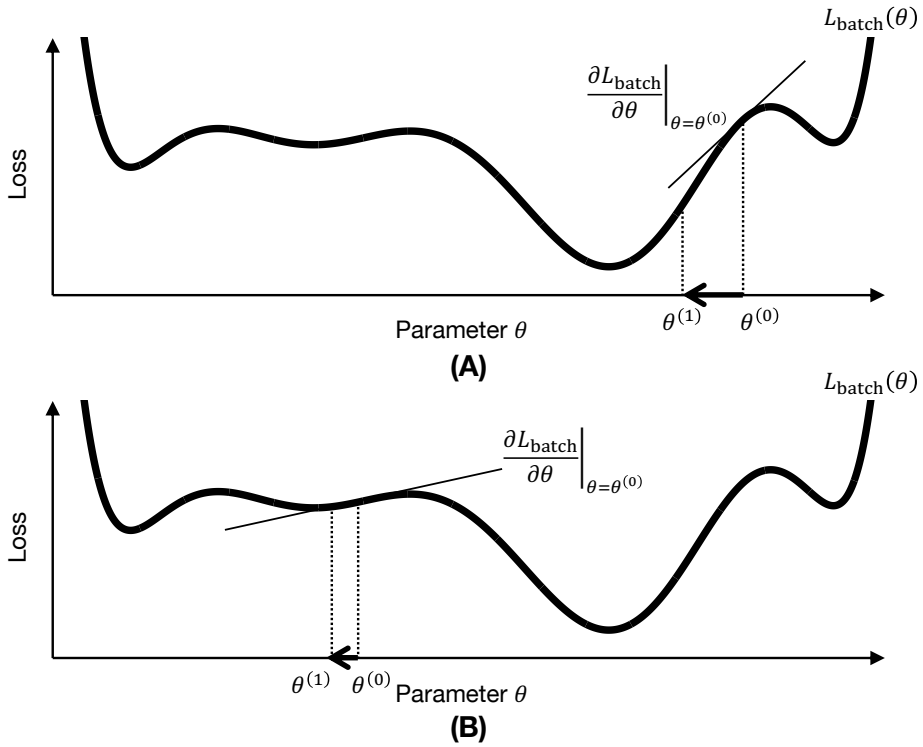


図 3.1 バッチ勾配降下法によるパラメータ更新の概念図。簡単のために，パラメータ θ は 1 次元であると仮定している。深い極小値に収束するか (A)，浅い極小値に収束するか (B) は，初期パラメータ $\theta^{(0)}$ の値に大きく依存する。

する操作を繰り返すことによって生成される。

$$\mathcal{I} = \mathcal{I}^{(1)} \sqcup \mathcal{I}^{(2)} \sqcup \cdots \sqcup \mathcal{I}^{(N/B)} \tag{3.5}$$

$$= \mathcal{I}^{(N/B+1)} \sqcup \mathcal{I}^{(N/B+2)} \sqcup \cdots \sqcup \mathcal{I}^{(2N/B)} = \cdots \tag{3.6}$$

ここで，$B \in \mathbb{N}$ はバッチサイズを表し，訓練用データセットのサンプル数 N が B で割り切れると仮定した。N が B で割り切れない場合，剰余 $N \bmod B$ を捨てるか使うかはオプションであり，剰余を使う場合，剰余に対応するミニバッチのバッチサイズを $N \bmod B$ にする等の処理を行う。なお，訓練データセットの学習を一周行うことを 1 エポック（epoch）と数える。1 エポックで行われるパラメータ更新の回数は，N と B に依存して決まる。

上述の手続きによって生成されたミニバッチを用いて，確率的勾配降下法では次式のよ

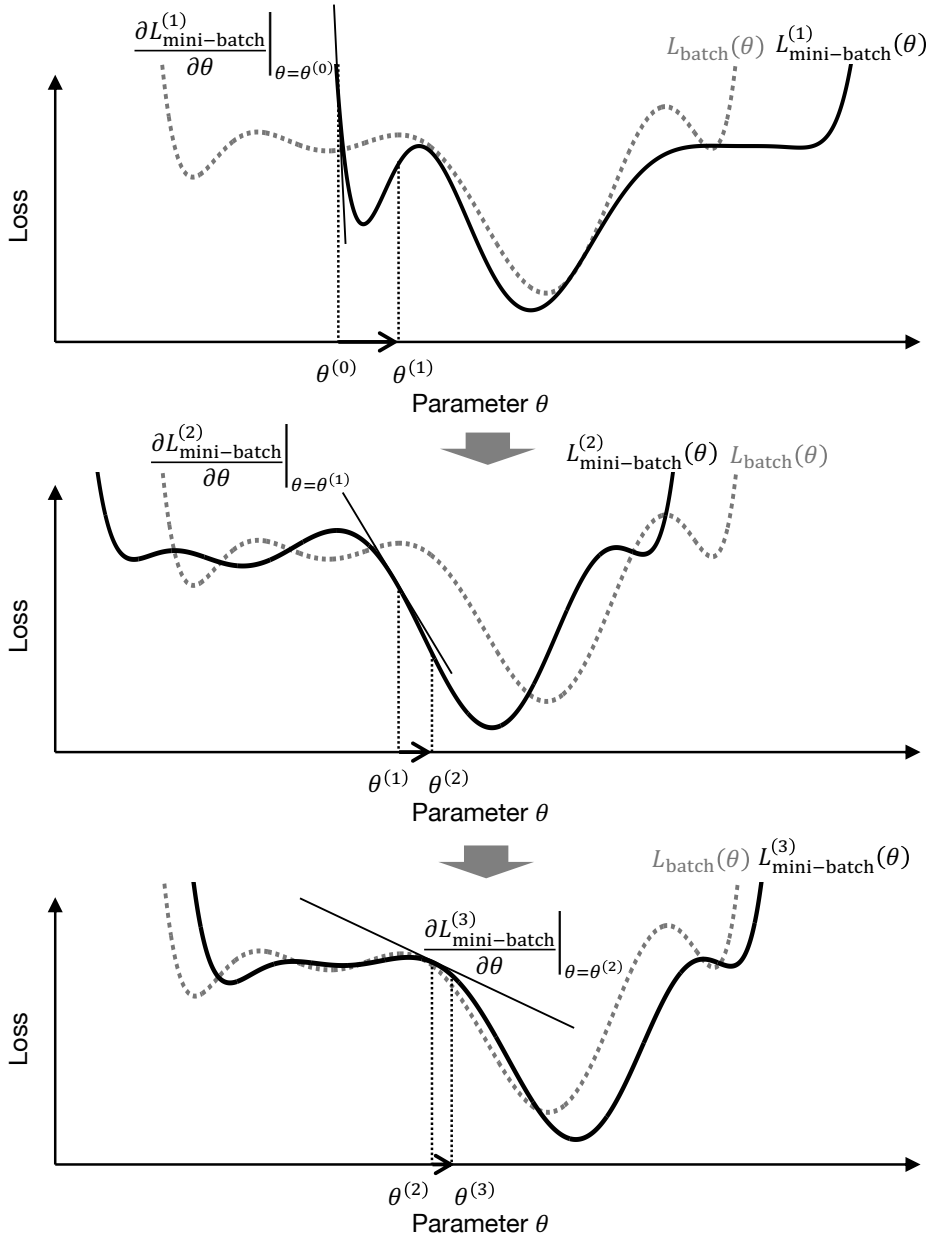


図 3.2 確率的勾配降下法によるパラメータ更新の概念図。簡単のために，パラメータ θ は 1 次元であると仮定している。

うにパラメータが更新される。

$$\theta^{(t)} = \theta^{(t-1)} - \eta \left(\frac{\partial L^{(t)}_{\text{mini-batch}}}{\partial\theta}\bigg|_{\theta=\theta^{(t-1)}} \right)^{\top} \tag{3.7}$$

$$L^{(t)}_{\text{mini-batch}}(\theta) = g\left(\{f(\boldsymbol{x}^{(i)};\theta)\}_{i\in\mathcal{I}^{(t)}}, \{\boldsymbol{t}^{(i)}\}_{i\in\mathcal{I}^{(t)}} \right) \tag{3.8}$$

確率的勾配降下法では，パラメータ更新の度に勾配計算に使われる損失関数 $L^{(t)}_{\text{mini-batch}}$ が変化するため，訓練データセット全体に対する損失関数 L_{batch} の浅い極小値に収束するリスクが低くなる（図 3.2）。バッチサイズが小さいほど，パラメータ更新の度の損失関数の変動が大きくなり，より良い極小値[*3]に収束しやすくなるということが示唆されている [158, 159]。しかし，バッチサイズが小さすぎると，損失関数の変動が大きくなりすぎてパラメータの収束を妨げる場合もあるため，バッチサイズは適切に設定する必要がある。

3.2.3　確率的勾配降下法の派生手法

前節で触れたように，確率的勾配降下法では，パラメータの収束が学習率やバッチサイズの値に大きく依存する上に，ミニバッチ選択のランダム性のためにパラメータ更新量がばらつく。また，確率的勾配降下法では，すべてのパラメータに対して同じ学習率が適用されるが，パラメータ毎に異なる学習率を適用した方が良いことが指摘されている [160]。本節では，上記の問題点の解消を目的とした確率的勾配降下法の派生手法を紹介する。なお，簡単のため，本節では t 回目の更新におけるパラメータ $\theta \in \mathbb{R}^D$ $(D \in \mathbb{N})$ の勾配を $\boldsymbol{g}^{(t)}$ と表記する。

$$g_i^{(t)} := \left. \frac{\partial L^{(t)}_{\text{mini-batch}}}{\partial \theta_i} \right|_{\theta=\theta^{(t-1)}} \tag{3.9}$$

モメンタム法（momentum） [161] は，パラメータ更新量のばらつきの抑制を目的とした手法であり，次式のようにパラメータが更新される。

$$m_i^{(t)} = \alpha m_i^{(t-1)} + \eta g_i^{(t)} \tag{3.10}$$

$$\theta_i^{(t)} = \theta_i^{(t-1)} - m_i^{(t)} \tag{3.11}$$

ここで，$\boldsymbol{m}^{(t)} \in \mathbb{R}^D$ $(\boldsymbol{m}^{(0)} = \boldsymbol{0})$，$\alpha \in (0, 1)$ はハイパーパラメータ，$\eta > 0$ は学習率である。式 (3.11) を解くと，$m_i^{(t)}$ は過去の勾配の加重移動平均であることがわかる（但し，重みの平均は 1 ではない）。

$$m_i^{(t)} = \sum_{k=1}^{t} \eta \alpha^{t-k} g_i(k) \tag{3.12}$$

モメンタム法は，パラメータ更新量を過去の勾配の加重移動平均とすることで，パラメータ更新量の正負の振動を抑制している。

[*3] より正確には，極小値付近でパラメータを摂動させても損失が大きく変化しないような，平坦な（flat）極小値のこと。平坦な極小値に収束すると，汎化性能（後述）が高くなりやすいと考えられている [158]。

RMSProp [162, 163] は，パラメータ毎にその更新量を適応的に調節する手法であり，次式のようにパラメータが更新される[*4]。

$$v_i^{(t)} = \alpha v_i^{(t-1)} + (1-\alpha)(g^{(t)})^2 \tag{3.13}$$

$$\theta_i^{(t)} = \theta_i^{(t-1)} - \frac{\eta}{\sqrt{v_i^{(t)}} + \epsilon} g^{(t)} \tag{3.14}$$

ここで，$\boldsymbol{v}^{(t)} \in \mathbb{R}^D_{\geq 0}$ $(\boldsymbol{v}^{(0)} = \boldsymbol{0})$，$\alpha \in (0,1)$ はハイパーパラメータ，$\epsilon > 0$ はゼロ除算を防ぐ微小量，$\eta > 0$ は学習率である。$\sqrt{v_i^{(t)}}$ は，加重移動平均のアプローチで計算された過去の勾配の標準偏差である。RMSProp は，過去の勾配の標準偏差を用いることによって，頻繁に大きな勾配を取るパラメータに関してはその更新量を抑制する一方で，あまり大きな勾配を取らないパラメータに関してはその更新量を増加させる。

Adam [164] は，モメンタム法と RMSProp の思想を組み合わせたような手法であり，次式のようにパラメータが更新される。

$$m_i^{(t)} = \beta_1 m_i^{(t-1)} + (1-\beta_1) g_i^{(t)} \tag{3.15}$$

$$v_i^{(t)} = \beta_2 v_i^{(t-1)} + (1-\beta_2)(g^{(t)})^2 \tag{3.16}$$

$$\hat{m}_i^{(t)} = \frac{m_i^{(t)}}{1-\beta_1^t} \tag{3.17}$$

$$\hat{v}_i^{(t)} = \frac{v_i^{(t)}}{1-\beta_2^t} \tag{3.18}$$

$$\theta_i^{(t)} = \theta_i^{(t-1)} - \frac{\eta}{\sqrt{\hat{v}_i^{(t)}} + \epsilon} \hat{m}_i^{(t)} \tag{3.19}$$

ここで，$\boldsymbol{m}^{(t)}, \hat{\boldsymbol{m}}^{(t)} \in \mathbb{R}^D$ $(\boldsymbol{m}^{(0)} = \hat{\boldsymbol{m}}^{(0)} = \boldsymbol{0})$，$\boldsymbol{v}^{(t)}, \hat{\boldsymbol{v}}^{(t)} \in \mathbb{R}^D_{\geq 0}$ $(\boldsymbol{v}^{(0)} = \hat{\boldsymbol{v}}^{(0)} = \boldsymbol{0})$，$\beta_1, \beta_2 \in (0,1)$ はハイパーパラメータ，$\epsilon > 0$ はゼロ除算を防ぐ微小量，$\eta > 0$ は学習率である。$\hat{\boldsymbol{m}}^{(t)}$ と $\hat{\boldsymbol{v}}^{(t)}$ は，学習の初期段階（t が小さいとき）におけるバイアスを補正した平均と分散に対応する。

本論文の実験では，パラメータの更新方法として Adam を使用した。また，Adam のハイパーパラメータの値としては，Kingma と Ba [164] によって推奨されている $\beta_1 = 0.9$，$\beta_2 = 0.999$，$\epsilon = 1 \times 10^{-8}$ を設定した。

3.3 数理モデル①：Multilayer Perceptron

MLP は最も基本的なネットワークであり，その起源は Perceptron [155] にある。しかし，Perceptron から順を追って考えるより，線形回帰を発展させる方が見通しが良いた

[*4] RMSProp の定義は文献によって異なるが [162, 163]，ここでは Ruder の文献 [162] に倣った。

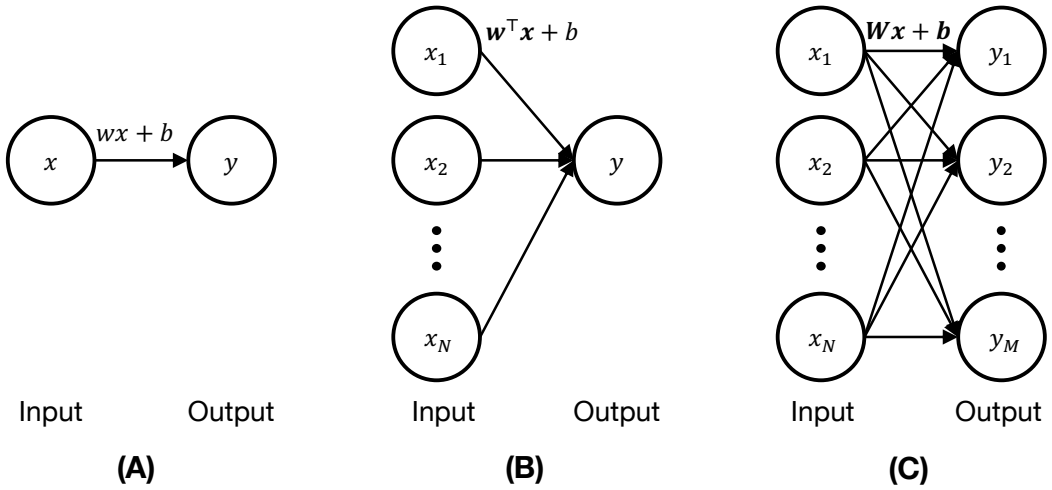


図 3.3 線形回帰モデル。(A) は 1 入力 1 出力（単回帰），(B) は多入力 1 出力（重回帰），(C) は多入力多出力のモデルに対応する。

め，本節では，線形回帰の導入から始め，MLP へと発展させる。より具体的には，単回帰，重回帰，ロジスティック回帰，MLP の順にモデルを発展させていくが，このようなモデルの発展のさせ方は，Stanford 大学の非常勤教授（adjunct professor）である人工知能研究者の Andrew Ng による講義 [165] を参考にした。

3.3.1　線形回帰

ある説明変数 $x \in \mathbb{R}$ から目的変数 $y \in \mathbb{R}$ を予測する回帰問題を解く場合，最も簡単な数理モデルは，図 3.3 (A) に示す 1 入力 1 出力の単回帰モデルであり，次式で表される。

$$y(x; w, b) = wx + b \tag{3.20}$$

ここで，$w \in \mathbb{R}$ は重み（weight），$b \in \mathbb{R}$ はバイアス（bias）を表す。説明変数を多次元化し，N 次元（$N \in \mathbb{N}$）実ベクトル $\boldsymbol{x} \in \mathbb{R}^N$ から $y \in \mathbb{R}$ を予測する場合には，図 3.3 (B) に示す多入力 1 出力の重回帰モデルを用いる。

$$y(\boldsymbol{x}; \boldsymbol{w}, b) = \boldsymbol{w}^\top \boldsymbol{x} + b \tag{3.21}$$

ここで，$\boldsymbol{w} \in \mathbb{R}^N$ は重み，$b \in \mathbb{R}$ はバイアスを表す。説明変数だけでなく目的変数も多次元化し，N 次元実ベクトル $\boldsymbol{x} \in \mathbb{R}^N$ から M 次元実ベクトル $y \in \mathbb{R}$ を予測する場合には，図 3.3 (C) に示す多入力多出力のモデルを用いる。

$$\boldsymbol{y}(\boldsymbol{x}; \boldsymbol{W}, \boldsymbol{b}) = \boldsymbol{W}\boldsymbol{x} + \boldsymbol{b} \tag{3.22}$$

ここで，$\boldsymbol{W} \in \mathbb{R}^{M \times N}$ は重み，$\boldsymbol{b} \in \mathbb{R}^M$ はバイアスを表す。

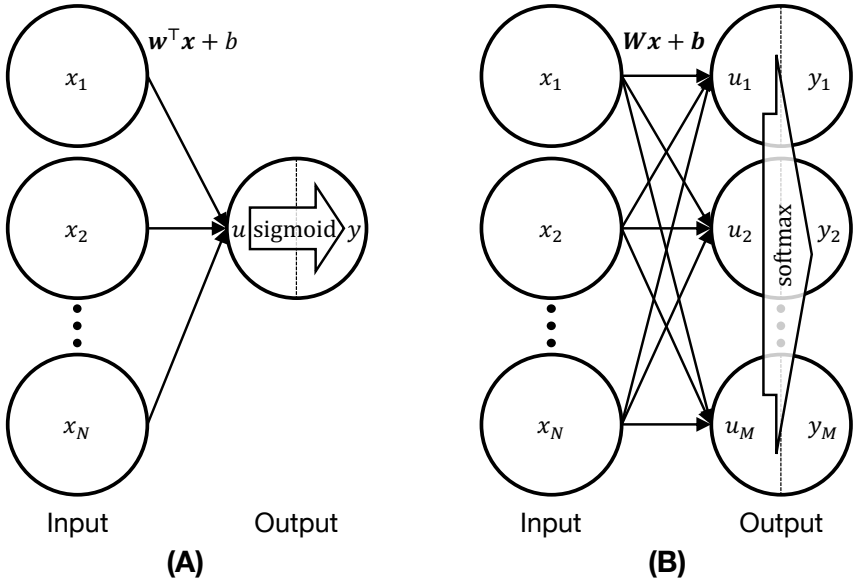


図 3.4 ロジスティック回帰モデル。(A) は 2 値分類，(B) は M クラス分類のためのモデルに対応する。

3.3.2 ロジスティック回帰

ロジスティック回帰 [166] は，分類問題のための数理モデルであり，線形回帰モデルの出力を確率と見做せるように修正することによって導出される。2 値分類問題のためのロジスティック回帰モデル（図 3.4 (A)）は，1 出力線形回帰モデルの出力の範囲を $(0, 1)$ に制限して確率と見做すことによって得られる。

$$y(\boldsymbol{x}; \boldsymbol{w}, b) = \mathrm{sigmoid}(\boldsymbol{w}^\top \boldsymbol{x} + b) \tag{3.23}$$

$$\mathrm{sigmoid}(u) := \frac{1}{1 + e^{-u}} \tag{3.24}$$

ここで，$\boldsymbol{x} \in \mathbb{R}^N$ は説明変数，$\boldsymbol{w} \in \mathbb{R}^N$ は重み，$b \in \mathbb{R}$ はバイアスを表し，$\mathrm{sigmoid} : \mathbb{R} \to (0, 1)$ は sigmoid 関数と呼ばれる関数である。sigmoid 関数によって，出力 y の範囲は $(0, 1)$ に制限されるため，y を確率と見做すことが可能となる。

M クラス（$M \in \mathbb{N}$）分類問題のためのロジスティック回帰モデル（図 3.4 (B)）は，M 出力線形回帰モデルの出力を離散確率分布と見做せるよう修正することによって得られる。

$$\boldsymbol{y}(\boldsymbol{x}; \boldsymbol{W}, \boldsymbol{b}) = \mathrm{softmax}(\boldsymbol{W}\boldsymbol{x} + \boldsymbol{b}) \tag{3.25}$$

ここで，$\boldsymbol{x} \in \mathbb{R}^N$ は説明変数，$\boldsymbol{W} \in \mathbb{R}^{M \times N}$ は重み，$\boldsymbol{b} \in \mathbb{R}^M$ はバイアスを表す。

$\mathrm{softmax} : \mathbb{R}^n \to (0,1)^n$ $(\forall n \in \mathbb{N})$ は softmax 関数と呼ばれ，次のように定義される。

$$\mathrm{softmax}(\boldsymbol{u})_i := \left(\frac{e^{u_i}}{\sum_j e^{u_j}} \right) \tag{3.26}$$

$\mathrm{softmax}(\boldsymbol{u})_i \in (0,1)$ は $\mathrm{softmax}(\boldsymbol{u}) \in (0,1)^n$ の第 i 成分を表す。定義から明らかであるが，softmax 関数は次の性質を満たす。

$$\|\mathrm{softmax}(\boldsymbol{u})\|_{l^1} = 1 \tag{3.27}$$

$$u_i > 0 \qquad (i = 1, 2, \ldots, n) \tag{3.28}$$

ここで，$\|\cdot\|_{l_1}$ は l_1 ノルムを表す。上記の性質は，softmax 関数は，任意の実ベクトルを l^1 ノルムが 1 である非負の実ベクトルに変換することを意味する。そのため，softmax 関数の出力は離散確率分布と見做すことができ，出力の最大値を取る成分が最尤クラスであると解釈される。

3.3.3 Multilayer Perceptron

線形回帰は説明変数と目的変数の間の線形性を仮定し，ロジスティック回帰は説明変数によってクラスの集合が線形分離可能であることを仮定する。しかし，問題の対象となる現象は，線形関係で近似的に表現可能であるとは限らない。線形関係で近似できない現象は，線形回帰やロジスティック回帰では十分にモデリングできない。この問題を解決するために考案されたのが MLP である。MLP は，アフィン変換と非線形変換の反復によって構成されており，非線形な現象のモデリングを行うことができる。

MLP は，回帰問題にも分類問題にも応用することができるが，ここでは，例として多クラス分類への応用を考える。図 3.5 に示す M クラス（$M \in \mathbb{N}$）分類問題のための MLP モデルでは，以下の演算が行われる。

$$\boldsymbol{y}(\boldsymbol{x}; \boldsymbol{W}^{(1)}, \boldsymbol{b}^{(1)}, \boldsymbol{W}^{(2)}, \boldsymbol{b}^{(2)}) = \left(\mathrm{softmax} \circ f^{(2)} \circ \tanh \circ f^{(1)} \right)(\boldsymbol{x}) \tag{3.29}$$

$$f^{(k)}(\boldsymbol{z}; \boldsymbol{W}^{(k)}, \boldsymbol{b}^{(k)}) = \boldsymbol{W}^{(k)}(z) + \boldsymbol{b}^{(k)} \qquad (k = 1, 2) \tag{3.30}$$

ここで，$\boldsymbol{W}^{(1)} \in \mathbb{R}^{L \times N}, \boldsymbol{W}^{(2)} \in \mathbb{R}^{M \times L}$ は重み，$\boldsymbol{b}^{(1)} \in \mathbb{R}^L, \boldsymbol{b}^{(2)} \in \mathbb{R}^M$ はバイアスである[*5][*6]。この MLP モデルは，図 3.4 (B) に示すロジスティック回帰モデルの演算の中間

[*5] 式 (3.29) では，1 変数実関数であるはずの tanh 関数を，多次元ベクトルに作用するベクトル関数として見做している。このような記法は，深層学習の分野でよく用いられ，ここでは tanh 関数をベクトルの各成分に作用させることを意味する（$\tanh(\boldsymbol{x})_i = \tanh(x_i)$）。以降，本論文では特に断らずにこの記法を用いる。

[*6] 回帰問題に応用したい場合は，最後の softmax 関数を除けば良い。2 値分類問題に応用したい場合は，最後の出力を 2 次元とするか，最後の出力を 1 次元とした上で softmax 関数を sigmoid 関数に置換すれば良い。

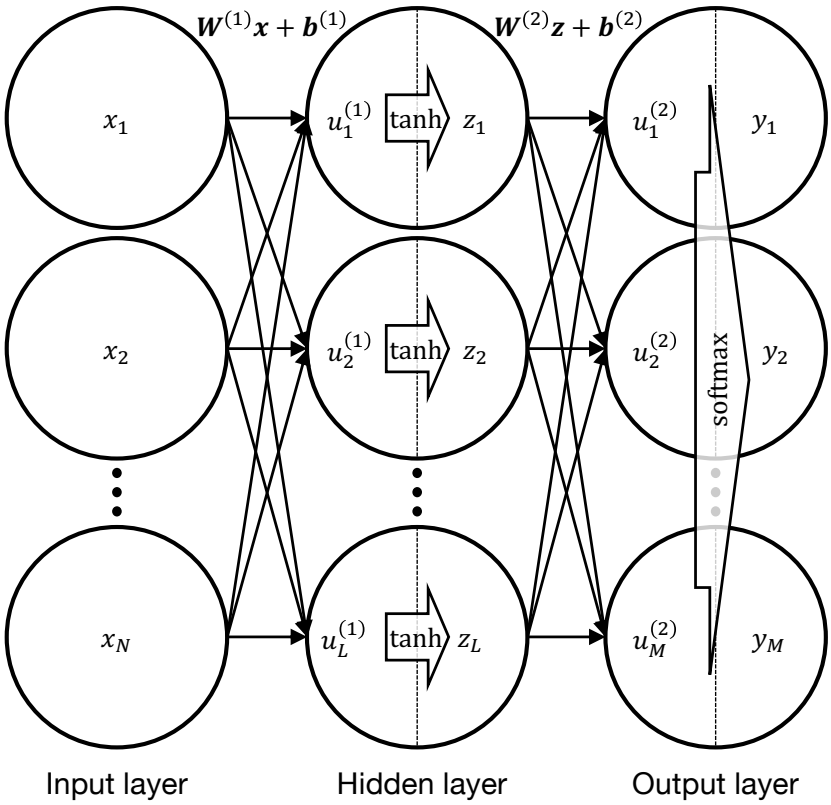


図 3.5 M クラス分類問題のための 3 層 MLP

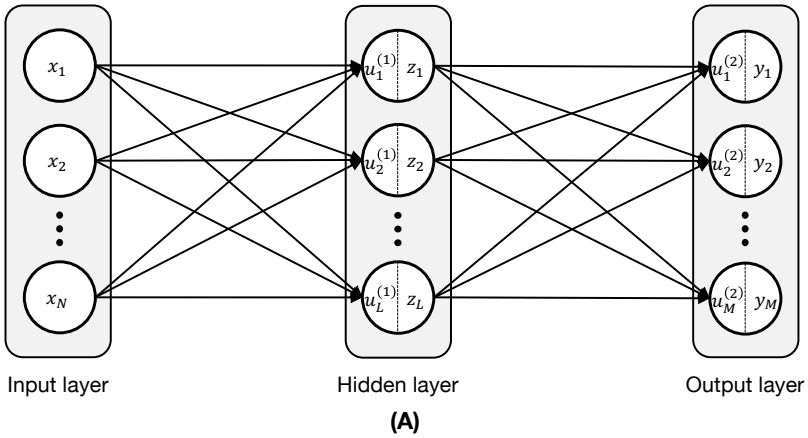
に，非線形関数である tanh 関数とアフィン変換を挿入することによって得られる[*7]。

MLP の名称に含まれる層（layer）という言葉は，ニューロン（neuron）の層を意味している。MLP を生物学的なモデルとして捉える場合，図 3.5 中の一つ一つの丸記号（◯）はニューロンを表し，MLP はニューロンの層が積み重なったものであると考える。このような解釈に基づくと，図 3.5 に示される MLP は 3 層構造であると見做され，特に，最初の層を入力層（input layer），最後の層を出力層（output layer），それ以外の中間の層を隠れ層（hidden layer）と呼ぶ。この例では隠れ層の数は一つとしたが，一般に，MLP は一つ以上の隠れ層を含む。また，tanh 関数やシグモイド関数，softmax 関数のような非線形関数は，ニューロンの発火現象の数理モデリングに着想を得ていることから，活性化関数（activation function）と呼ばれる。

3.3.4 ニューロンの層と全結合層

先述のように，MLP という名称中の「層」はニューロンの層を表しており，この見方では，演算の入出力に現れるベクトルに着目している。その一方で，現代の深層学習で

[*7] アフィン変換だけでなく非線形関数も挿入されているということが非常に重要である。仮に非線形変換がなかった場合，$\boldsymbol{W}^{(2)}(\boldsymbol{W}^{(1)}\boldsymbol{x}+\boldsymbol{b}^{(1)})+\boldsymbol{b}^{(2)}=(\boldsymbol{W}^{(2)}\boldsymbol{W}^{(1)})\boldsymbol{x}+(\boldsymbol{W}^{(2)}\boldsymbol{b}^{(1)}+\boldsymbol{b}^{(2)})$ となる。つまり，2 つのアフィン変換の合成は 1 つのアフィン変換と等価になるだけである。



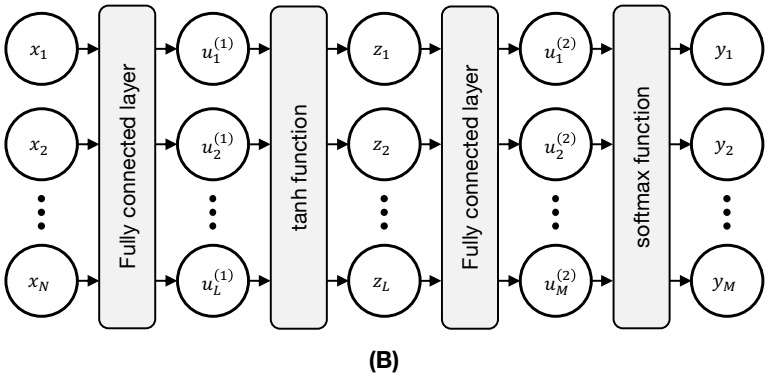


図 3.6 M クラス分類問題のための 3 層 MLP（図 3.5）の二つの表現方法。ニューロンの層の数は 3 つである一方 (A)，全結合層の数は 2 つである (B)。

は，演算そのものを「層」と呼び，MLP に含まれるアフィン変換を特に全結合層（fully connected layer）と呼ぶ。そのため，図 3.6 に示すように，3 層 MLP は，ニューロンの層の数は 3 つである一方，全結合層の数は 2 つである。また，演算に着目する現代の深層学習では，入出力間に現れる途中の演算結果を表すために，隠れ層ではなく，隠れ表現（hidden representation）という用語を用いる。

現代的な DNN の多くは，生物学的な意味での “Neural Network” の模倣から逸脱しており，DNN をニューロンの層と解釈することは困難かつ不適切な場合がある。演算そのものを層と見做した方が，生物学的モデリングの文脈に縛られずに柔軟である他，DNN の構造を理解しやすくなるという利点がある。よって，本論文ではこれ以降，特に断らない限り，「層」という用語は演算を意味するために使用する。

3.4　誤差逆伝播法

DNN の学習には，確率的勾配降下法またはその派生手法が用いられる。しかし，DNN は大量のパラメータを持つ深い入れ子構造の合成関数であり，その勾配を数値計

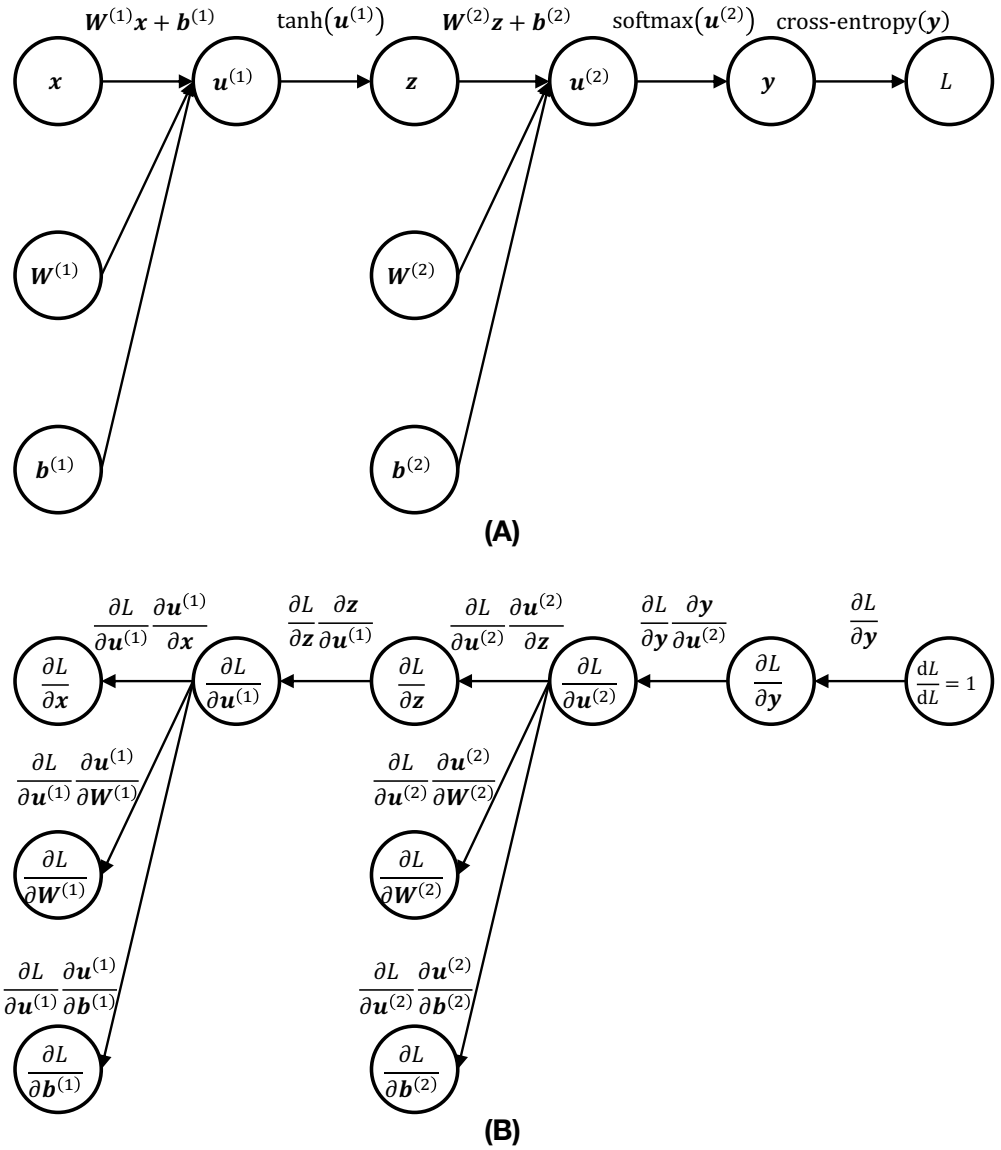


図 3.7 M クラス分類問題のための 3 層 MLP（図 3.5）の順伝播 (A) と逆伝播 (B) の計算グラフ

算する方法は自明ではない。最も素朴なアプローチの一つである数値微分（numerical differentiation）は，パラメータ数に比例して計算コストが増加する上に，計算誤差が大きくなりやすい。そこで，DNN の勾配は，一般に誤差逆伝播法（backpropagation）という手法を用いて，高効率かつ高精度で数値計算される。

誤差逆伝播法は，微分の連鎖律（chain rule）に基づき，損失関数の勾配を逐次的に計算する手法である。例えば，損失関数が，入力 $\boldsymbol{x}$ に対する合成関数 $L(\boldsymbol{x}) = f^{(K)}(f^{(K-1)}(\cdots f^{(2)}(f^{(1)}(\boldsymbol{x}))\cdots))$ で与えられ，中間の出力が $\boldsymbol{z}^{(k)}(\boldsymbol{x}) = f^{(k)}(f^{(k-1)}(\cdots f^{(2)}(f^{(1)}(\boldsymbol{x}))\cdots))$ で表されたとする（但し，$K \in \mathbb{N}$,

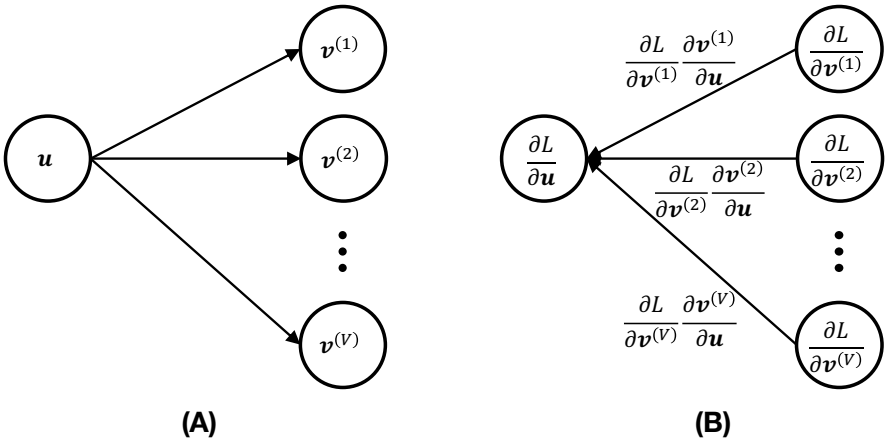


図 3.8 出力が分岐する場合の順伝播 (A) と逆伝播 (B) の計算グラフ。逆伝播では，$\partial L/\partial \boldsymbol{u} = \sum_{k=1}^{V} \partial L/\partial \boldsymbol{v}^{(k)}$ という計算が行われる。

$k = 0, 1, \ldots, K$，$\boldsymbol{z}^{(0)} = \boldsymbol{x}$，$f^{(k)} : \mathbb{R}^{D^{(k-1)}} \to \mathbb{R}^{D^{(k)}}$，$D^{(k)} \in \mathbb{N}$，$D^{(K)} = 1$)。具体的に入力として $\hat{\boldsymbol{x}}$ が与えられたとき，中間出力を $\hat{\boldsymbol{z}}^{(k)} = \boldsymbol{z}^{(k)}(\hat{\boldsymbol{x}})$ と表すと，外側の関数から内側の関数に向かって，逐次的に勾配が計算される。

$$
\begin{aligned}
\left.\frac{\partial L}{\partial \boldsymbol{z}^{(K)}}\right|_{\boldsymbol{z}^{(K)}=\hat{\boldsymbol{z}}^{(K)}} &= 1, \\
\left.\frac{\partial L}{\partial \boldsymbol{z}^{(K-1)}}\right|_{\boldsymbol{z}^{(K-1)}=\hat{\boldsymbol{z}}^{(K-1)}} &= \left.\frac{\partial L}{\partial \boldsymbol{z}^{(K)}}\right|_{\boldsymbol{z}^{(K)}=\hat{\boldsymbol{z}}^{(K)}} \left.\frac{\partial \boldsymbol{z}^{(K)}}{\partial \boldsymbol{z}^{(K-1)}}\right|_{\boldsymbol{z}^{(K-1)}=\hat{\boldsymbol{z}}^{(K-1)}}, \\
\left.\frac{\partial L}{\partial \boldsymbol{z}^{(K-2)}}\right|_{\boldsymbol{z}^{(K-2)}=\hat{\boldsymbol{z}}^{(K-2)}} &= \left.\frac{\partial L}{\partial \boldsymbol{z}^{(K-1)}}\right|_{\boldsymbol{z}^{(K-1)}=\hat{\boldsymbol{z}}^{(K-1)}} \left.\frac{\partial \boldsymbol{z}^{(K-1)}}{\partial \boldsymbol{z}^{(K-2)}}\right|_{\boldsymbol{z}^{(K-2)}=\hat{\boldsymbol{z}}^{(K-2)}}, \ldots, \\
\left.\frac{\partial L}{\partial \boldsymbol{z}^{(0)}}\right|_{\boldsymbol{z}^{(0)}=\hat{\boldsymbol{x}}} &= \left.\frac{\partial L}{\partial \boldsymbol{z}^{(1)}}\right|_{\boldsymbol{z}^{(1)}=\hat{\boldsymbol{z}}^{(1)}} \left.\frac{\partial \boldsymbol{z}^{(1)}}{\partial \boldsymbol{z}^{(0)}}\right|_{\boldsymbol{z}^{(0)}=\hat{\boldsymbol{x}}}
\end{aligned}
$$

このように，誤差逆伝播法は，一度計算した勾配の値を再利用することによって高い計算効率を実現している。勾配の値は出力から入力に向かって計算されるため，この勾配計算を逆伝播（backward propagation; backpropagation）と呼び，対比的に，入力から出力への計算を順伝播（forward propagation）と呼ぶ。

誤差逆伝播法は，任意の微分可能な（differentiable）関数の合成関数に適用可能である。例えば，図 3.5 に示す 3 層 MLP には分岐する入力が含まれるが，順伝播と逆伝播は図 3.7 のように行うことができる。また，図 3.8 に示すように出力が分岐する場合にも，誤差逆伝播法は適用可能である。微分可能でありさえすれば誤差逆伝播法は適用可能であるため，近年では，微分可能な時間周波数計算機構 [149,167,168] やデータ拡張 [169,170] 等，特定の問題に特化した様々な演算機構を DNN に組み込む研究も行われている。

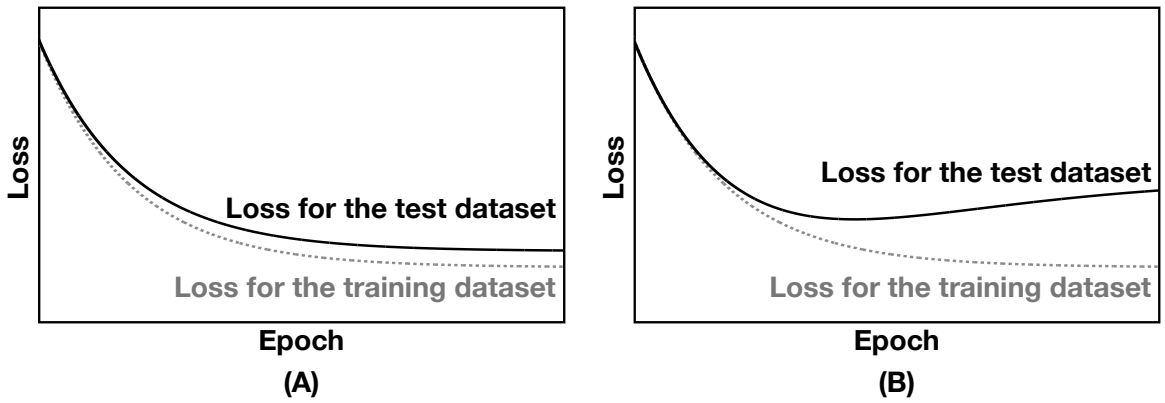


図 3.9 過剰適合が生じている学習曲線の概念図（岡谷 [157] を元に作成）

3.5　学習上の困難さとその緩和

ここまで，深層学習の基礎について説明を行ったが，DNN を学習させる上で生じる困難さについては触れてこなかった。本節では，DNN の学習上の困難さと，その困難さを緩和する手法を紹介する。

3.5.1　汎化性能と過剰適合

これまで説明したように，DNN は訓練データセットに対する損失が小さくなるように訓練される。しかし，いくら訓練データセットに含まれるデータに対しては高い性能を実現できたとしても，訓練データセットに含まれない未知のデータに対する性能，すなわち汎化性能（generalization performance）が低いならば，その DNN は実用的ではない。そこで，訓練データセットと中身が重複しないデータから構成されるテストデータセット（test dataset）を準備し，テストデータセットに対する性能を DNN の評価に用いる [157]。

図 3.9 は，学習の経過に伴う損失の変化を示す学習曲線（learning curve）の概念図である。図 3.9 に示すように，典型的には，学習の経過と共に，訓練データセットに対する損失と，テストデータセットに対する損失は乖離していく。このように損失の乖離が生じている状態を過剰適合（overfitting）と呼ぶ。図 3.9 (A) のように，テストデータセットに対する損失が上昇しない場合もあれば，図 3.9 (B) のように，訓練データセットに対する損失は減少するにも関わらず，テストデータセットに対する損失が上昇する場合もある。DNN の汎化性能を高める上では，過剰適合を抑制することが重要であるため，

図 3.9 (B) のようにテストデータセットに対する損失が上昇する前に，DNN の学習を早期終了（early stopping）することが望ましい。そこで，訓練データセットともテストデータセットとも中身が重複しない検証データセット（validation dataset）を準備し，検証データセットに対する損失を早期終了の判断基準に用いる [157]。

DNN は，膨大な次元のパラメータを持つため，容易に重度の過剰適合が生じる。そのため，高い汎化性能を実現するためには，過剰適合を抑制するための工夫が必要となる。

3.5.2 勾配消失と勾配爆発

DNN は，その名の通り深い層構造を有する数理モデルである。しかし，DNN が深くなれば深くなるほど，入力に近い層において，勾配の値がほとんど 0 になってしまう勾配消失（vanishing gradient）や，その逆に勾配の値が発散してしまう勾配爆発（exploding gradient）が生じるリスクが高くなる [171, 172]。

例えば，$K \in \{2, 3, \dots\}$ 層の MLP が，$f(\boldsymbol{x}) = f^{(K)}(f^{(K-1)}(\cdots f^{(2)}(f^{(1)}(\boldsymbol{x}))\cdots))$，$f^{(k)}(\boldsymbol{x}; \boldsymbol{W}^{(k)}) = g(\boldsymbol{W}^{(k)}\boldsymbol{x})$ のように与えられたとする（簡単のため，バイアス項は省略している）。ここで，g は要素毎に作用する活性化関数であり，$\boldsymbol{W}^{(k)}$ の次元は式が成立するように適切に与えられている。また，隠れ表現を $\boldsymbol{u}^{(k)}(\boldsymbol{x}) = \boldsymbol{W}^{(k)}\left(f^{(k-1)}(f^{(k-2)}(\cdots f^{(2)}(f^{(1)}(\boldsymbol{x}))\cdots))\right)$，$\boldsymbol{z}^{(k)}(\boldsymbol{x}) = g(\boldsymbol{u}^{(k)}(\boldsymbol{x}))$ と表し，損失関数が L で与えられたとする。このとき，入力層の重みに対する損失関数の勾配 $\partial L/\partial \boldsymbol{W}^{(1)}$ は，連鎖律を用いて次のように計算される。

$$
\begin{aligned}
\frac{\partial L}{\partial \boldsymbol{W}^{(1)}} &= \frac{\partial L}{\partial \boldsymbol{z}^{(K)}}\frac{\partial \boldsymbol{z}^{(K)}}{\partial \boldsymbol{u}^{(K)}}\left(\prod_{k=2}^{K}\frac{\partial \boldsymbol{u}^{(k)}}{\partial \boldsymbol{z}^{(k-1)}}\frac{\partial \boldsymbol{z}^{(k-1)}}{\partial \boldsymbol{u}^{(k-1)}}\right)\frac{\partial \boldsymbol{u}^{(1)}}{\partial \boldsymbol{W}^{(1)}} \\
&= \frac{\partial L}{\partial \boldsymbol{z}^{(K)}}\frac{\partial \boldsymbol{z}^{(K)}}{\partial \boldsymbol{u}^{(K)}}\left(\prod_{k=2}^{K}\boldsymbol{W}^{(k)}g'(\boldsymbol{u}^{(k-1)})\right)\frac{\partial \boldsymbol{u}^{(1)}}{\partial \boldsymbol{W}^{(1)}}
\end{aligned}
\tag{3.31}
$$

但し，g' は g の導関数を表し，式変形において，行列 $\boldsymbol{A}$ とベクトル $\boldsymbol{u}$ の積 $\boldsymbol{A}\boldsymbol{u}$ に関して $\partial(\boldsymbol{A}\boldsymbol{u})/\partial\boldsymbol{u} = \boldsymbol{A}$ となることを利用した。

式 (3.31) からわかるように，層が深いほど，すなわち K が大きいほど，$\prod_{k=2}^{K}\boldsymbol{W}^{(k)}g'(\boldsymbol{u}^{(k-1)})$ が $\partial L/\partial \boldsymbol{W}^{(1)}$ の値に大きな影響を与える。活性化関数 g が恒等関数であり，$k \to \infty$ のとき，任意の重み $W^{(k)}$ のスペクトル半径（固有値の絶対値の最大値）が 1 未満ならば $\partial L/\partial \boldsymbol{W}^{(1)} \to 0$（勾配消失），1 を超えるならば $\partial L/\partial \boldsymbol{W}^{(1)} \to \infty$（勾配爆発）となる [172]。また，活性化関数が sigmoid 関数や tanh 関数である場合，その微分値は 1 以下であり，加えて原点付近以外ではさらに微分値が小さいため，特に勾配消失が起こりやすい。

DNN の深化は表現力向上に繋がるが，勾配消失や勾配爆発が生じてしまったら，特に

入力に近い層のパラメータを適切に更新できない。そのため，DNN を深化させる場合には，勾配消失や勾配爆発の影響を抑えることが重要になる。

3.5.3 正則化

過剰適合の抑制を目的としてパラメータに制約を加える手法のことを正則化（regularization）と呼ぶ。l^2 正則化（l^2 regularization）は正則化手法の一つであり，次式のように損失にパラメータの l^2 ノルムの二乗を加える。

$$L^{(t)}_{\text{mini-batch}}(\theta) = g\left(\{f(\boldsymbol{x}^{(i)};\theta)\}_{i\in\mathcal{I}^{(t)}}, \{\boldsymbol{t}^{(i)}\}_{i\in\mathcal{I}^{(t)}}\right) + \lambda \left\|\theta\right\|^2_{l^2} \tag{3.32}$$

$\lambda > 0$ は正則化の強さを表すハイパーパラメータであり，正則化定数（regularization constant）と呼ばれ，その他の記号は式 (3.8) と同様のものを表す。l^2 正則化は，パラメータの l^2 ノルムが過剰に大きくなるのを抑制するが，バイアスは大きな値を取る必要がある場合もあるため，一般に l^2 正則化は重みに対してのみ適用される。l^2 正則化は，重みの値が大きいほど効果が強くなることから，荷重減衰（weight decay）とも呼ばれる。

Dropout [173] は正則化手法の一種であり，学習時に隠れ表現の一部をランダムに無効化することによって，過剰適合の抑制を図る。Dropout を適用する場合，学習時は，隠れ表現ベクトルの要素の値を，ハイパーパラメータである一定の割合 p で，ミニバッチごとにランダムに 0 に置換する。その一方で，テスト時は，0 への置換を行わない代わりに，隠れ表現ベクトルの値を p 倍することによって値の大きさを補償する[*8]。ランダムな無効化によって元の DNN から小さな別の DNN を作り出すことに相当し，Dropout は，学習時はランダムな無効化によって元の DNN から多数の小さな DNN を作り出すと共に訓練し，テスト時は多数の小さな DNN の平均化を行うことによって，擬似的にアンサンブル学習を実現して過剰適合を抑制していると考えられている [173]。

3.5.4 活性化関数

現代の深層学習では，代表的な活性化関数の一つとして，次式で定義される Rectified Linear Unit (ReLU) [174, 175] が広く使用されている。

$$\mathrm{ReLU}(x) = \begin{cases} x & (x \geq 0) \\ 0 & (x < 0) \end{cases} \tag{3.33}$$

また，ReLU の微分は，次式のようになる。

$$\frac{\mathrm{d}\,\mathrm{ReLU}}{\mathrm{d}x}(x) = \begin{cases} 1 & (x > 0) \\ 0 & (x < 0) \end{cases} \tag{3.34}$$

[*8] 学習時に隠れ表現ベクトルの値を $1/p$ 倍にする実装もある。

ReLU は数学的には $x = 0$ で微分不可能であるが，実装上は，$x = 0$ における微分値は 0 か 1 に割り当てる。先述の通り，伝統的な活性化関数である sigmoid 関数や tanh 関数は微分値が小さいため，勾配消失を引き起こす原因の一つであった。その一方で，ReLU は正の領域で微分値が 1 であるために勾配消失を引き起こしにくい上に，sigmoid 関数や tanh 関数よりも効果的な活性化関数であることが実証されている [175]。

ReLU は現代のデファクトスタンダードであるが，ReLU の派生となる活性化関数も多数提案されており [176–179]，次式で定義される Swish [176] は ReLU の派生の一つである。

$$\mathrm{Swish}(x) = x \cdot \mathrm{sigmoid}(x) \tag{3.35}$$

Swish は，自動探索技術を用いて特定された関数であり，DNN 中の ReLU を Swish で置換することによって，DNN の性能が向上することが実証されている [176]。

3.5.5 Batch Normalization

Batch Normalization [180] は，隠れ表現ベクトルの平均と標準偏差を正規化する手法であり，典型的には，全結合層と活性化関数の間で適用される[*9]。隠れ表現ベクトルのミニバッチ $\{\boldsymbol{z}^{(b)} \in \mathbb{R}^N\}_{b=1}^B$ $(N, B \in \mathbb{N})$ が与えられたとき，Batch Normalization は，$\{\boldsymbol{z}^{(b)}\}_{b=1}^B$ に対する平均 $\boldsymbol{\mu} \in \mathbb{R}^N$ と標準偏差 $\boldsymbol{\sigma} \in \mathbb{R}^N$ を用いて，正規化されたミニバッチ $\{\tilde{\boldsymbol{z}}^{(b)} \in \mathbb{R}^N\}_{b=1}^B$ を次のように計算する。

$$\tilde{z}_i^{(b)} = \frac{z_i^{(b)} - \mu_i}{\sqrt{\sigma_i^2 + \epsilon}} \cdot \gamma_i + \beta_i \tag{3.36}$$

$$\mu_i = \frac{1}{B} \sum_{b'=1}^{B} z_i^{(b')} \tag{3.37}$$

$$\sigma_i = \sqrt{\frac{1}{B} \sum_{b'=1}^{B} (z_i^{(b')} - \mu_i)^2} \tag{3.38}$$

ここで，$\epsilon > 0$ はゼロ除算を防ぐ微少量，$\gamma \in \mathbb{R}_{>0}^N$ と $\beta \in \mathbb{R}^N$ は $\gamma = 1$，$\beta = 0$ で初期化される学習可能なパラメータであり，それぞれスケーリングと平行移動の役割を担う。なお，Batch Normalization を適用する場合，直前の全結合層のバイアスは計算過程で無効化されるため，一般的に全結合層のバイアスは省略される。

テスト時は，ミニバッチの平均と標準偏差の代わりに，訓練データセット全体に対する平均と標準偏差の推定値を用いる。通常，平均と標準偏差の推定値は加重移動平均を

[*9] CNN の場合は，畳み込み層と活性化関数の間で適用される他，Batch Normalization の定義も少し異なる。詳細は第 3.7.5 節を参照。

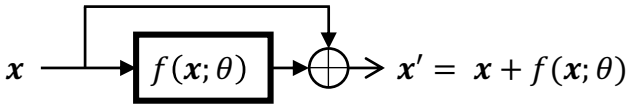


図 3.10 残差ブロックの概念図

使って算出される。t 回目のパラメータ更新後の平均と標準偏差の推定値を $\hat{\boldsymbol{\mu}}^{(t)}, \hat{\boldsymbol{\sigma}}^{(t)}$，$t$ 番目のミニバッチの平均と標準偏差を $\boldsymbol{\mu}^{(t)}, \boldsymbol{\sigma}^{(t)}$ と記す場合，推定値の更新式は次のようになる。

$$\hat{\boldsymbol{\mu}}^{(t)} = \alpha\hat{\boldsymbol{\mu}}^{(t-1)} + (1-\alpha)\boldsymbol{\mu}^{(t)} \tag{3.39}$$

$$\hat{\boldsymbol{\sigma}}^{(t)} = \alpha\hat{\boldsymbol{\sigma}}^{(t-1)} + (1-\alpha)\boldsymbol{\sigma}^{(t)} \tag{3.40}$$

ここで，α は重み付けを調節するハイパーパラメータである。$\alpha = 0.9$ や $\alpha = 0.99$ などに設定されることが多く，本論文の実験では $\alpha = 0.9$ と設定した。

Batch Normalization は，勾配消失や勾配爆発を抑制するだけでなく，学習率などのハイパーパラメータに対する頑健性を高めたり，損失低下の速度を加速したりする等，様々な点で DNN の学習を改善する効果が確かめられている [180]。その一方で，Batch Normalization の有効性が生じる理由に関しては様々な説があるが [181–183]，未だ完全な理解には至っておらず，現在も議論の的である。

3.5.6　スキップ接続

残差ネットワーク（residual network） [184, 185] は，途中の演算を迂回するような構造を取り入れたネットワークである。残差ネットワークは，図 3.10 に示す残差ブロック（residual block） [186] と呼ばれる構造を積み重ねることによって構成され，残差ブロック内の迂回路のことをスキップ接続（skip connection） [185] と呼ぶ[*10]。残差ブロック内の関数 $f(\boldsymbol{x};\theta)$ は，全結合層や Batch Normalization，活性化関数等の組み合わせで構成される。「残差（residual）」という用語は，関数 $f(\boldsymbol{x};\theta)$ で計算するものが，残差ブロックの出力 $\boldsymbol{x}'$ から入力 $\boldsymbol{x}$ を引いた残り（residue）$\boldsymbol{x}' - \boldsymbol{x}$ であることに由来している。

スキップ接続は，出力から入力に向かって確実に勾配を伝えることによって勾配消失を解消し，スキップ接続がない場合と比べて，さらに深いネットワークの学習を可能にした [184]。また，残差ネットワーク内の一部の残差ブロックを削除したり，残差ブロックの順番を入れ替えたりした場合でも，性能は大きく低下しないことが知られており [187]，学習時にランダムで残差ブロックを削除する確率的深度変動法（stochastic depth）とい

[*10] 残差ブロックを残差ユニット（residual unit），スキップ接続をショートカット（shortcut）とも呼ぶ [185]。

う正則化手法が提案されている [187]。

3.5.7　勾配ノルムクリッピング

勾配ノルムクリッピング（gradient norm clipping） [172] は，勾配算出後のパラメータ更新直前で，勾配のノルムが閾値以下になるように勾配を修正する手法である。勾配ノルムクリッピングでは，閾値 $\theta > 0$ がハイパーパラメータとなる。典型的には，DNN の全パラメータの勾配を平坦化されたベクトル $\boldsymbol{g} \in \mathbb{R}^N$ $(N \in \mathbb{N})$ と見做し，$\|\boldsymbol{g}\|_{l^2} \leq \theta$ ならば特に処理を行わず，$\|\boldsymbol{g}\|_{l^2} > \theta$ ならば勾配を $(\theta / \|\boldsymbol{g}\|_{l^2})\boldsymbol{g}$ に修正する。勾配ノルムクリッピングは，勾配爆発を直接的に抑制する効果がある [172]。

3.6　数理モデル②：Recurrent Neural Network

RNN は，再帰構造を含むニューラルネットワークの総称であり，系列データを扱うことに特化している。MLP は時間構造を陽に考慮しない構造であるが，RNN は系列構造を陽に考慮する構造である。また，MLP では原則固定長入力しか扱えないが[*11]，RNN では自然に可変長の系列データを扱うことができる。本節では，RNN の中でも単純な構造を導入した後，単純な RNN の学習で生じる問題と，その緩和策を取り入れた RNN を説明する。

3.6.1　Simple Recurrent Network

Elman ネットワーク [188] は，Simple Recurrent Network（SRN）の一種として知られる構造である。系列長が $T \in \mathbb{N}$ である系列データ $\{\boldsymbol{x}^{(t)} \in \mathbb{R}^L\}_{t=1}^T$ $(L \in \mathbb{N})$ が入力として与えられた時，Elman ネットワークは以下の計算を行う。

$$\boldsymbol{z}^{(t)} = g_z\left(\boldsymbol{W}_z\boldsymbol{x}^{(t)} + \boldsymbol{U}_z\boldsymbol{z}^{(t-1)} + \boldsymbol{b}_z\right) \tag{3.41}$$

$$\boldsymbol{y}^{(t)} = g_y\left(\boldsymbol{W}_y\boldsymbol{z}^{(t)} + \boldsymbol{b}_y\right) \tag{3.42}$$

ここで，$\boldsymbol{W}_z \in \mathbb{R}^{M\times L}, \boldsymbol{U}_z \in \mathbb{R}^{M\times M}, \boldsymbol{W}_y \in \mathbb{R}^{N\times M}$ は重み，$\boldsymbol{b}_z \in \mathbb{R}^M, \boldsymbol{b}_y \in \mathbb{R}^N$ はバイアス，$\{\boldsymbol{z}^{(t)} \in \mathbb{R}^M\}_{t=0}^T$ $(M \in \mathbb{N})$ は隠れ表現の系列，$\{\boldsymbol{y}^{(t)} \in \mathbb{R}^N\}_{t=1}^T$ $(N \in \mathbb{N})$ は出力系列，g_z と g_y は活性化関数である[*12]。Elman ネットワークは，図 3.11 (A) のように再帰構造を組み込むことによって，陽に時間方向の依存関係を考慮する。その一方で，図 3.11 (A) を時間方向に展開することによって，図 3.11 (B) のような多層構造のネットワークのよ

[*11] パディング等で対処する方策はある。

[*12] Elman による原典 [188] では，$\boldsymbol{z}^{(0)}$ のすべての要素の値は 0.5，g_z の値域は 0 以上 1 以下とされる。

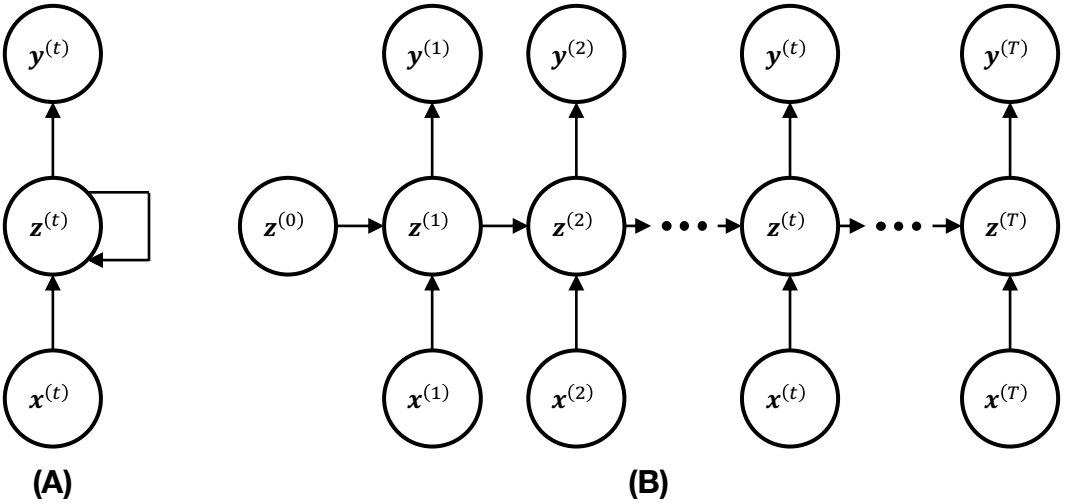


図 3.11 Elman ネットワークの計算グラフ。Elman ネットワークは再帰構造を持つが (A)，時間方向に展開することによって多層構造のネットワークのようにも見做せる (B)。

うに表現でき，誤差逆伝播法による勾配計算が可能である [189]。

通常，Elman ネットワークを含む RNN は過去から未来に向かって再帰処理を行うが，未来から過去に向かう再帰処理を組み合わせる手法がある。このように，2 つの向きで再帰処理を行う RNN を双方向 RNN（bidirectional RNN）と呼ぶ [190]。双方向 RNN は未来の情報を利用するためにリアルタイム処理には使えない制約があるが，単一方向の RNN よりも性能が高くなることがある [191]。

一般に，RNN は系列データを入力とし，系列データを出力する構造であるが，応用においては，系列データの入力から固定長ベクトルを出力したい場面がある。例えば，本研究で取り組む音声データからの GRBAS 尺度の評点推定や，テキストデータからの感情推定等，系列データを入力とする回帰問題や分類問題では，出力を固定長ベクトルに変換する必要が生じる。このような場合，出力系列の全時間に平均を計算したり，SRN の出力系列の最後の時点のベクトルを用いたりすることによって，可変長の系列データを固定長の隠れ表現ベクトルに変換して対応する。

3.6.2 勾配消失問題への対処

RNN は時間方向に展開することによって多層構造のネットワークと見做せるため，第 3.5.2 節の説明と同様の理由で，SRN を用いた学習では勾配爆発，とりわけ勾配消失が生じやすい [171, 172]。時間長が長くなるほどそのリスクは大きくなり，SRN が長期的依存関係を学習するのは困難である。

Long Short-Term Memory (LSTM) [192, 193] は，上記の問題点を解消するために考案された RNN である。系列長が $T \in \mathbb{N}$ である系列データ $\{\boldsymbol{x}^{(t)} \in \mathbb{R}^N\}_{t=1}^T$ $(N \in \mathbb{N})$ が入力として与えられたとき，LSTM は以下の計算を行う。

$$\boldsymbol{i}^{(t)} = \text{sigmoid}\left(\boldsymbol{W}_i \boldsymbol{x}^{(t)} + \boldsymbol{U}_i \boldsymbol{z}^{(t-1)} + \boldsymbol{b}_i\right) \tag{3.43}$$

$$\boldsymbol{f}^{(t)} = \text{sigmoid}\left(\boldsymbol{W}_f \boldsymbol{x}^{(t)} + \boldsymbol{U}_f \boldsymbol{z}^{(t-1)} + \boldsymbol{b}_f\right) \tag{3.44}$$

$$\boldsymbol{o}^{(t)} = \text{sigmoid}\left(\boldsymbol{W}_o \boldsymbol{x}^{(t)} + \boldsymbol{U}_o \boldsymbol{z}^{(t-1)} + \boldsymbol{b}_o\right) \tag{3.45}$$

$$\hat{\boldsymbol{c}}^{(t)} = \tanh\left(\boldsymbol{W}_c \boldsymbol{x}^{(t)} + \boldsymbol{U}_c \boldsymbol{z}^{(t-1)} + \boldsymbol{b}_c\right) \tag{3.46}$$

$$\boldsymbol{c}^{(t)} = \boldsymbol{i}^{(t)} \odot \hat{\boldsymbol{c}}^{(t)} + \boldsymbol{f}^{(t)} \odot \boldsymbol{c}^{(t-1)} \tag{3.47}$$

$$\boldsymbol{z}^{(t)} = \boldsymbol{o}^{(t)} \odot \tanh\left(\boldsymbol{c}^{(t)}\right) \tag{3.48}$$

ここで，$\boldsymbol{W}_f, \boldsymbol{W}_i, \boldsymbol{W}_o, \boldsymbol{W}_c \in \mathbb{R}^{M \times N}, \boldsymbol{U}_f, \boldsymbol{U}_i, \boldsymbol{U}_o, \boldsymbol{U}_c \in \mathbb{R}^{M \times M}$ は重み，$\boldsymbol{b}_f, \boldsymbol{b}_i, \boldsymbol{b}_o, \boldsymbol{b}_c \in \mathbb{R}^M$ はバイアスであり，$\{\boldsymbol{c}^{(t)} \in \mathbb{R}^M\}_{t=0}^T$ はメモリセル（memory cell），$\{\boldsymbol{z}^{(t)} \in \mathbb{R}^M\}_{t=0}^T$ は隠れ状態（hidden state）と呼ばれる $(M \in \mathbb{N})$。LSTM のメモリセルと隠れ状態は，それぞれ Elman ネットワークの隠れ表現と出力系列に対応する役割を担い，通常，$\boldsymbol{c}^{(0)} = \boldsymbol{z}^{(0)} = \boldsymbol{0}$ で初期化される。メモリセルの更新量は，式 (3.43) の入力ゲート（input gate）と式 (3.44) の忘却ゲート（forget gate）によって，式 (3.46) と (3.47) のように制御される。sigmoid 関数の値域は $(0, 1)$ であり，sigmoid 関数の出力が 0 に近いほどゲートが閉じ，出力が 1 に近いほどゲートが開いている状態と考えることができる。入力ゲートが閉じ，かつ忘却ゲートが開いている限り，メモリセルの値は変化しない。このようなメモリセルの機構によって，LSTM では長期的な依存関係の学習と勾配消失の抑制の能力が SRN よりも改善している。なお，式 (3.45) の出力ゲート（output gate）は，出力系列である隠れ表現に，メモリセルの情報をどれだけ反映するかを式 (3.48) のように制御する。

Gated Recurrent Unit (GRU) [194] は，LSTM の計算を一部簡略化した RNN である。系列長が $T \in \mathbb{N}$ である系列データ $\{\boldsymbol{x}^{(t)} \in \mathbb{R}^N\}_{t=1}^T$ $(N \in \mathbb{N})$ が入力として与えられたとき，GRU は以下の計算を行う。

$$\boldsymbol{u}^{(t)} = \text{sigmoid}\left(\boldsymbol{W}_u \boldsymbol{x}^{(t)} + \boldsymbol{U}_u \boldsymbol{z}^{(t-1)} + \boldsymbol{b}_u\right) \tag{3.49}$$

$$\boldsymbol{r}^{(t)} = \text{sigmoid}\left(\boldsymbol{W}_r \boldsymbol{x}^{(t)} + \boldsymbol{U}_r \boldsymbol{z}^{(t-1)} + \boldsymbol{b}_r\right) \tag{3.50}$$

$$\hat{\boldsymbol{z}}^{(t)} = \tanh\left(\boldsymbol{W}_z \boldsymbol{x}^{(t)} + \boldsymbol{r}^{(t)} \odot (\boldsymbol{U}_z \boldsymbol{z}^{(t-1)}) + \boldsymbol{b}_z\right) \tag{3.51}$$

$$\boldsymbol{z}^{(t)} = \boldsymbol{u}^{(t)} \odot \hat{\boldsymbol{z}}^{(t)} + (1 - \boldsymbol{u}^{(t)}) \odot \boldsymbol{z}^{(t-1)} \tag{3.52}$$

ここで，$\boldsymbol{W}_u, \boldsymbol{W}_r, \boldsymbol{W}_z \in \mathbb{R}^{M \times N}, \boldsymbol{U}_u, \boldsymbol{U}_r, \boldsymbol{U}_z \in \mathbb{R}^{M \times M}$ は重み，$\boldsymbol{b}_u, \boldsymbol{b}_r, \boldsymbol{b}_z \in \mathbb{R}^M$ はバイアス，$\{\boldsymbol{z}^{(t)} \in \mathbb{R}^M\}_{t=0}^T$ は隠れ状態である $(M \in \mathbb{N})$。LSTM では 3 つのゲートが存在す

るのに対し，GRU に存在するのは式 (3.49) の更新ゲート（update gate）と式 (3.50) の初期化ゲート（reset gate）と呼ばれる 2 つのゲートである。また，GRU の隠れ状態は，LSTM のメモリセルと隠れ状態の両方の役割を兼ねている。課題による差異はあるものの，LSTM と GRU の両者の性能は顕著には変わらないことが比較実験で示唆されている [195, 196]。

3.7　数理モデル③：Convolutional Neural Network

音声波形等の系列データや画像等の行列データでは，成分の位置（添字）が重要な意味を持つ。しかし，全結合層は，図 3.12 (A) のように，位置に関係なくすべての成分を同等に扱う仕組みであり，成分の位置関係を陽には考慮しない。それに対し，本節で説明する CNN の主要な構成要素である畳み込み層（convolution layer）は，図 3.12 (B) のように畳み込み演算を用いることによって，相対的な位置関係に基づいて現れるパターンの抽出に特化している。また，畳み込み層では，相対的な位置関係に基づいて結合重みの値が共有されるため，相対的な位置関係が重要なデータに関しては，MLP よりも CNN の方が学習を効率的に行うことができると考えられる。実際，CNN は音声合成 [67] や画像分類 [184] 等の課題で高い性能を発揮していることに加え，（2 次元）CNN はランダムなノイズ画像よりも自然画像に現れる特徴を捉えやすいことが示唆されている [197]。

本節では，CNN を構成する要素である畳み込み層とプーリング層（pooling layer）に加え，CNN 用に定義された Batch Normalization を説明する。なお，本節では，数式を簡潔に表現するため，テンソル[*13]等の添字は 0 から始める。

3.7.1　1 次元畳み込み層

畳み込み層は，フィルタバンクを用いた畳み込み変換を行う層である。畳み込み層には様々なバリエーションがあり，設定可能なハイパーパラメータも多く存在するが，まず初めに，畳み込み層の中でも 1 次元畳み込み層について，特に最も素朴な定義を説明する。$\boldsymbol{X} \in \mathbb{R}^{C_{\mathrm{input}} \times L_{\mathrm{input}}}$ が入力として与えられたとき，重み $\boldsymbol{W} \in \mathbb{R}^{C_{\mathrm{output}} \times C_{\mathrm{input}} \times L_{\mathrm{kernel}}}$ ($L_{\mathrm{kernel}} \leq L_{\mathrm{input}}$) とバイアス $\boldsymbol{b} \in \mathbb{R}^{C_{\mathrm{output}}}$ を持つ（最も素朴な）1 次元畳み込み層は以下のように $\boldsymbol{Y} \in \mathbb{R}^{C_{\mathrm{output}} \times L_{\mathrm{output}}}$ を出力する。

$$Y_{c,l} = \sum_{c'=0}^{C_{\mathrm{input}}-1} \left(\sum_{l'=0}^{L_{\mathrm{kernel}}-1} X_{c',l+l'} W_{c,c',l'} \right) + b_c \tag{3.53}$$

[*13] 1 階のテンソルはベクトル，2 階のテンソルは行列に相当する。

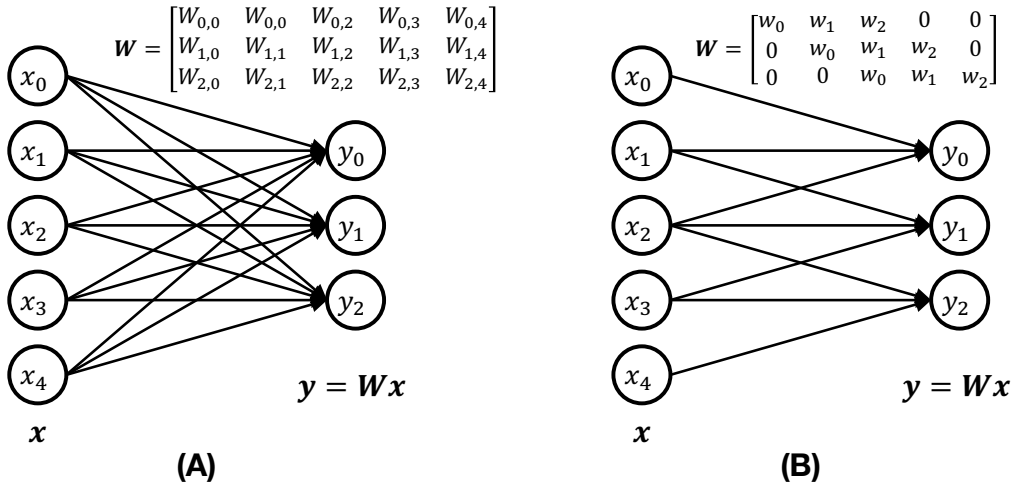


図 3.12 (A) は全結合層，(B) は簡略版の畳み込み層を表す。全結合層では位置に依存しない処理が行われず，各矢印が表す重みの値は独立している。それに対し，畳み込み層では位置に依存した処理が行われ，相対的な配置に基づいて重みの値が共有される。

$\boldsymbol{X}$ は，ベクトル列 $\{\boldsymbol{x}^{(c)} \in \mathbb{R}^{L_{\text{input}}} : x_i^{(c)} = \boldsymbol{X}_{c,i}\}_{c=0}^{C_{\text{input}}-1}$ を重ねたものであると見做され，各 $\boldsymbol{x}^{(c)}$ をチャンネル（channel）と呼ぶ。これは，$\boldsymbol{Y}$ についても同様であり，$\boldsymbol{X}$ のチャンネルを特に入力チャンネル（input channel），$\boldsymbol{Y}$ のチャンネルを特に出力チャンネル（output channel）と呼ぶ。$\boldsymbol{X}$ と $\boldsymbol{Y}$ の最初の次元はチャンネル数であり[*14]，C_{input} が入力チャンネル数，C_{output} が出力チャンネル数に対応する。畳み込み層は，図 3.13 のように，チャンネル数を 1 から C_{output} に変換する C_{input} 個のフィルタバンクのように見做せ，C_{input} 個のフィルタバンクの出力（およびバイアス）の和が，畳み込み層としての出力になっていると考えることができる。あるいは，図 3.14 のように，入力の各チャンネルに対して作用する C_{input} チャンネルのフィルタバンクが C_{output} 個あり，各フィルタバンクの出力（およびバイアス）の和が，出力のチャンネルの 1 つに対応していると考えることもできる。フィルタバンクを表す重み $\boldsymbol{W}$ は，単にフィルタ，あるいはカーネル（kernel）とも呼ばれ，L_{kernel} をフィルタサイズやカーネルサイズ等と呼ぶ。

何も処理を行わない場合，出力チャンネルのサイズは $L_{\text{output}} = L_{\text{input}} - (L_{\text{kernel}} - 1)$ であり，入力チャンネルのサイズより $L_{\text{kernel}} - 1$ だけ小さくなる。このようにサイズが小さくなるのを防ぐため，多くの場合，畳み込み層では式 (3.53) の処理をする前に入力チャンネルのサイズを $L_{\text{kernel}} - 1$ だけ大きくする。このような操作をパディング（padding）と呼ぶ。典型的には，ゼロパディング（zero-padding）と呼ばれる，入力チャンネルの最初の成分の前と，最後の成分の後に値が 0 の成分を追加する処理を行う。

パディングとは逆に，ストライド（stride）は出力チャンネルのサイズを小さくするこ

[*14] 最後の次元でチャンネル数を表す流儀もある。

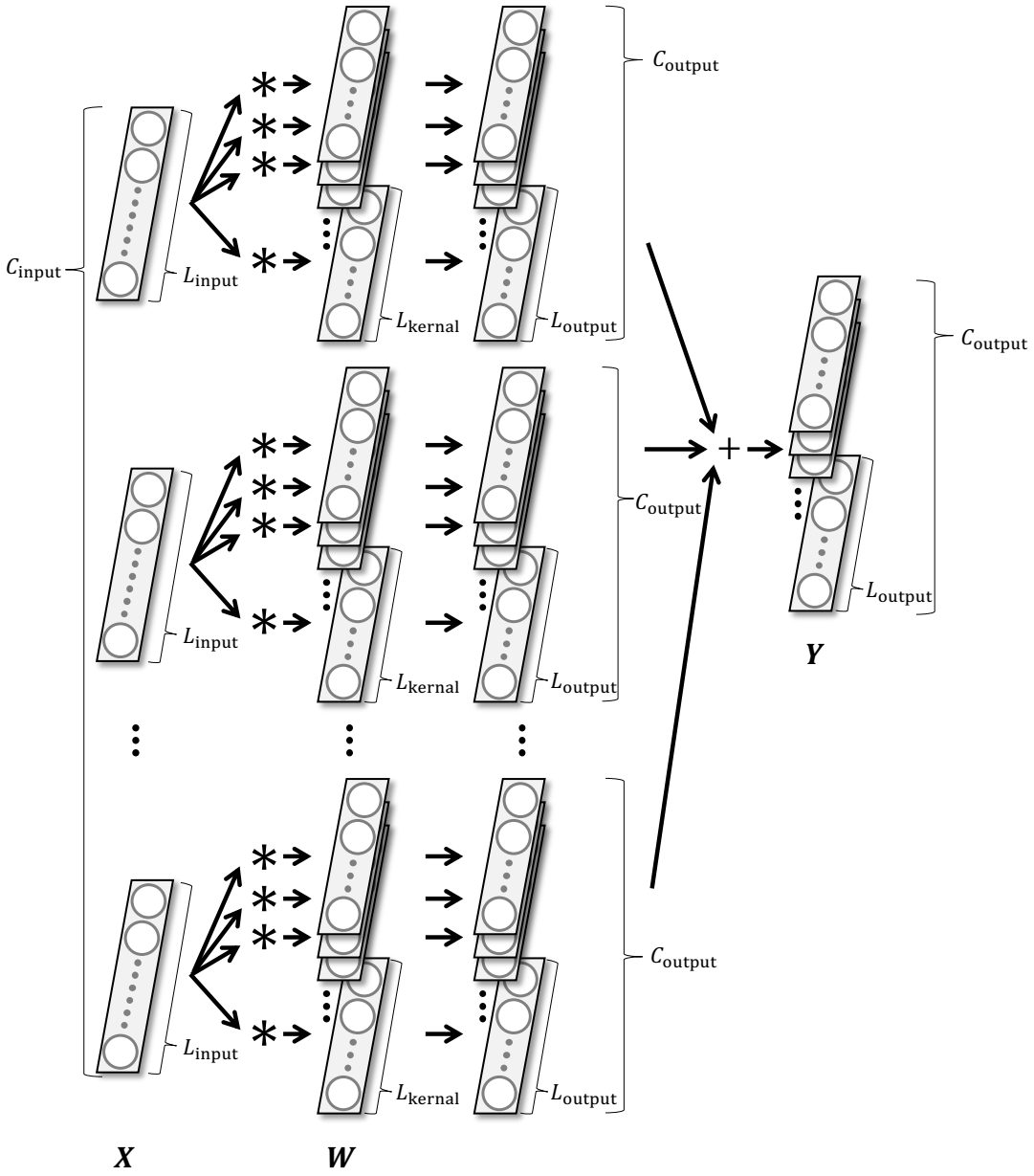


図 3.13 畳み込み層の重みを，チャンネル数を 1 から C_{output} に変換するフィルタバンクが C_{input} 個集まったものと解釈した場合の，1 次元畳み込み層で行われる演算の概念図。但し，バイアスは省略されている。

とに関わるハイパーパラメータである。ストライドが $s \in \mathbb{N}$ の 1 次元畳み込みは，次式のように行われる。

$$Y_{c,l} = \sum_{c'=0}^{C_{\text{input}}-1} \left(\sum_{l'=0}^{L_{\text{kernel}}-1} X_{c',sl+l'} W_{c,c',l'} \right) + b_c \tag{3.54}$$

パディングが行われない場合，出力チャンネルのサイズは $L_{\text{output}} = 1 + \lfloor (L_{\text{input}} - L_{\text{kernel}})/s \rfloor$（$\lfloor \cdot \rfloor$ は床関数）となり，$L_{\text{input}} \gg L_{\text{kernel}}$ ならば $L_{\text{output}} \approx L_{\text{input}}/s$

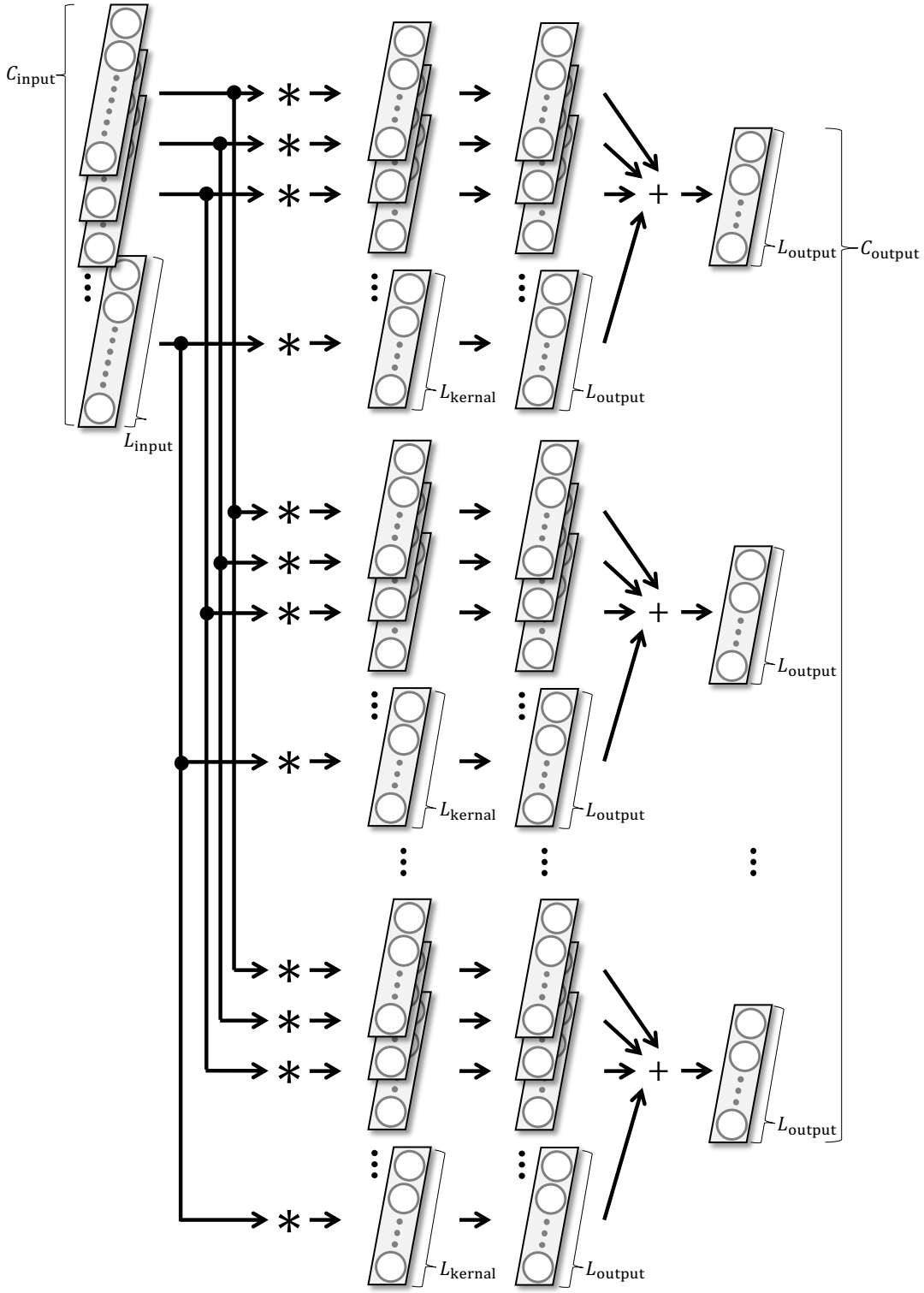


図 3.14 畳み込み層の重みを，入力の各チャンネルに対して作用する C_{input} チャンネルのフィルタバンクが C_{output} 個集まったものと解釈した場合の，1 次元畳み込み層で行われる演算の概念図。但し，バイアスは省略されている。

となる。$s = 1$ のとき，式 (3.54) は式 (3.53) に一致するが，チャンネルのサイズをダウ

ンサンプリングしたい場合，$s \geq 2$ に設定する場合がある。

畳み込み層で行われる「畳み込み」は，通常の文脈では「相関（correlation）」に対応する演算であり，通常の意味の「畳み込み」にするならば，式 (3.53) の $W_{c,p,l}$ を $W_{c,p,(L_{\mathrm{kernel}}-1)-l}$ に置換する必要がある。しかし，$\boldsymbol{W}$ は学習によって適応的に更新されるパラメータであるために，（既に向きが決まっている確定的な信号を畳み込むのとは異なり）$\boldsymbol{W}$ の成分の並びの向きに固執する必要はない。そのため，特別な事情がない場合，CNN の文脈においては，畳み込みと相関を同一視しても良い。

3.7.2 N 次元畳み込み層

ここまで，1 次元畳み込み層に限定して説明してきたが，1 次元畳み込み層は N 次元畳み込み層に拡張される。N 次元畳み込み層は，入出力チャンネルが N 階のテンソルである畳み込み層である。$\boldsymbol{X} \in \mathbb{R}^{C_{\mathrm{input}} \times L^{(0)}_{\mathrm{input}} \times \cdots \times L^{(N-1)}_{\mathrm{input}}}$ が入力として与えられたとき，重み $\boldsymbol{W} \in \mathbb{R}^{C_{\mathrm{output}} \times C_{\mathrm{input}} \times L^{(0)}_{\mathrm{kernel}} \times \cdots \times L^{(N-1)}_{\mathrm{kernel}}}$ $(L^{(n)}_{\mathrm{kernel}} \leq L^{(n)}_{\mathrm{input}}, n = 0, \ldots, N-1)$ とバイアス $\boldsymbol{b} \in \mathbb{R}^{C_{\mathrm{output}}}$ を持つ N 次元畳み込み層は以下のように $\boldsymbol{Y} \in \mathbb{R}^{C_{\mathrm{output}} \times L^{(0)}_{\mathrm{output}} \times \cdots \times L^{(N-1)}_{\mathrm{output}}}$ を出力する（簡単のため，ストライドは省略している）。

$$
\begin{aligned}
&Y_{c,l^{(0)},\ldots,l^{(N-1)}} \\
&= \sum_{c'=0}^{C_{\mathrm{input}}-1} \left(\sum_{l'^{(0)}=0}^{L^{(0)}_{\mathrm{kernel}}-1} \cdots \sum_{l'^{(N-1)}=0}^{L^{(N-1)}_{\mathrm{kernel}}-1} X_{c',l^{(0)}+l'^{(0)},\ldots,l^{(N-1)}+l'^{(N-1)}} W_{c,c',l'^{(0)},\ldots,l'^{(N-1)}} \right) + b_c
\end{aligned}
\tag{3.55}
$$

なお，パディングやストライドに関しては，1 次元畳み込み層と同様に考える。

N 次元畳み込み層から構成される CNN を N 次元 CNN と呼び，1 次元 CNN は波形データ等，2 次元 CNN は画像データ等を扱うために使用されることが多い。時間周波数表現は数値計算上画像データのように取り扱うことができるため，話者認識 [75, 76]，感情音声分類 [198, 199]，言語識別 [200, 201]，音源分離 [72, 202] 等の様々な課題で，時間周波数表現は 2 次元 CNN への入力として使用されている。その一方で，通常の画像と異なり，時間周波数表現では（特に周波数軸方向の）絶対位置に関する情報が重要である。畳み込み層は相対的な位置関係を重視する構造であり，2 次元 CNN に時間周波数を入力するのは一見不適切と考えられるかもしれない。しかし，実際の応用で高い性能を実現していることに加え [72, 75, 76, 198–202]，畳み込み層で付与されるゼロパディングを手掛かりに，2 次元 CNN は絶対的な位置の情報を獲得できることが実証されていることから [203]，時間周波数表現を 2 次元 CNN に入力することは有用であると考えられる。なお，時間周波数表現は時間波形にフィルタバンクを畳み込むことによって得られるため，1 次元 CNN を使って時間周波数表現を求めるフィルタバンクを学習するというアプローチもある [149, 167, 168]。

3.7.3　畳み込み層の派生

畳み込み層の中でも，カーネルサイズが $1 \times 1 \times \cdots \times 1$ であるものを特に点別畳み込み層（pointwise convolution layer）と呼ぶ。$\boldsymbol{X} \in \mathbb{R}^{C_{\text{input}} \times L^{(0)} \times \cdots \times L^{(N-1)}}$ が入力として与えられたとき，重み $\boldsymbol{W} \in \mathbb{R}^{C_{\text{output}} \times C_{\text{input}}}$ とバイアス $\boldsymbol{b} \in \mathbb{R}^{C_{\text{output}}}$ を持つ N 次元点別畳み込み層は以下のように $\boldsymbol{Y} \in \mathbb{R}^{C_{\text{output}} \times L^{(0)} \times \cdots \times L^{(N-1)}}$ を出力する（簡単のため，ストライドは省略している）。

$$Y_{c,l^{(0)},\ldots,l^{(N-1)}} = \sum_{c'=0}^{C_{\text{input}}-1} X_{c',l^{(0)},\ldots,l^{(N-1)}} W_{c,c'} + b_c \tag{3.56}$$

点別畳み込み層は，チャンネル間での各成分の足し合わせを行う一方で，チャンネル内での畳み込み演算が省略されている。

点別畳み込み層とは対照的に，深さ別畳み込み層（depthwise convolution layer）は，図 3.15 のように，チャンネル内での畳み込み演算を行う代わりに，チャンネル間の足し合わせが省略されている層である。$\boldsymbol{X} \in \mathbb{R}^{C \times L^{(0)}_{\text{input}} \times \cdots \times L^{(N-1)}_{\text{input}}}$ が入力として与えられたとき，重み $\boldsymbol{W} \in \mathbb{R}^{C \times L^{(0)}_{\text{kernel}} \times \cdots \times L^{(N-1)}_{\text{kernel}}}$ $(L^{(n)}_{\text{kernel}} \leq L^{(n)}_{\text{input}}, n = 0, \ldots, N-1)$ とバイアス $\boldsymbol{b} \in \mathbb{R}^{C}$ を持つ N 次元深さ別畳み込み層は以下のように $\boldsymbol{Y} \in \mathbb{R}^{C \times L^{(0)}_{\text{output}} \times \cdots \times L^{(N-1)}_{\text{output}}}$ を出力する（簡単のため，ストライドは省略している）。

$$\begin{aligned} &Y_{c,l^{(0)},\ldots,l^{(N-1)}} \\ &= \left(\sum_{l'^{(0)}=0}^{L^{(0)}_{\text{kernel}}-1} \cdots \sum_{l'^{(N-1)}=0}^{L^{(N-1)}_{\text{kernel}}-1} X_{c,l^{(0)}+l'^{(0)},\ldots,l^{(N-1)}+l'^{(N-1)}} W_{c,l'^{(0)},\ldots,l'^{(N-1)}} \right) + b_c \end{aligned} \tag{3.57}$$

点別畳み込み層ではチャンネル内の関係性を，層別畳み込み層ではチャンネル間の関係性を表現できないため，通常，これらの層は他の種類の畳み込み層と組み合わせて使用される。一つの使用例としては，畳み込み演算の計算量削減を目的に，通常の畳み込み層を深さ別畳み込み層と点別畳み込み層のセット（但し，各畳み込み層の後で活性化関数を適用する）で置換することがある [204]。入力チャンネル数が C_{input}，出力チャンネル数が C_{output}，カーネルサイズが $L^{(0)} \times \cdots \times L^{(N-1)}$ である畳み込み層では，入力の各位置に対して $C_{\text{input}} C_{\text{output}} \prod_{n=0}^{N-1} L^{(n)}$ 回の乗算が行われる。一方，この畳み込み層を，チャンネル数が C_{input} でカーネルサイズが $L^{(0)} \times \cdots \times L^{(N-1)}$ である層別畳み込み層と，入力チャンネル数が C_{input} で出力チャンネル数が C_{output} である点別畳み込み層とのセットでは，入力の各位置に対して $C_{\text{input}} \prod_{n=0}^{N-1} L^{(n)} + C_{\text{input}} C_{\text{output}}$ 回の乗算が行われる。つまり，畳み込み層を深さ別畳み込み層と点別畳み込み層のセットで置換することによって，各位置に対する乗算回数は $1/C_{\text{output}} + 1/\left(\prod_{n=0}^{N-1} L^{(n)}\right)$ 倍になり，出力チャンネル数やカーネルサイズが大きいほど計算量を大きく削減することができる。

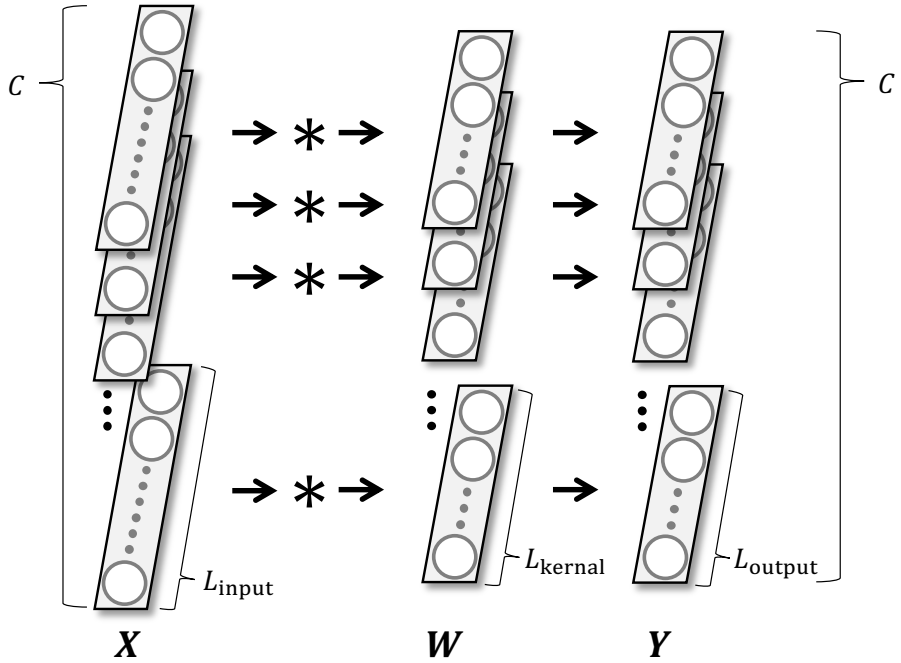


図 3.15 1 次元深さ別畳み込み層で行われる演算。但し，バイアスは省略されている。

3.7.4 プーリング層

プーリング層は，プーリング（pooling）と呼ばれる処理を行う層である。プーリングとは，局所的な範囲内の値を要約しつつ，ダウンサンプリングによって入力サイズを小さくする処理である。深さ別畳み込み層のように，プーリングはチャンネル毎に独立して行われる。値の要約方法としては，平均値か最大値を取るという方法を採ることが多い。

$\boldsymbol{X} \in \mathbb{R}^{C \times L^{(0)}_{\text{input}} \times \cdots \times L^{(N-1)}_{\text{input}}}$ が入力として与えられたとき，カーネルサイズが $L^{(0)}_{\text{kernel}} \times \cdots \times L^{(N-1)}_{\text{kernel}}$ でストライドが $s^{(0)} \times \cdots \times s^{(N-1)}$ であるような平均プーリング（average pooling）は，以下のように $\boldsymbol{Y} \in \mathbb{R}^{C_{\text{output}} \times L^{(0)}_{\text{output}} \times \cdots \times L^{(N-1)}_{\text{output}}}$ を出力する。

$$
\begin{aligned}
&Y_{c,l^{(0)},\ldots,l^{(N-1)}} \\
&= \frac{1}{\prod_{n=0}^{N-1} L^{(n)}_{\text{kernel}}} \left(\sum_{l'^{(0)}=0}^{L^{(0)}_{\text{kernel}}-1} \cdots \sum_{l'^{(N-1)}=0}^{L^{(N-1)}_{\text{kernel}}-1} X_{c,s^{(0)}l^{(0)}+l'^{(0)},\ldots,s^{(N-1)}l^{(N-1)}+l'^{(N-1)}} \right)
\end{aligned} \tag{3.58}
$$

また，同様の条件が与えられたとき，最大プーリング（max pooling）は次の計算を行う。

$$
\begin{aligned}
Y_{c,l^{(0)},\ldots,l^{(N-1)}} = \max_{\substack{l'^{(0)} \in \{0,\ldots,L^{(0)}_{\text{kernel}}-1\} \\ \vdots \\ l'^{(N-1)} \in \{0,\ldots,L^{(N-1)}_{\text{kernel}}-1\}}} X_{c,s^{(0)}l^{(0)}+l'^{(0)},\ldots,s^{(N-1)}l^{(N-1)}+l'^{(N-1)}}
\end{aligned} \tag{3.59}
$$

図 3.16 の例に示すように，プーリングのカーネルサイズとストライドは同じ値に設定されることが多い。従来は，プーリング層が入力チャンネルのサイズを小さくする役割を

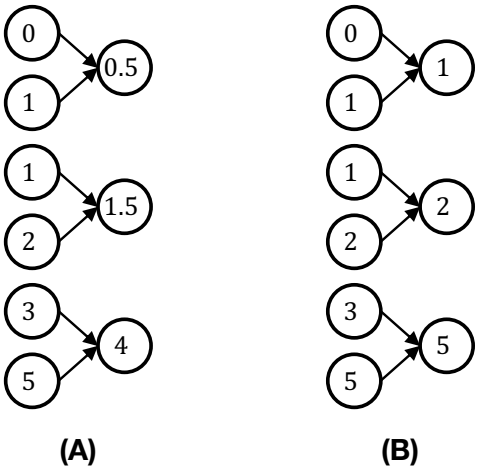


図 3.16 1 次元平均プーリング (A) と 1 次元最大プーリング (B) の例。ここでは，カーネルサイズとストライドはどちらも 2 としている。

担ってきたが [205, 206]，近年では，この役割は大きなストライドを持つ畳み込み層で代用されることが増えている [184, 204]。その一方で，大域プーリング（global pooling）と呼ばれる，カーネルサイズが入力チャンネルのサイズと同じであるプーリングが多用されるようになっている [184]。大域プーリングは，可変サイズ（$C \times L^{(0)} \times \cdots \times L^{(N-1)}$）の入力を固定サイズ（$C \times 1 \times \cdots \times 1$）にしたい場面等で使用される。

3.7.5 Batch Normalization

MLP のための Batch Normalization ではベクトルの成分毎に正規化が行われるのに対し，CNN のための Batch Normalization ではチャンネル毎に正規化が行われる [180]。隠れ表現テンソルのミニバッチ $\{\boldsymbol{Z}^{(b)} \in \mathbb{R}^{C \times L^{(0)} \times \cdots \times L^{(N-1)}}\}_{b=0}^{B-1}$ $(B, C, N, L^{(0)}, \ldots, L^{(N-1)} \in \mathbb{N})$ が与えられたとき，N 次元 Batch Normalization は，$\{\boldsymbol{Z}^{(b)}\}_{b=0}^{B-1}$ に対する平均 $\boldsymbol{\mu} \in \mathbb{R}^C$ と標準偏差 $\boldsymbol{\sigma} \in \mathbb{R}^C$ を用いて，正規化されたミニバッチ $\{\tilde{\boldsymbol{Z}}^{(b)} \in \mathbb{R}^{C \times L^{(0)} \times \cdots \times L^{(N-1)}}\}_{b=0}^{B-1}$ を次のように計算する。

$$\tilde{Z}^{(b)}_{c,l^{(0)},\ldots,l^{(N-1)}} = \frac{Z^{(b)}_{c,l^{(0)},\ldots,l^{(N-1)}} - \mu_c}{\sqrt{\sigma_c^2 + \epsilon}} \cdot \gamma_c + \beta_c \tag{3.60}$$

$$\mu_c = \frac{1}{B} \sum_{b'=0}^{B-1} \left(\frac{1}{\prod_{n=0}^{N-1} L^{(n)}} \sum_{l'^{(0)}=0}^{L^{(0)}-1} \cdots \sum_{l'^{(N-1)}=0}^{L^{(N-1)}-1} Z^{(b')}_{c,l'^{(0)},\ldots,l'^{(N-1)}} \right) \tag{3.61}$$

$$\sigma_c = \sqrt{\frac{1}{B} \sum_{b'=0}^{B-1} \left(\frac{1}{\prod_{n=0}^{N-1} L^{(n)}} \sum_{l'^{(0)}=0}^{L^{(0)}-1} \cdots \sum_{l'^{(N-1)}=0}^{L^{(N-1)}-1} (Z^{(b')}_{c,l'^{(0)},\ldots,l'^{(N-1)}} - \mu_c)^2 \right)} \tag{3.62}$$

ここで，$\epsilon > 0$ はゼロ除算を防ぐ微少量，$\gamma \in \mathbb{R}^C_{>0}$ と $\beta \in \mathbb{R}^C$ は $\gamma = 1$，$\beta = 0$ で初期化される学習可能なパラメータである。なお，典型的には，N 次元 Batch Normalization は畳み込み層と活性化関数の間で適用され，N 次元 Batch Normalization の直前の畳み込み層のバイアスは一般的に省略される。

3.8 まとめ

本章では，学習アルゴリズムの根幹である教師あり学習と確率的勾配降下法，DNN の数理モデルである MLP と CNN，RNN，および，DNN の学習上の困難さを緩和する方法について説明を行った。後述する本論文の実験では，本章で説明した技術を土台にして DNN を構築する。

第 4 章

病的音声データベースの構築

4.1　はじめに

これまでの章では，時間周波数表現と深層学習に関する理論的説明を行った。本章では，本研究を遂行する上で必須となる，病的音声データベースの構築について述べる。声質の自動評価に応用可能な病的音声データベースを作成するためには，病的音声データを収集することに加えて，各音声データに対して人間が声質の聴覚心理的評価を行う必要がある。第 1.2 節でも説明したように，声質評価は主観的な作業であるため，評者内変動や評者間変動の問題が存在する。機械学習による自動推定の性能を，人間の評者内信頼性や評者間信頼性と比較するためにも，評者の人数は多いことが望ましい。しかし，音声データの声質評価は大きな労力を要するため，大量の音声データに多数の評者による声質評価を行うことは困難である。

本研究では，まず，病的音声データを収集し，収集したすべての音声データに対して 1 名の評者による声質評価を行う。次に，サンプル数が少なくても多様な音声データが含まれるように考慮して，1 名の評者による声質評価の結果に基づき，300 サンプルの音声データを選択した。最後に，選択した音声データに対して，複数名の評者による声質評価を行った。このような手順を踏むことによって，音声データに複数名の評者による声質評価の結果が付与された，多様な声質の音声データを含む病的音声データベースを構築した。なお，第 5 章では，1 名の評者による声質評価の結果が付与された 2,545 サンプルの病的音声データを，第 7 章では，複数名の評者による声質評価の結果が付与された 300 サンプルの病的音声データを使用した。

4.2　音声データの収集

音声データとしては，九州大学病院耳鼻咽喉科で録音された音声データと，九州大学大橋キャンパスで録音された音声データを用いた。前者の音声データは，九州大学病院耳鼻咽喉科に来院した患者が発声したものを，同病院の関係者が録音したものであり，その録音自体に筆者は関わっていない。その一方で，後者の音声データは，基本的に正常な発声が可能である九州大学大橋キャンパスの学生が発声し，筆者が録音したものである。本節では，音声データの録音環境と，発話内容について説明する。

2017 年 7 月 26 日時点において，九州大学病院耳鼻咽喉科では，防音室でマイクロフォン（RODE: NT2-A）を用いて録音が行われていた。しかし，本研究で用いた九州大学病院耳鼻咽喉科の音声データは，2008 年から 2017 年にかけて録音されたものであり，録音時期によってどのような設備が使用されていたかは不明である。録音時期によって絶対的な録音レベルや SN 比が大きく異なっていた他，標本化周波数が 96 kHz のデータと 50 kHz のデータ，およびモノラルのデータとステレオのデータが混在していたため，使用機材や設定は一貫していなかった可能性がある。九州大学大橋キャンパスでは，防音室でマイクロフォン（Brüel & Kjær: Model 4189）を用いて，あるいは，無響室でマイクロフォン（Audio-Technica: AT4040）を用いて録音を行い，標本化周波数は 48 kHz としてモノラルで符号化を行った。なお，九州大学大橋キャンパスでの録音は 2017 年に行った。このように，音声データの録音条件はある程度多様なものであった。

九州大学病院耳鼻咽喉科で録音されていた音声データの発話内容には，録音時期による差異があるものの，ほぼすべての音声データに，持続母音/a/，/e/の発声，および『ジャックと豆の木』の冒頭部分の音読が含まれていた。そのため，九州大学大橋キャンパスの音声収録での発話内容も同様に，持続母音/a/，/e/の発声，および『ジャックと豆の木』の冒頭部分の音読とした。本研究では，これらの発話内容のうち，持続母音/a/の起声（オンセット）から終声（オフセット）までを切り出し，切り出した持続母音/a/を声質評価の対象として用いた。

4.3　単一評者による声質評価

本節では，収集したすべての音声データに対して 1 名の評者による声質の聴覚心理的評価を行った内容について述べる。音声データの量が多かったため，聴覚心理的評価は，2017 年から 2018 年にかけて断続的に実施した。評者は耳鼻咽喉科医であり，2017 年時点で音声を専門とした経験年数は 5 年であった。聴覚心理的評価の内容は，持続母音/a/の各音声データに対する GRBAS 尺度のすべての項目の評点を付与するというも

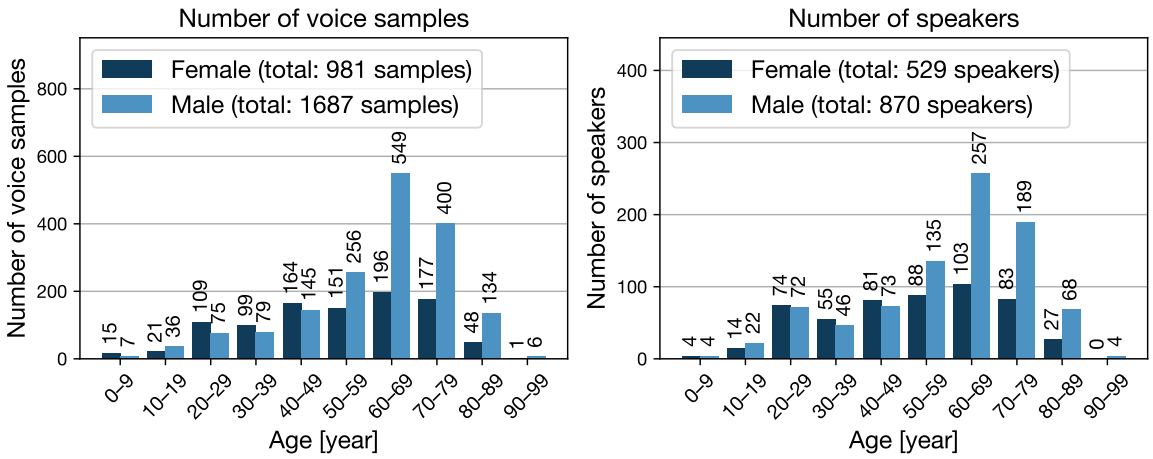


図 4.1 1 名の耳鼻咽喉科医によって聴覚心理的評価が行われた音声データに関する，年齢と性別の分布（左：音声サンプル数，右：話者数）。なお，話者数の分布に関して，複数の音声データを発声していた話者の年齢は，それらの音声データ発声時の年齢の平均値を取り，小数点以下を切り捨てた値とした。

のであり，GRBAS 尺度の評点は Microsoft Excel のスプレッドシートに直接入力した。音声は，パソコン（Apple: MacBook Air 13-inch Mid 2013）に接続したオーディオインタフェース（cakewalk by Roland: UA-25EX）を経由して，ヘッドフォン（SONY: MDR-CD900ST）から再生した。音量は，評者が適切だと考える音量に適宜調節した。また，第 4.2 節で述べたように音声データの元々の標本化周波数は一定ではなく，評価開始当初は 48 kHz のデータと 50 kHz のデータが混在していたが，途中から 48 kHz にリサンプリングして統一した。なお，どちらもナイキスト周波数は 24 kHz 以上であり，これは，例えば電話の通話帯域と比べて十分に広いため，標本化周波数の違いに対して聴覚的な差異はほとんど生じないと考えられる。

すべての音声データを聴取した結果，音声データの中には，明らかに音質が悪いものや，音声以外の雑音の音量が大きいものが含まれていた。そのため，聴感的に音質が悪い音声データや，SN 比が A 特性音圧レベルで 30 dB 未満である音声データは除外した。なお，SN 比の基準は，SN 比が A 特性音圧レベルで 30 dB を下回る場合に，シマ（shimmer）等の声質に関連する音響特徴量の計算結果の信頼性が，大きく低下しうることを指摘した Deliyski らの論文 [207] に基づき設定した。

上記の除外作業を行った後，音声データの総数は最終的に 2,668 サンプル（女性：981 サンプル，男性：1,687 サンプル）となった。複数の音声サンプルで重複する話者もいたため，話者数は音声データの数より少なく，1,399 名（女性：529 名，男性 870 名）となった。また，九州大学病院耳鼻咽喉科で録音された音声データは 2,598 サンプル（女性：961 サンプル，509 名，男性：1,637 サンプル，820 名），九州大学大橋キャンパスで録音された音声障害ではない話者の音声データは 70 サンプル（女性：20 サンプル，20 名，男性：

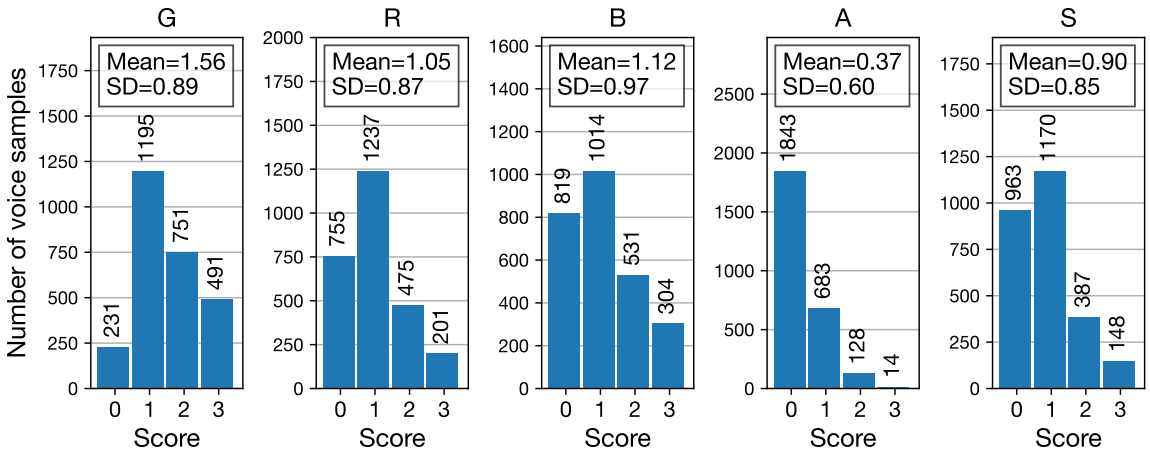


図 4.2 1 名の耳鼻咽喉科医によって聴覚心理的評価が行われた音声データに関する，GRBAS 尺度の評点の分布。なお，縦軸は話者数ではなく音声サンプル数を表す。

50 サンプル，50 名）であった。図 4.1 に話者の年齢と性別の分布を示す。GRBAS 尺度の評点の分布は，図 4.2 のようになり，分布に大きな偏りがあることが観察された。

4.4　音声データの選別

本節では，前節で述べた 1 名の耳鼻咽喉科医によって GRBAS 尺度の評点が付与された持続母音/a/のデータから，複数名の評者による聴覚心理的評価を実施するためのデータの選択方法について説明する。本章冒頭でも述べたように，音声データを選択する際は，サンプル数が少なくても，評点の分布の観点から多様な音声データが含まれるように考慮した。

なお，後述の複数名の評者による聴覚心理的評価の際には，評者間変動や評者内変動を低減するために，GRBAS 尺度の評点が明示的に与えられた音声刺激であるアンカー（anchor）を導入した。Kreiman らによって，評者がある音声の声質を評価する作業は，評者自身が持つ声質に関する内的基準（internal standard）とその音声を照らし合わせる作業であるとモデル化できることが提案されている [25]。このモデルに基づくと，評者間変動は，評者のこれまでの臨床などの経験によって形成された内的基準の違いによって生じると考えることができる [25, 33, 34]。また，評者内変動は，記憶や注意力の低下等の内部要因や，音響的文脈や聴取課題等の外部要因による内部基準の変動によって生じると考えることができる [25, 34, 38]。アンカーは，上述したような内的基準の個人間の差異や個人内の変動を低減するために，固定された外的基準（external standard）として導入されたものである [25, 38]。実際，アンカーを導入することによって，評者間信頼性は多くの場合で改善し [38, 208–210]，評者内信頼性は有意に改善しない場合もあれば [208–210]，有意に改善する場合もある [38, 208] ことが報告されている。このことから，アンカーを

表 4.1 アンカーとして用いた音声に関する情報。GRBAS 尺度の評点は，第 4.3 節で与えられたものである。疾患名は，基本的に音声障害診療ガイドラインによる分類 [9] に従った。但し，疾患名「喉頭悪性腫瘍（上皮内癌を含む）」は「喉頭癌」と表記した。なお，R2 のアンカーの疾患名「異形成（白板症を含む）」と「喉頭癌」に関しては，疑いはあったが，生検で確定診断されたものではない。

			評点					
アンカー	性別	年齢	G	R	B	A	S	疾患名
R1	男性	30–39 歳	1	1	0	0	0	片側声帯麻痺（手術後）
R2	男性	40–49 歳	2	2	0	0	1	異形成（白板症を含む），喉頭癌，急性喉頭炎
R3	男性	50–59 歳	3	3	1	0	1	声帯ポリープ
B1	女性	50–59 歳	1	0	1	0	0	片側声帯麻痺
B2	女性	50–59 歳	2	0	2	0	1	片側声帯麻痺（手術後）
B3	女性	80–89 歳	3	0	3	1	2	両側声帯麻痺（手術後）

導入することによって，病的音声データベースに付与される聴覚心理的評価の評点の信頼性が向上することが期待される。

音声データの選別の順序に関して，アンカーを選定した後で，複数名の評者によって声質を評価するための音声データを選定した。また，ここで選定されなかった残りの音声データの中から，機械学習モデルが声質自動評価の訓練のために使う音声データを抽出した。以下，音声データの選別について順に説明する。

4.4.1　アンカーの選定

アンカーは，第 4.3 節で声質評価を行った評者が，自身が評価した 2,668 サンプルの音声データの中から選定した。アンカーは，実際に持続母音/a/の音声を聴取しながら，R1（項目 R の評点が 1），R2，R3，B1，B2，B3 の声質を代表するような音声を探して選定した。選定した音声に関する情報については，表 4.1 に示す。

4.4.2　複数評者によって声質を評価するための音声データの選定

アンカーの選定に引き続き，複数評者によって声質を評価するための音声データを選定した。ここで選定対象の候補となる音声データからは，アンカーの音声データの話者によって発声されたものを除外した。また，耳鼻咽喉科の実際の臨床場面で聴くような音声データに限定することを意図して，九州大学大橋キャンパスで録音された音声データを除外した。加えて，採録回数（各話者に関して存在する音声データのサンプル数）が複数回あった話者の音声データに関しては，できる限り治療が行われていない状態の初診時に近いものを使うことを意図して，時系列順で 2 回目以降に採録された音声データは除外し

た。これらの除外作業を行った後，以下の手順で，音声データ 300 サンプルを機械的に抽出した。

1. まず，G0 の音声をランダムに 30 サンプル（男女 15 サンプルずつ）抽出した。
2. 次に，残った音声サンプルの中に存在する GRBAS 尺度の全項目の評点の組み合わせパターンに関して，各組み合わせパターンに対して 1 サンプルずつ音声データをランダムに抽出した。なお，組み合わせパターンの総数は，純粋に考えると 256 ($= 4^5$) であるが，G3R0B0A0S0 のような評点は存在しないことや，G0 の音声は除外したため，実際のパターンの総数は 256 より少なかった。抽出の際は，採録回数が 1 回である話者の音声データを優先的に選択し，採録回数が 1 回である話者の音声データの中には存在しない評点パターンに関してのみ，採録回数が複数回である話者の音声データを選択した。ここで採録回数が 1 回である話者の音声データを優先した理由に関しては，第 4.4.3 節で述べる。
3. 最後に，G1，G2，G3 の音声データのサンプル数がそれぞれ 90 サンプルずつになるように，音声データを抽出した。抽出の際は，サンプル数が少ない評点（例えば，R2 や A3 等）の音声データを優先的に選択した。

また，各評者の評者内信頼性を調べるためには，各評者が同じ音声データを 2 回評価する必要があるが，300 サンプルの音声データのすべてを 2 回評価すると，実験参加者の負担が大きく増加してしまう。そこで，この 300 サンプルの音声データの中から，各評者によって 2 回評価される 50 サンプルの音声データを，以下の手順で機械的に抽出した。

1. まず，G0 の音声をランダムに 5 サンプル（女性：2 サンプル，男性：3 サンプル）抽出した。
2. 次に，サンプル数が少ない評点（例えば，R2 や A3 等）の音声データを優先しつつ，GRBAS 尺度の全項目の評点の組み合わせパターンの重複もないように，残り 45 サンプルの音声データをランダムに抽出した。

抽出した音声データの年齢と性別の分布を図 4.3 (A)，(B) に，GRBAS 尺度の評点の分布を図 4.4 (A)，(B) に示す。GRBAS 尺度の各項目に関して，抽出後の評点の分布の標準偏差は，元々の評点の分布（図 4.2）の標準偏差よりも大きくなっており，上述した抽出アルゴリズムによって，評点の分布の偏りがある程度軽減されたことが確認された。また，抽出した 300 サンプルの音声データについては，話者が罹患していた疾患について調査した。調査の結果，疾患の分布は表 4.2 のようになり，喉頭癌や声帯結節のような声帯の物性変化を伴う音声障害，片側声帯麻痺のような神経異常による音声障害，過緊張性発声障害のような機能的な音声障害など，様々な音声障害が含まれていたことがわかった。

4.4.3　第 5 章で機械学習モデルの訓練のために使用する音声データの抽出

アンカーの話者でも，第 4.4.2 節で抽出された 300 サンプルの話者でもない話者によって発声された音声データは合計で 2,245 サンプル（女性：798 サンプル，393 名，男性：1,447 サンプル，701 名）となった。第 5 章では，この 2,245 サンプルを訓練データセットとして，第 4.4.2 節で選定した 300 サンプルをテストデータセットとして使用する。訓練データセットとして用いた音声データの年齢と性別の分布を図 4.3 (C) に，GRBAS 尺度の評点の分布を図 4.4 (C) に示す。

第 4.4.2 節で 300 サンプルを抽出した際に，採録回数が複数回である音声データよりも採録回数が 1 回である音声データを優先したのは，第 5 章の訓練用データセットのサンプル数を増やすためである。訓練データセットとテストデータセットに同じ話者の音声データが含まれていた場合，データ漏洩（data leakage）によってテストデータセットに対する性能が不当に高くなってしまう可能性がある。そのため，訓練データセットとテストデータセットでの話者の重複は避ける必要がある。但し，訓練データセットの抽出では話者の重複を避けるよう十分に注意したが，第 5 章の実験が終わった後で患者情報を正確に調べた結果，1 名の話者について，複数評者によって声質が評価される音声データ（2 回評価されない）と第 5 章の訓練データセットの間で重複があることが判明した。しかし，1 名という数値は全体に対して 1% 未満であり，結果に与えた影響はごく軽微であったと考えられる。

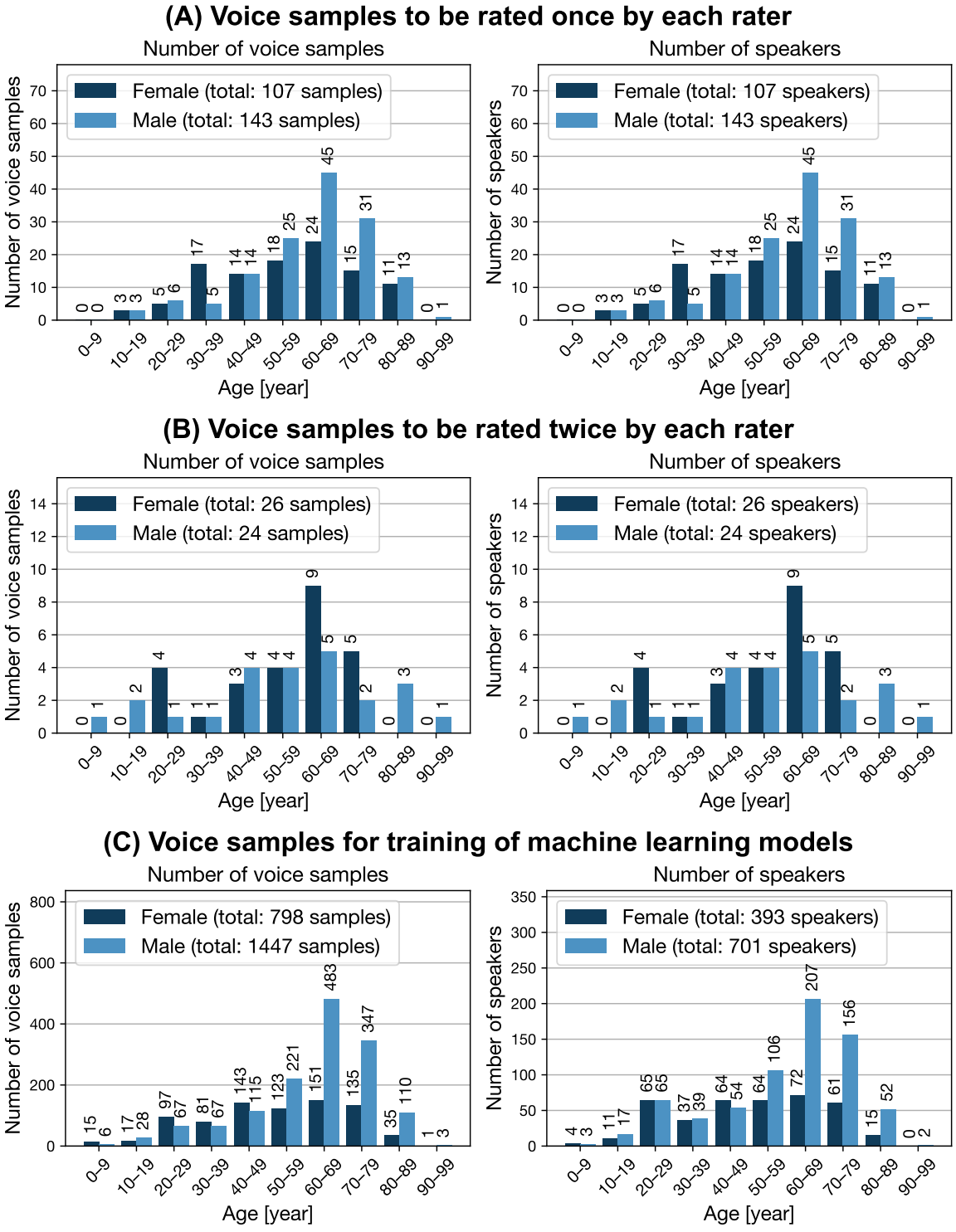


図 4.3 1 名の耳鼻咽喉科医によって聴覚心理的評価が行われた音声データに関する，年齢と性別の分布。なお，複数の音声データを発声していた話者の年齢は，それらの音声データ発声時の年齢の平均値を取り，小数点以下を切り捨てた値とした。(A) は複数名の各評者によって 1 回のみ評価された音声データ，(B) は複数名の各評者によって 2 回評価された音声データ，(C) は第 5 章の声質自動評価実験の訓練データセットとして使用した音声データに関する分布を表す。

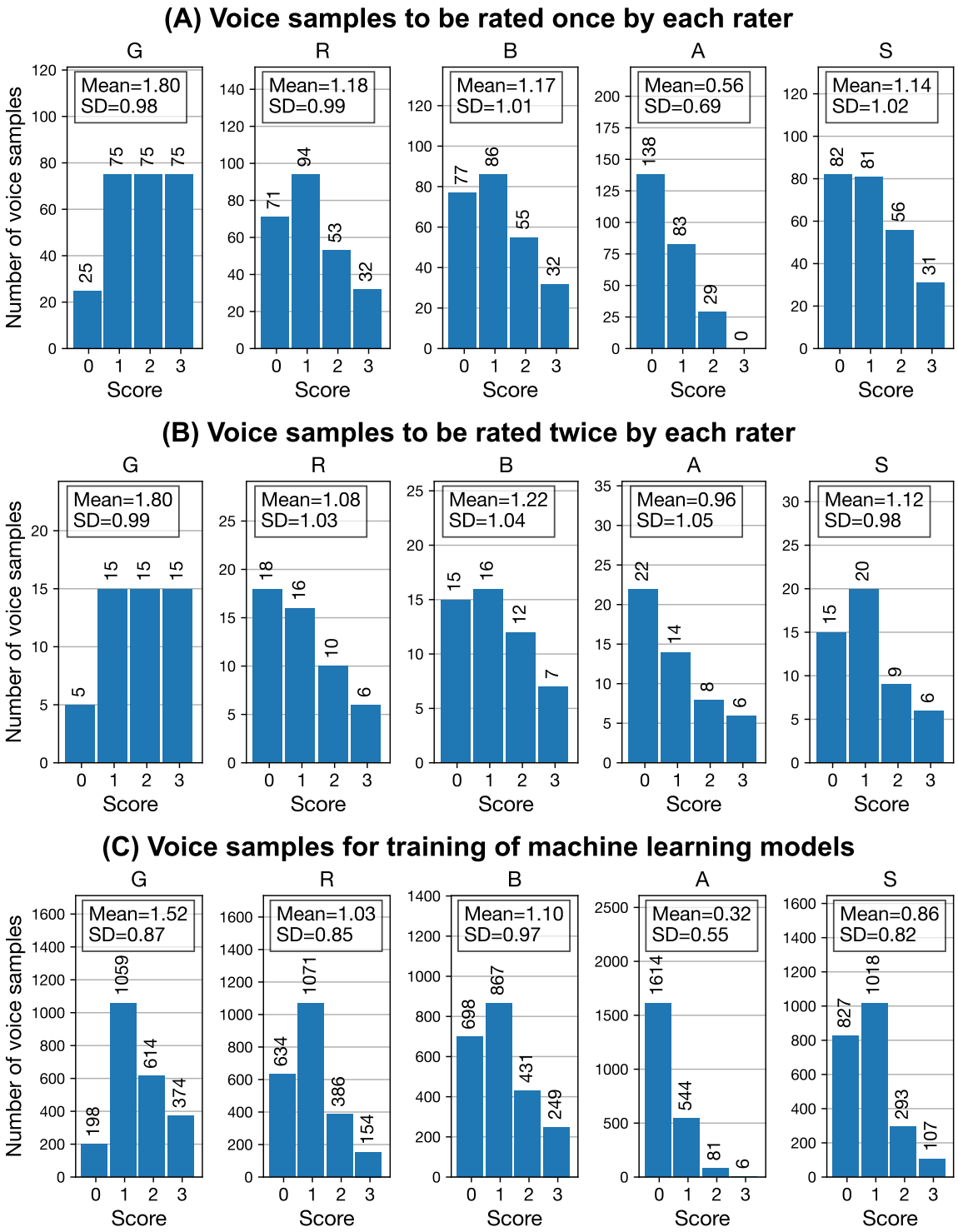


図 4.4 1 名の耳鼻咽喉科医によって聴覚心理的評価が行われた音声データに関する，GRBAS 尺度の評点の分布。(A) は複数名の各評者によって 1 回のみ評価された音声データ，(B) は複数名の各評者によって 2 回評価された音声データ，(C) は第 5 章の声質自動評価実験の訓練データセットとして使用した音声データに関する分布を表す。

表 4.2 複数評者による聴覚心理的評価の対象とした 300 サンプルの音声データについての，話者が罹患していた疾患の度数分布。なお，300 サンプルの音声データはすべて異なる話者によって発声されたため，話者数と音声サンプル数は同じである。複数の疾患を罹患していた話者に関しては，罹患したすべての疾患を数えたため，表中の数値の合計は話者数である 300 よりも多い。疾患名と疾患コードは，基本的に音声障害診療ガイドラインによる分類 [9] に従った。但し，疾患名「喉頭悪性腫瘍（上皮内癌を含む）」は「喉頭癌」と表記し，表の末尾には，音声障害診療ガイドラインに載っていない疾患も載せている。疾患名「正常範囲内」は，特記すべき音声障害がなかったことを表す。表中の「治療」は，「手術または放射線治療」のことを指す。「治療前」の列の括弧内の数値は，治療歴があるが病気が再発した話者数を表す。例えば，「病名コード」が 1120 である行の「治療前」の数値「12 (1)」は，発声当時に喉頭癌に罹患していた話者数が 12 名であり，内 1 名が再発患者であったことを表す。また，「病名コード」が 1120 である行の「治療後」の数値「22」は，発声当時に喉頭癌を治療済みであり，かつ喉頭癌が再発していなかった話者数が 22 名であったことを表す。

		話者数	
疾患コード	疾患名	治療前	治療後
1110	異形成（白板症を含む）	10 (0)	8
1120	喉頭癌	12 (1)	22
1130	喉頭乳頭腫	2 (2)	1
1140	その他の腫瘍	11 (1)	3
1210	声帯結節	15 (1)	2
1220	声帯ポリープ	12 (3)	3
1230	声帯嚢胞	4 (0)	1
1240	ポリープ様声帯	9 (0)	6
1250	声帯瘢痕	1 (0)	1
1260	声帯溝症	2 (0)	0
1270	喉頭肉芽腫	2 (0)	0
1280	その他の声帯粘膜異常	11 (0)	4
1430	加齢性声帯萎縮	3 (0)	0
1500	喉頭の瘢痕・狭窄	1 (0)	0
1510	声門下狭窄	1 (0)	0
1520	声門狭窄 / 喉頭狭窄	0 (0)	1
2210	急性喉頭炎	10 (0)	0

2300	咽喉頭逆流症	1 (0)	0
3000	喉頭の外傷	1 (0)	0
4230	膠原病	1 (0)	0
5100	呼吸器疾患	1 (1)	0
5110	気管支喘息	1 (0)	0
5210	胃食道逆流症	1 (0)	0
6110	心因性発声障害	4 (0)	0
6210	うつ病 / 大うつ病性障害	1 (0)	0
6400	その他の心理的疾患・精神疾患	2 (0)	0
7110	上喉頭神経麻痺	3 (0)	0
7120	片側声帯麻痺	49 (0)	27
7130	両側声帯麻痺	9 (0)	2
7150	その他の末梢神経・神経筋接合部障害	2 (0)	1
7210	内転型痙攣性発声障害	4 (0)	0
7220	外転型痙攣性発声障害	1 (0)	0
7240	音声振戦	3 (0)	0
7250	パーキンソン病	1 (0)	0
7260	その他の中枢神経障害	4 (0)	0
8110	過緊張性発声障害	24 (0)	0
8120	低緊張性発声障害	4 (0)	0
8200	変声障害	3 (0)	0
8300	仮声帯発声	3 (0)	0
8400	奇異性声帯運動	1 (0)	0
8500	その他の音声障害	10 (0)	1
9000	原因不明の音声障害	1 (0)	0
-	嚥下障害	13 (0)	0
-	吃音	3 (0)	0
-	上咽頭癌・上咽頭腫瘤	0 (0)	2
-	その他の疾患	2 (0)	0
-	正常範囲内	32 (0)	0

4.5　複数評者による声質評価：聴覚実験

前節で抽出した 300 サンプルの持続母音/a/の音声データを用いて，17 名の評者による聴覚実験を実施した。なお，この聴覚実験では，病的音声データベースの構築だけではなく，評価の様々な側面を調べることも目的として，実験参加者の課題は GRBAS 尺度の評価以外にも以下に示す項目を含めた。

- GRBAS 尺度の評価：下記の各項目をそれぞれ 0，1，2，3 のいずれかで評価
  - 項目 G（Grade of hoarseness）
  - 項目 R（Roughness）
  - 項目 B（Breathiness）
  - 項目 A（Asthenia）
  - 項目 S（Strain）
- CAPE-V [41] の声質に関わる定量評価項目の評価：下記の各項目をビジュアルアナログスケールで評価
  - Overall Severity：GRBAS 尺度の項目 G に対応
  - Roughness：GRBAS 尺度の項目 R に対応
  - Breathiness：GRBAS 尺度の項目 B に対応
  - Strain：GRBAS 尺度の項目 S に対応
- CAPE-V で定性的に記述するように設定されている声質の判定：下記の各項目を有りか無しかの二値で評価
  - 二重声（diplophonia）
  - ボーカルフライ（fry）
  - 裏声（falsetto）
  - 無力性（asthenia）
  - ピッチ不安定（pitch instability）
  - 音声振戦（tremor）
  - 唾液貯留音（wet/gurgly）
  - その他（other，直接書き下す）
- 音声障害の疾患名の推定：下記の各項目を疑い有りか疑い無しかの二値で評価
  - 喉頭癌（laryngeal cancer）
  - 片側声帯麻痺（片側反回神経麻痺; uni-lateral recurrent laryngeal nerve paralysis/palsy; uni-lateral RLNP）
  - 声帯ポリープ（polyp）・声帯結節（nodule）

  - 過緊張性発声障害（muscle tension dysphonia; MTD）
  - ポリープ様声帯（Reinke 浮腫; Reinke's eduma）
  - 痙攣性発声障害（spasmodic dysphonia; SD）
- 確信度（confidence）：GRBAS 尺度の評価の確信の度合いをビジュアルアナログスケールで評価

CAPE-V [41] は，GRBAS 尺度に影響を受けて，ASHA で考案された声質の聴覚心理的評価法である。CAPE-V で定量評価される声質の項目は，Overall Severity，Roughness，Breathiness，Strain の 4 種類であり，それぞれ GRBAS 尺度の項目 G，R，B，S に対応する。GRBAS 尺度では各項目を離散尺度（0，1，2，3）で評価するのに対し，CAPE-V では連続尺度（0〜100）で評価する。また，CAPE-V では，定量評価する項目に含まれない声質に関しては定性的に記述するようになっており，この聴覚実験では，それらの声質の項目の判定に関しても調査対象とした。加えて，聴覚のみを使ってどれだけの精度で音声障害の疾患名を推定できるか調査するため，疾患推定も検討に入れた。また，自身の GRBAS 尺度の評点に対する確信度も回答項目とした。推定対象の疾患名としては，実際に来院する患者が罹患していることが多いものを設定した。本節では，実験参加者，実験手続き，実験結果について順に述べる。

4.5.1　実験参加者

表 4.3 に，聴覚心理的評価実験に評者として参加した，実験参加者の職業と経験年数の一覧を示す。実験参加者は，アンカーを選定した評者を含めて耳鼻咽喉科医 8 名，言語聴覚士 4 名，耳鼻咽喉科医でも言語聴覚士でもない非専門家 5 名であった。実験参加者は，専門性の違いと，音声を専門とした臨床の経験年数によって，経験年数が 2 年以上の専門家群，経験年数が 2 年未満の専門家群，非専門家群の 3 つに分けた。なお，非専門家群のうち 2 名は，「聴能形成」と呼ばれる大学の授業を受講した経験があった。聴能形成 [211] は，「音を聴く態度を修得し，物理的特徴と関連させた音の記憶を蓄える」ことを目的とし，1969 年に九州芸術工科大学音響設計学科で創設され，2023 年現在は九州大学芸術工学部音響設計コースで行われている授業である。聴能形成では，聴覚によって知覚した主観的な音情報を，音の高さ，大きさ，音色等を表す物理量と対応づける訓練が行われることに加え，音響心理学的知識に関する講義が行われる。このことから，聴能形成を受講した人は，受講していない人と比べて微妙な声質の違いを聴き分ける能力が高い可能性があることに留意する必要がある。

表 4.3 300 サンプルの持続母音/a/の音声データを用いた聴覚心理的評価実験に，評者として参加した実験参加者の職業と経験年数。「経験年数」は，音声を専門とした臨床経験の年数を表し，評者番号は評者年数が降順になるように設定した。専門性と経験年数（2 年以上か 2 年未満か）で群分けを行った。

評者番号	専門性	経験年数	職業	備考
1	専門家	20	耳鼻咽喉科医	—
2	専門家	15	耳鼻咽喉科医	—
3	専門家	13	耳鼻咽喉科医	—
4	専門家	11	言語聴覚士	—
5	専門家	10	言語聴覚士	—
6	専門家	7	耳鼻咽喉科医	第 4.3 節で声質評価を行い， 第 4.4.1 節でアンカーを選定した評者
7	専門家	4	耳鼻咽喉科医	—
8	専門家	2	耳鼻咽喉科医	—
9	専門家	1.5	耳鼻咽喉科医	—
10	専門家	1	耳鼻咽喉科医	—
11	専門家	0	言語聴覚士	—
12	専門家	0	言語聴覚士	—
13	非専門家	0	その他	聴能形成を受講済み
14	非専門家	0	その他	聴能形成を受講済み
15	非専門家	0	その他	—
16	非専門家	0	その他	—
17	非専門家	0	その他	—

4.5.2　実験手続き

専門家群と非専門家群では，聴覚心理的評価についての前提知識が異なるため，非専門家群に対してのみ，事前に GRBAS 尺度の定義付けや，評価の訓練を行った。本節では，専門家群と非専門家群で共通の手続きについて説明した後，非専門家群に対して行った事前手続きについて述べる。

専門家群と非専門家群で共通の手続き

実験参加者は，まず初めに，R1，R2，R3，B1，B2，B3 のアンカーを順に聴取した後，6 回の練習試行に続き，350 回の本試行を行った。6 回の練習試行では，音声データとしてアンカーを用いた。350 回の本試行の内，最初の 300 回で重複のない 300 サンプルの音声データによる 1 回目の評価を行い，最後の 50 回で再評価用の音声データを用いた 2 回目の評価を行った。なお，音声データの提示順序は実験参加者ごとにランダマイズした。

実験のためのコンピュータプログラムは，心理学を含む行動科学の実験のためのソフト

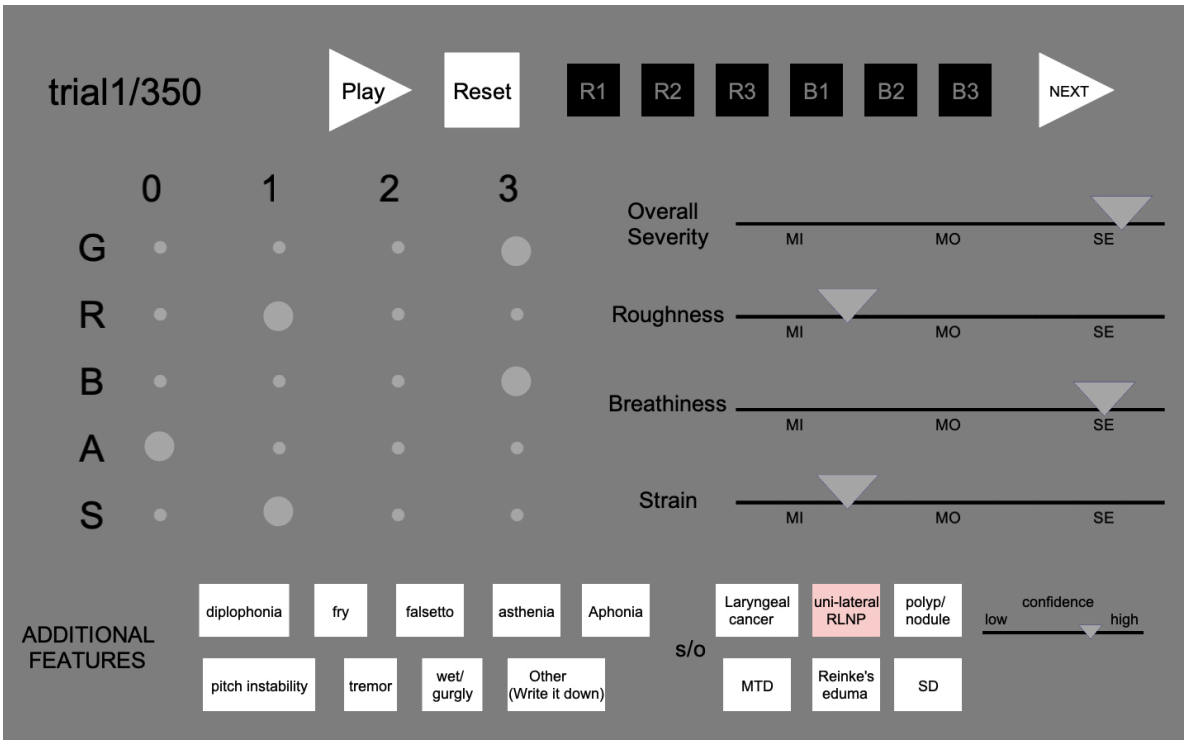


図 4.5 聴覚実験で実験参加者へ提示した GUI のスクリーンショット。この例では，評価対象の持続母音/a/の音声データが，評者によって，G3，R1，B3，A0，S1 と評価され，片側声帯麻痺が疑われている。なお，非専門家群の実験参加者に対しては，練習試行と本試行の際は CAPE-V の定性的な声質の特徴（ADDITIONAL FEATURES）と疾患推定（s/o）の入力部を非表示にし，GRBAS 尺度の評価の訓練の際はさらに CAPE-V の定量評価項目も非表示にした。

ウェアパッケージである PsychoPy [212] を用いて作成した。実験参加者には，図 4.5 に示すグラフィカルユーザインタフェース（GUI）を用いて，各音声データに対する回答を入力するように指示した。CAPE-V の定性的な声質の特徴と疾患名は，同時に複数項目選択することを可能とした。また，非専門家群に関しては，CAPE-V の定性的な声質の特徴と疾患名は評価項目の対象外とし，これらの項目は GUI 上でも非表示にした。各試行で，実験参加者は，評価対象の音声データとアンカーを任意の回数聴取することが可能であった。但し，一度回答を完了すると，以前の音声データの修正はできないようにした。各試行に対する回答時間は制限せず，任意のタイミングで休憩をとることが可能であったが，被験者ごとに，実験全体は 1 日間以内で終えるようにした。音声データは，パソコン（Apple: MacBook Air 13-inch Early 2015，あるいは Apple: MacBookPro 13-inch 2016 four Thunderbolt 3 ports）に接続したオーディオインタフェース（ZOOM: UAC-2）を経由して，ヘッドフォン（Sennheiser: HDA 200）から再生した。再生時の音量は，練習試行の間に被験者自身で調整を行い，本試行では音量を変更しないように教示した。なお，実験開始前に，6 サンプルのアンカーと，300 サンプルの評価用の音声デー

タは，A 特性重み付け音圧レベル（等価騒音レベル L_{Aeq}）が等しくなるようにデジタル信号処理で正規化した他，標本化周波数は 48 kHz に統一した。

非専門家群に対する GRBAS 尺度の定義付けと訓練

非専門家群に対しては，先述のアンカー提示，練習試行，および本試行を行う前に，GRBAS 尺度の定義付けと，GRBAS 尺度の評価の訓練を続けて実施した。ここでは，その内容について説明する。

GRBAS 尺度の定義付けとしては，日本音声言語医学会より出版されている資料を用いて，文章による定義付けと，聴覚による定義付けを行った。まず，文章による定義付けとして，“音声障害診療ガイドライン 2018 年版” [9] と，“動画で見る音声障害 ver. 2.0（DVD-ROM）” [213] に付属する冊子に記載されている GRBAS 尺度の定義と評価方法を抜粋したものを実験参加者に提示した。その後，“嗄声のサンプルテープ” [214] と “動画で見る音声障害 ver. 2.0（DVD-ROM）” [213] より抜粋した音声データを用いて，聴覚による GRBAS 尺度の定義付けを行った。表 4.4 に，聴覚的な定義付けのために用いた音声データの一覧を示す。音声振戦による基本周波数の揺らぎは，GRBAS 尺度の評価の対象とはしないことにされている [213]。しかし，非専門家群では音声振戦による揺らぎをいずれかの声質の評点に割り当てる可能性が想定されたため，定義付けのための音声データの中に，音声振戦の罹患者によって発声されたものを加えた。なお，“嗄声のサンプルテープ” [214] の音声データは，GRBAS 尺度のどの項目に対応する声質であるかに関する説明は付与されている一方で，その評点は明示されていないため，ここでは具体的な評点に関する定義付けは行わなかった。

GRBAS 尺度の評価の訓練は，既に説明した図 4.5 に示す GUI を用いて実施した。この訓練は，実際に GRBAS 尺度の具体的な評点の付け方を学ぶことを目的とした。訓練のための音声データとしては，“動画で見る音声障害 ver. 2.0” [213] 内に収録されているものを用いた。“動画で見る音声障害 ver. 2.0” [213] の各音声データには，日本音声言語医学会によって既に GRBAS 尺度の評点が付与されている。“動画で見る音声障害 ver. 2.0” 内に存在する項目 R と B のすべての組み合わせについて，それぞれ最も典型的であると考えられる音声データを，アンカーを選定した耳鼻咽喉科医が選定した。また，その際，“動画で見る音声障害 ver. 2.0” 内に含まれる項目 A と S のすべての評点が含まれるようにした。表 4.5 に，実際に選定した訓練用の音声データを示す。訓練用の音声データは，ランダムな順序で非専門家群の実験参加者に提示し，各音声データに対して GRBAS 尺度の評価を行うことを指示した。各音声データの評価が終わる度，正解となる GRBAS 尺度の評点を表示し，音声データを再び再生し，どのような音声にどのような評点が対応するか確認するよう促した。なお，正解となる GRBAS 尺度の評点としては，“動画で見る音声障害 ver. 2.0” [213] で付与されている日本音声言語医学会による評点を用いた。

表 4.4 非専門家に対する GRBAS 尺度の定義付けのために使用した音声データの一覧。なお，“嗄声のサンプルテープ” [214] の音声データは，GRBAS 尺度のどの項目に対応する声質であるかに関する説明は付与されている一方で，その評点は明示されていない。

声質の特徴	出典
R	“嗄声のサンプルテープ” [214]
R	“嗄声のサンプルテープ” [214]
B，A	“嗄声のサンプルテープ” [214]
B	“嗄声のサンプルテープ” [214]
A	“嗄声のサンプルテープ” [214]
S	“嗄声のサンプルテープ” [214]
R，B	“嗄声のサンプルテープ” [214]
R，S	“嗄声のサンプルテープ” [214]
B，A	“嗄声のサンプルテープ” [214]
B，S	“嗄声のサンプルテープ” [214]
R，B，S	“嗄声のサンプルテープ” [214]
R，B，S	“嗄声のサンプルテープ” [214]
音声振戦	“動画で見る音声障害 ver. 2.0” [213] の収録番号 054 の音声データ

表 4.5 非専門家に対する GRBAS 尺度の訓練のために使用した音声データと GRBAS 尺度の評点の一覧。音声データとしては，“動画で見る音声障害 ver. 2.0” [213] 内に収録されたものを用い，GRBAS 尺度の評点としては，“動画で見る音声障害 ver. 2.0” [213] 内で各音声データに付与されたものを用いた。

	評点				
収録番号	G	R	B	A	S
062	0	0	0	0	0
031	1	1	0	0	1
013	2	1	1	0	1
020	2	1	2	0	0
037	2	2	0	0	1
039	2	2	1	0	0
003	3	2	2	1	0
048	3	2	3	0	2
024	3	3	0	0	0
058	3	3	1	1	0
065	3	3	2	0	3

4.5.3　結果

本節では，第 7 章の機械学習で使用した病的音声データセットの正解ラベルに関連する，GRBAS 尺度の評価に関する結果について報告する。

GRBAS 尺度の評価の評者内信頼性を定量化する指標としては，2 名の評者間信頼性を表す二次重み付け Cohen のカッパ係数 [215–217] を用いた。2 名の評者の代わりに，ある評者による 1 回目と 2 回目の評価を用いることで，その評者の評者内信頼性を定量化することができる。Cohen のカッパ係数の値は大きいほど信頼性が高いことを表す。2 名の評者による評価が完全に一致する場合には，値が最大値の 1 となるのに対し，評価が完全にランダムなときよりも信頼性が低い場合には，値が 0 を下回ることもある。

GRBAS 尺度の評価の評者間信頼性としては，「各実験参加者と他の実験参加者の間の信頼性」と，「各群内の評者間信頼性」を調べた。前者は，各実験参加者が，（群の相違を問わずに）他の実験参加者とどれくらい近い評価基準を持っているかどうかを表す。この評者間信頼性は，二次重み付け Cohen のカッパ係数を用いて定量化した。その一方で，後者は，同じ群に属する評者同士が，どれくらい近い評価基準を持っているかどうかを表す。各群内の評者間信頼性は，3 名以上の評者間信頼性を定量化できるように二次重み付け Cohen のカッパ係数を一般化した，二次重み付け Conger のカッパ係数 [217, 218] を用いて定量化した。

GRBAS 尺度の評価の評者内信頼性

GRBAS 尺度の評価に関して，図 4.6 に，各実験参加者の評者内信頼性に対応する二次重み付け Cohen のカッパ係数 κ の結果を，図 4.7 に，実験参加者の評者内信頼性に対応する二次重み付け Cohen のカッパ係数 κ を各群に関して平均化した結果を示す。

まず，項目 G，R，B の結果に着目すると，経験年数が 2 年以上の専門家群，経験年数が 2 年未満の専門家群，非専門家群の順で κ が大きかった。その一方で，各群内の変動も大きく，項目 R に関して，経験年数が 2 年以上の専門家群内で κ の開きが最大で 0.33，項目 B に関して，非専門家群内で κ の開きが最大で 0.50 になった。また，経験年数が 2 年未満の専門家群や非専門家群の中にも，実験参加者全体で考えて κ が大きい実験参加者が存在した。特に，非専門家群に属する評者番号が 13 の実験参加者は，項目 G で 2 番目，項目 B で 1 番目に κ が大きかった。

続いて, 項目 A と S の結果に着目すると, 項目 G, R, B の結果と比べて, κ の群間変動や群内変動が大きかった。群間に平均値に関する差異があるかを判定する Kruskal–Wallis 検定を行ったところ，結果は表 4.6 のようになった。検定の結果，GRBAS 尺度のすべての項目で，P 値は有意水準 5% は下回らなかったものの，項目 A と S で P 値は有意水準

表 4.6 評者内信頼性に対応する二次重み付け Cohen のカッパ係数について，経験年数が 2 年以上の専門家群，経験年数が 2 年未満の専門家群，非専門家群の 3 群を Kruskal–Wallis 検定で群間比較した結果。

	G	R	B	A	S
H	1.659	2.498	1.137	5.873	5.213
P 値	0.436	0.287	0.566	0.053	0.074

10% を下回り，検定統計量 H が他の項目より大きかった。項目 A に関して，経験年数が 2 年未満の専門家群の評者の κ は総じて，他の群の評者の κ より小さい傾向があり，特に評者番号が 12 の評者では，κ が負の値になった。項目 S の κ の群内標準偏差に関して，経験年数が 2 年以上の専門家群では 0.080 であったのに対し，経験年数が 2 年未満の専門家群では 0.179，非専門家群では 0.298 となった。このことは，項目 S に関して，経験年数が 2 年未満の専門家群と非専門家群では評者の個人差が大きかったことを意味した。

GRBAS 尺度の評価の評者間信頼性：各実験参加者と他の実験参加者の間

GRBAS 尺度の評価に関して，図 4.8 に，各実験参加者と他の評者との間の評者間信頼性に対応する二次重み付け Cohen のカッパ係数 κ の結果を，図 4.9 に，各実験参加者と他の評者との間の評者間信頼性に対応する二次重み付け Cohen のカッパ係数 κ を各群に関して平均化した結果を示す。項目 G，R，B，A では，各群の κ の平均値の差異は高々 0.068 であった（図 4.9）。各実験参加者と他の実験参加者との間の評者間信頼性は，実験参加者が所属する群の違いの影響（図 4.9）よりも，個々人の違いの影響（図 4.8）が強く現れる傾向があった。特に，項目 R，B，A で最も評者間信頼性の κ が小さかった実験参加者は，経験年数が 2 年以上の専門家群に属していた（図 4.8）。また，項目 G，R，B では，経験年数が 2 年以上の専門家群の κ の平均値は，一番目に大きな値ではなかった（図 4.9）。その一方で，項目 S では，非専門家に属するすべての実験参加者の κ は，経験年数に関係なく専門家群に属するすべての実験参加者の κ よりも小さかった。

GRBAS 尺度の評価の評者間信頼性：各群内

GRBAS 尺度の評価に関して，図 4.10 に，各群内での評者間信頼性に対応する二次重み付け Conger のカッパ係数 κ の結果を示す。項目 G，R，B では，経験年数が 2 年未満の専門家群の κ の値が一番大きく，項目 R では経験年数が 2 年以上の専門家群の κ の値が一番小さかった。その一方で，項目 A と S では経験年数が 2 年以上の専門家群の κ の値が一番大きく，項目 A では経験年数が 2 年未満の専門家群，項目 S では非専門家群の κ の値が一番小さかった。

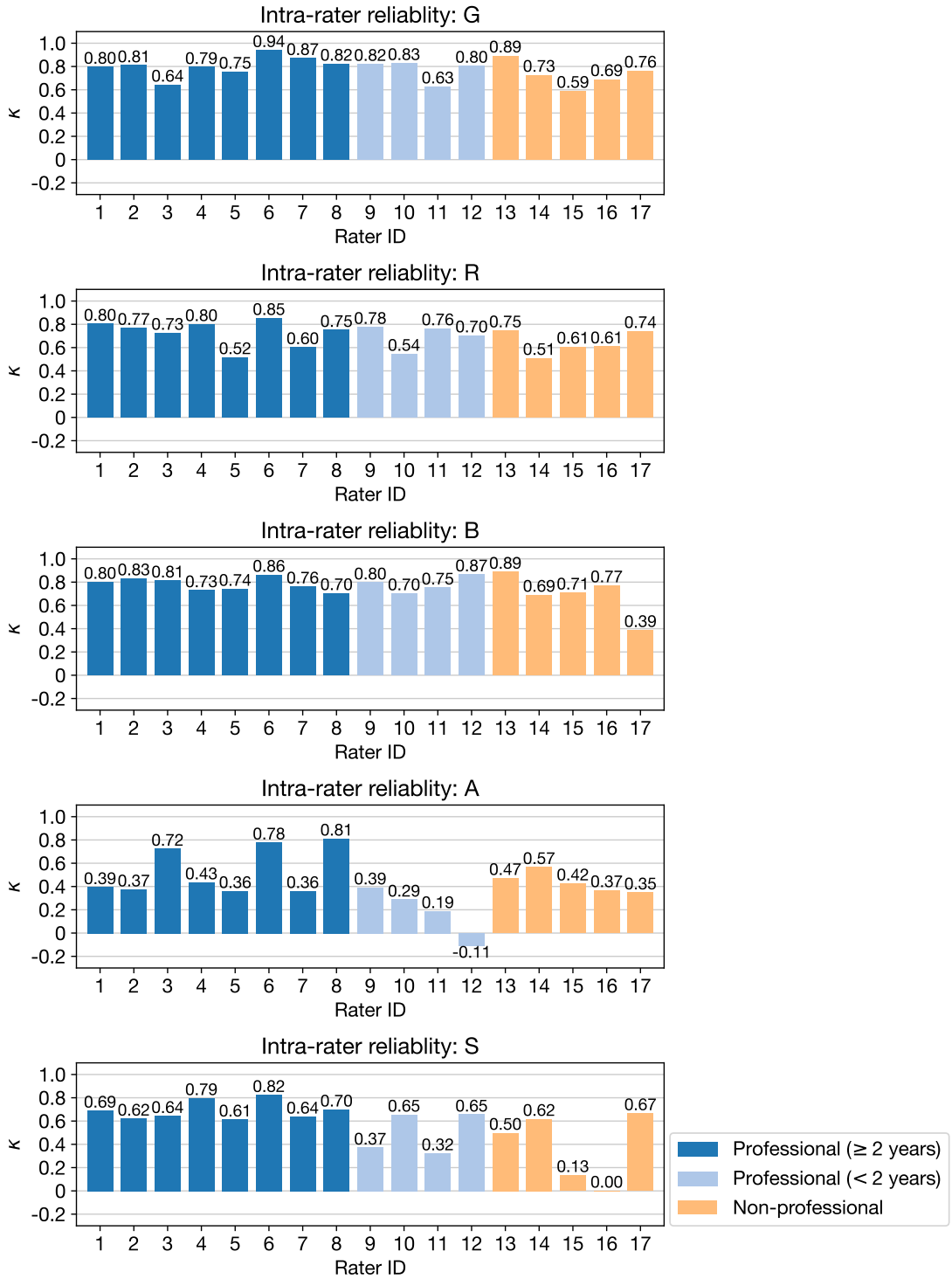


図 4.6 各実験参加者の評者内信頼性に対応する二次重み付け Cohen のカッパ係数 κ の結果の棒グラフ。κ の値の計算には，2 回評価された 50 サンプルの音声データに対する評点を用いた。

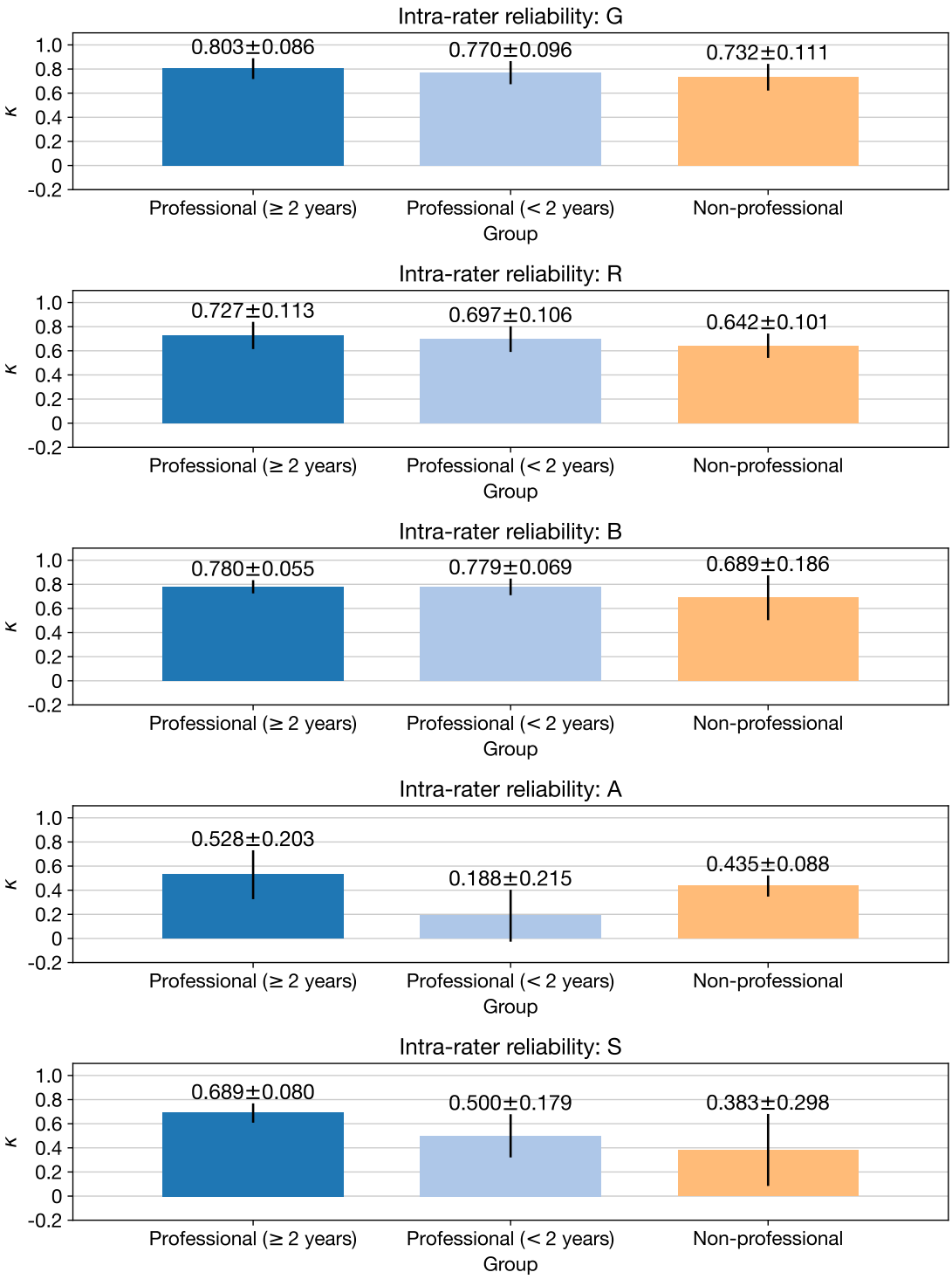


図 4.7 各実験参加者の評者内信頼性に対応する二次重み付け Cohen のカッパ係数 κ を，各群に関して平均化した結果の棒グラフ。κ の値の計算には，2 回評価された 50 サンプルの音声データに対する評点を用いた。エラーバーは標準偏差，棒の上の数値は「平均 $\pm$ 標準偏差」を表す。

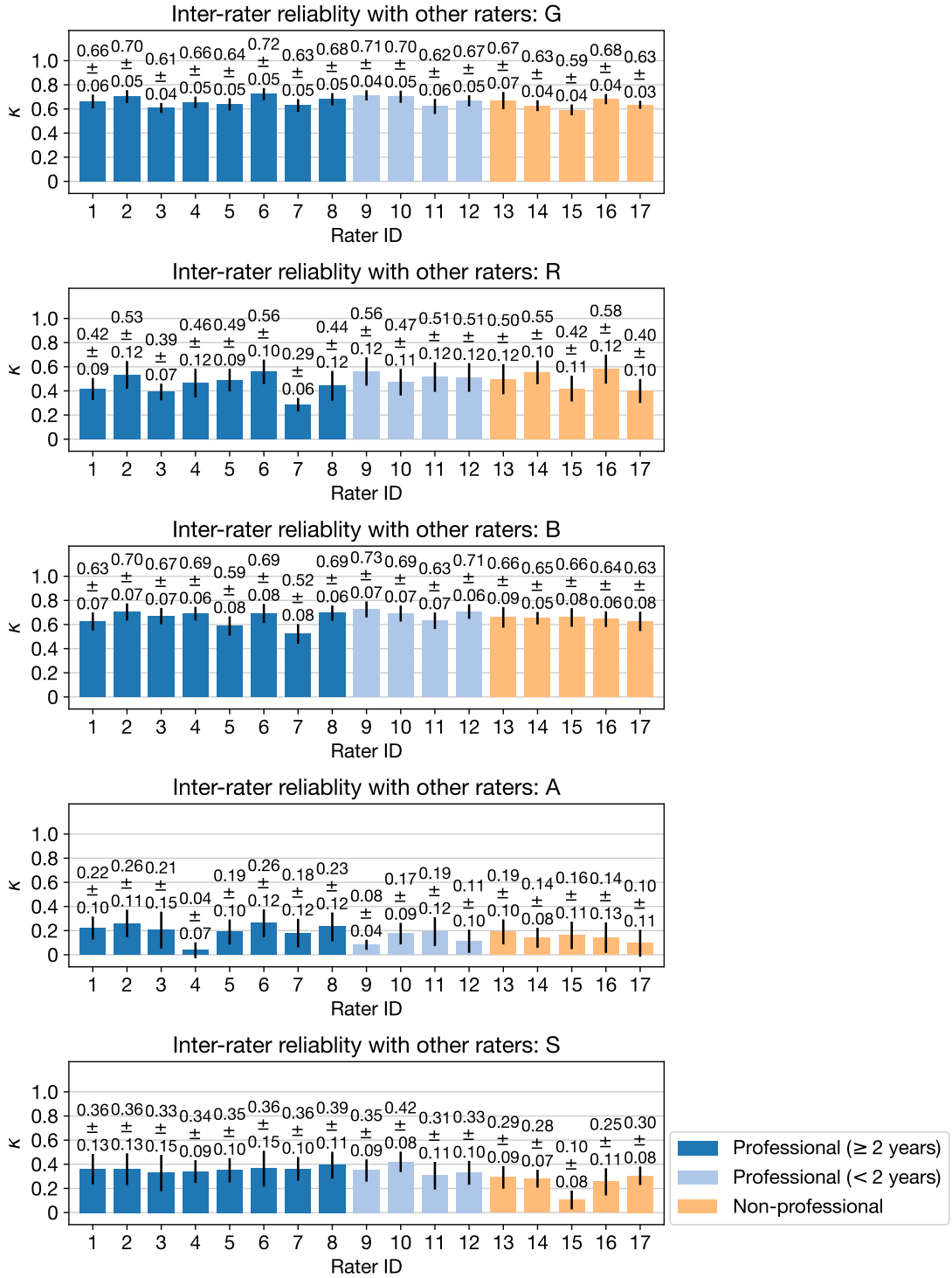


図 4.8 各実験参加者と他の実験参加者との間の評者間信頼性に対応する二次重み付け Cohen のカッパ係数 κ の結果の棒グラフ。κ の値の計算には，1 回目に与えられた 300 サンプルの音声データに対する評点を用いた。エラーバーは標準偏差，棒の上の数値は「平均 $\pm$ 標準偏差」を表す。実験参加者数を $N \in \mathbb{N}$，評者番号 k の実験参加者と評者番号 l の実験参加者の間の κ を $\kappa_{k,l}$ と記すとき，評者番号 k の実験参加者の κ の平均値と標準偏差は，評者番号 l $(\neq k)$ の実験参加者 16 名との総当たりで計算される $\left\{\kappa_{k,l} : l \in \{1, \ldots, N\},\ k \neq l\right\}$ を元に算出した。

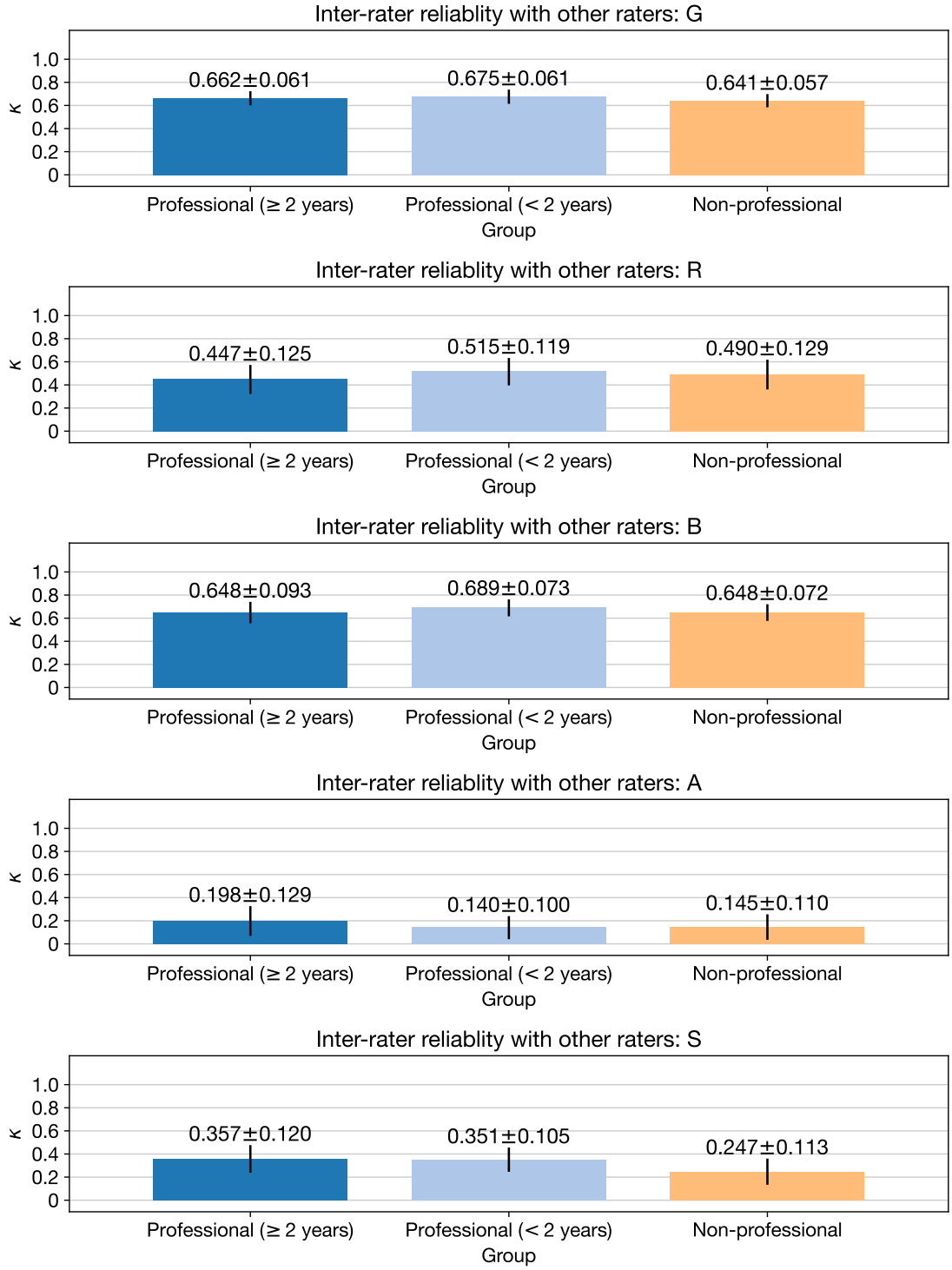


図 4.9 各実験参加者と他の評者との間の評者間信頼性に対応する二次重み付け Cohen のカッパ係数 κ の結果を，各群に関して平均化した結果の棒グラフ。κ の値の計算には，1 回目に与えられた 300 サンプルの音声データに対する評点を用いた。エラーバーは標準偏差，棒の上の数値は「平均 $\pm$ 標準偏差」を表す。実験参加者数を $N \in \mathbb{N}$，評者番号 k の実験参加者と評者番号 l の実験参加者の間の κ を $\kappa_{k,l}$ と記すとき，評者番号の集合が $\{k_n\}_{n=1}^{M}$ $(M \in \mathbb{N})$ である群の評者間信頼性の κ の平均値と標準偏差は，総当たりで計算される $\left\{\kappa_{k_n,l} : n \in \{1,\ldots,M\},\ l \in \{1,\ldots,N\},\ k_n \neq l\right\}$ を元に算出した。

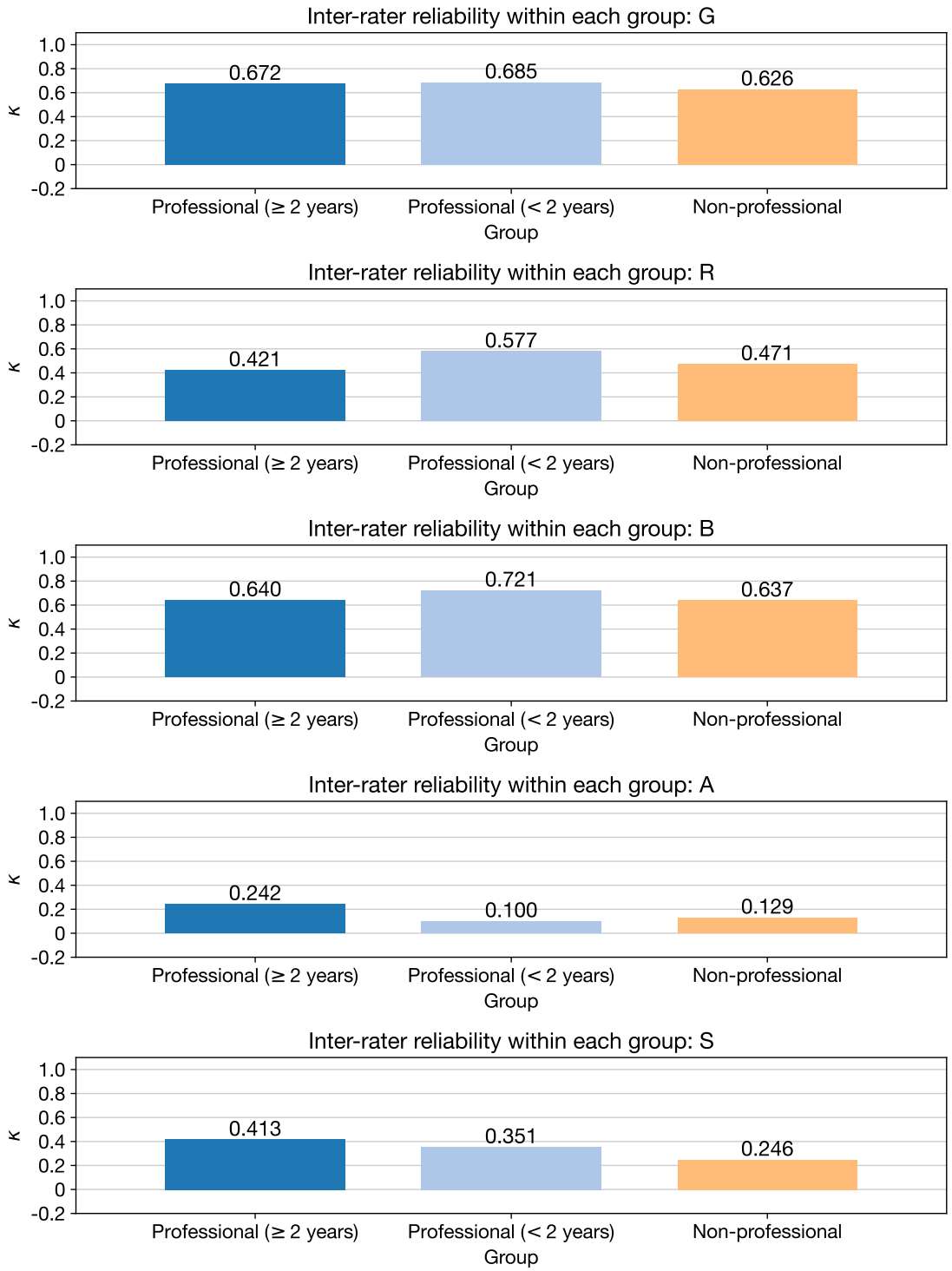


図 4.10 各群内の評者間信頼性に対応する二次重み付け Conger のカッパ係数 κ の結果の棒グラフ。

4.5.4　考察

評者内信頼性に関しては，GRBAS 尺度のすべての項目で，経験年数が 2 年以上の専門家群の κ が他の群の κ よりも大きい傾向が観察された。その一方で，経験年数が 2 年未満の専門家群や非専門家群の中には，評者内信頼性が低い実験参加者も存在すれば，評者内信頼性が高い実験参加者も存在した。特に，評者番号が 13 の実験参加者は，項目 G で 2 番目，項目 B で 1 番目に κ が大きかった。評者番号が 13 の実験参加者は，過去に聴能形成を受講した経験があり，聴能形成の受講によって，声質評価に関わる感性が先見的に形成されていた可能性があると考えられる。経験が浅い専門家群や非専門家群では評者内信頼性の高さは個人差が大きい一方で，経験を積んだ専門家群内には評者内信頼性が著しく低い評者が比較的少なかった。これらの結果を総合すると，臨床経験に依存しない評者内信頼性の初期値は個人差が大きいものの，臨床経験を積むことによって評者内信頼性が底上げされるということが示唆された。なお，De Bodt ら [36] による実験でも，経験年数が少ない専門家群よりも，経験年数が多い専門家群の方が，GRBAS 尺度の評者内信頼性が高かったことが報告されている。また，経験年数が 2 年以上の専門家群の κ と他の群の κ の差は，項目 G，R，B よりも，項目 A と S で大きかった。この結果は，項目 R と B ではアンカーが声質評価の外的基準として実験参加者の評者内信頼性向上に寄与した可能性と，項目 A と S に対する内的基準を形成するには臨床経験が重要である可能性を示唆していると考えられる。なお，非専門家群に関しては，事前訓練で 11 サンプルについて項目 A と S の具体的な評点の例を学んだ後は，アンカーを含めて項目 A と S の評点の具体的な付け方を学ぶ機会が無かったため，そもそも項目 A と S に対する内的基準を形成する経験が不足していたと考えられる。

各実験参加者と他の実験参加者の間の評者間信頼性に関して，項目 G，R，B，A では，群間の違いよりも，個々人の違いが強く現れる傾向があった。項目 G，R，B では経験年数が 2 年以上の専門家群の κ の平均値は，一番目に大きな値ではなく，項目 R，B，A で最も評者間信頼性の κ が小さかった実験参加者は，経験年数が 2 年以上の専門家群に属していた。Kreiman らによって，専門家は，それまでに受けた訓練や臨床経験の内容に基づいて，病的音声の声質評価に関する独自の内的基準を形成するために，必ずしも専門家同士の評価が類似するとは限らないことが指摘されている [33, 34]。本実験で，経験年数が 2 年以上の専門家群の評者間信頼性の κ が必ずしも高くなかったのは，Kreiman らが指摘したものと同じ原因である可能性がある。その一方で，項目 S では，非専門家に属するすべての実験参加者の κ は，経験年数に関係なく専門家群に属するすべての実験参加者の κ よりも小さかった。このことは，項目 S に関しては，専門的知識が身に付いていない状態の方が内的基準の個人差が大きく，専門的知識が身につくことによって個人差が小さ

くなる可能性があることを示唆している。

各群内の評者間信頼性に関しては，項目 G，R，B では，経験年数が 2 年未満の専門家群の κ が一番大きく，項目 R では経験年数が 2 年以上の専門家群の κ が一番小さかった。先述した内容と同様であるが，個々の専門家が異なる臨床経験を積むことによって，経験年数が 2 年以上の専門家群ではかえって評者間信頼性が下がった可能性がある。項目 A では，経験年数が 2 年以上の専門家群の κ が最も大きかった一方で，経験年数が 2 年未満の専門家群の κ が最も小さかった。このことは，臨床経験を積むことによって項目 A の内的基準の個人差は小さくなる一方で，臨床経験が少ない場合は，専門的知識を有していても項目 A の内的基準の個人差は大きい可能性を示唆している。項目 S では，非専門家群の κ が最も小さかった。先述の内容と重複するが，このことは，専門的知識を持つことによって項目 S の評価基準の個人差が小さくなる可能性を示唆している。

評者内信頼性，各実験参加者と他の実験参加者の間の評者間信頼性，各群内の評者間信頼性のすべてに関して，項目 A と S の κ は他の項目の κ よりも小さかった。いくつかの研究によって，項目 A と S は他の項目と比べて信頼性が低くなることが報告されており [36, 39]，GRBAS 尺度に影響を受けて考案された RBH 尺度（the RBH scale） [40] では項目 A と S が，CAPE-V [41] では項目 A が定量評価項目から除外されている。本実験で得られた結果は，これらの先行研究の結果を支持するものであると考えられる。

4.5.5　第 7 章の GRBAS 尺度の評点推定実験のために用いる評点の選定

GRBAS 尺度の評点推定実験のために用いるデータセットの正解ラベルとしては，何らかの意味で臨床的価値が高い評点を採用することが望ましいと考えられる。例えば，声質評価の内的基準が確立していない（評者内信頼性が低い）評者による評点や，実際の臨床で付与される評点と大きく異なる評点は，臨床的価値が低いため正解ラベルとして採用すべきでないと考えられる。今回の聴覚実験では，経験年数が 2 年以上の専門家群は，平均値に関して，GRBAS 尺度のすべての項目で評者内信頼性に対応する κ が最も大きく，項目 A と S で評者間信頼性に対応する κ が最も大きくなった。項目 G，R，B では評者間信頼性に対応する κ が最も大きくはならなかったが，臨床経験を積むことによって内的基準が独自化したことが原因の可能性があり，必ずしも臨床的価値が低いわけではないと考えられた。その一方で，経験年数が 2 年未満の専門家群と非専門家群に関しては，項目 A と S で評者内信頼性に対応する κ が著しく小さい実験参加者が存在し，評者間信頼性に対応する κ は経験年数が 2 年以上の専門家群より小さい傾向が見られた。以上を踏まえて，第 7 章で用いる GRBAS 尺度の評点推定のデータセットには，臨床的価値が比較的高いと考えられる，経験年数が 2 年以上の専門家群によって付与された評点を採用した。

4.6 まとめ

本章では，専門家によって GRBAS 尺度の評点が付与された病的音声データベースを構築した。収集したすべての音声データに対して 1 名の耳鼻咽喉科医による GRBAS 尺度の評点付与を行った後，音声データの多様性が高くなるように注意を払いつつ 300 サンプルの音声データを選別し，選別した音声データに対して 17 名の評者による GRBAS 尺度の評点付与を行った。17 名の評者によって与えられた評点に関しては，評者内信頼性と評者間信頼性の分析を行った。最終的に，臨床的価値が比較的高いと考えられる，経験年数が 2 年以上の 8 名の専門家（耳鼻咽喉科医 6 名，言語聴覚士 2 名）による評点を，GRBAS 尺度の評点推定実験で使うために採用した。本章で構築した病的音声データベースは，次章以降の GRBAS 尺度の評点推定実験で使用した。

第 5 章

瞬時周波数を用いた GRBAS 尺度の自動評価

5.1 はじめに

健常音声と比較して，病的音声は，非定常性や雑音成分が強いこと，逆に言えば正弦波成分が弱いことで特徴づけられることが多い。そのため，慣習的な音響特徴量の多くは，非定常性や雑音成分を定量化することを意図して設計されている [44, 46, 47, 219]。本研究で着目している位相情報に関して言えば，短時間フーリエ変換の位相は，正弦波成分が存在する周波数帯域で特徴的なパターンを表す性質がある（例えば，図 2.1 (C)，(D) を参照)。Bayram と Kamasak [220] や，Yatabe と Oikawa [221] は，この性質を応用した雑音除去（denoising）手法を提案している。

本章では，聴覚心理的評価の評点の自動推定における，スペクトルの位相情報の有効性を調べることを目的として実施した実験について述べる。具体的には，1 名の耳鼻咽喉科医によって評価された 2,545 サンプルの病的音声データを元に，振幅および位相に基づく時間周波数表現と RNN を用いて，GRBAS 尺度の評点の自動推定実験を行った。位相に基づく時間周波数表現としては，Bayram と Kamasak の研究 [220] と，Yatabe と Oikawa [221] の研究を参考にした，位相の時間方向の変化に着目した表現を使用した。

5.2 手法

5.2.1 スペクトログラム

第 2.2.1 節でも述べたように，スペクトログラムという言葉は，広義では，短時間フーリエ変換に基づいて計算される時間周波数表現のことを表す。本章では，スペクトログ

ラムという言葉を，「瞬時位相修正短時間フーリエ変換（instantaneous phase corrected STFT; iPC-STFT）」 [125, 135, 221] に基づいて計算される時間周波数表現を表すために用いる。瞬時位相修正（instantaneous phase correction）は位相の直観的な解釈や技術的応用を促進する手法であり，瞬時位相修正短時間フーリエ変換は瞬時位相修正と短時間フーリエ変換を組み合わせたものである。本章では，瞬時位相修正短時間フーリエ変換に基づいて計算されるスペクトログラムの振幅と位相について考える。また，周波数軸のメル尺度への変換に関しても，振幅と位相の両方について検討を行う。

瞬時位相修正

瞬時位相修正 [125, 135, 221] は，正弦波の短時間フーリエ変換の位相に表れる性質に着目した手法である。$\boldsymbol{w} \in \mathbb{R}^W (W \in \mathbb{N})$ を窓関数とする離散信号 $\boldsymbol{x} \in \mathbb{R}^X (X \in \mathbb{N})$ の短時間フーリエ変換 $\mathcal{F}^{(\boldsymbol{w},a,D)}\boldsymbol{x} \in \mathbb{C}^{M\times N} (M, N \in \mathbb{N})$ は，以下のように定義される（式 (2.11) の再掲）。

$$\left(\mathcal{F}^{(\boldsymbol{w},a,D)}\boldsymbol{x}\right)[m,n] := \sum_{l=an-\lfloor W/2\rfloor}^{an+\lfloor (W-1)/2\rfloor} \boldsymbol{x}[l]\boldsymbol{w}[l-an]\exp\left(-\frac{2\pi iml}{D}\right) \tag{5.1}$$

ここで，窓関数の添字集合は $\{-\lfloor W/2\rfloor, -\lfloor W/2\rfloor+1, \ldots, -1, 0, 1, \ldots, \lfloor (W-1)/2\rfloor\}$ で与えられ，$i = \sqrt{-1}$，$m \in \{0, 1, \ldots, M-1\}$ は周波数の添字，$n \in \{0, 1, \ldots, N-1\}$ は時間の添字，$a \in \mathbb{N}$ はシフト長，$D \in \mathbb{N}$ $(D \geq W)$ は DFT 点数を表す。離散信号が，$\boldsymbol{x}[n] = \exp(2\pi i\xi n/D) (\xi \in \mathbb{R})$ と表される複素正弦波であるとき，$\xi = m$ を満たす周波数添字 m が存在するならば，次の関係が成り立つ。

$$\left(\mathcal{F}^{(\boldsymbol{w},a,D)}\boldsymbol{x}\right)[\xi, n+1] = \left(\mathcal{F}^{(\boldsymbol{w},a,D)}\boldsymbol{x}\right)[\xi, n] = \sum_{l=-\lfloor W/2\rfloor}^{\lfloor (W-1)/2\rfloor} \boldsymbol{w}[l] \tag{5.2}$$

しかし，$\xi = m$ を満たす周波数添字 m が存在しない場合，ξ を何らかの周波数添字に割り当てることはできない。任意の周波数添字 m に対しては，次の関係が成り立つ。

$$\left(\mathcal{F}^{(\boldsymbol{w},a,D)}\boldsymbol{x}\right)[m, n+1]\exp\left(-\frac{2\pi ia\delta}{D}\right) = \left(\mathcal{F}^{(\boldsymbol{w},a,D)}\boldsymbol{x}\right)[m, n] \tag{5.3}$$

ここで，$\delta = \xi - m$ は複素正弦波の周波数と周波数添字の偏差を表す。上式より，短時間フーリエ変換の（適切にアンラップされた）位相 $\phi \in \mathbb{R}^{M\times N}$ に関しては次の関係が成り立つ。

$$\phi[m, n+1] = \phi[m, n] + \frac{2\pi a\delta}{D} \tag{5.4}$$

偏差 δ は, 周波数ビン m を基準に考えたときの, 複素正弦波 $\boldsymbol{x}[n] = \exp(2\pi i\xi n/D) (\xi \in \mathbb{R})$ の相対瞬時周波数に対応する。このことから，正弦波に関しては，あるフレームの位相

と相対瞬時周波数がわかれば，その次のフレームの位相が予測可能であることが示唆される。

瞬時位相修正は，正弦波の周波数と，周波数添字に対応する周波数の間の不一致の低減を目的とした処理である [135]。この目的のために，瞬時位相修正では相対瞬時周波数が利用される。$\boldsymbol{w} \in \mathbb{R}^W (W \in \mathbb{N})$ を窓関数とする離散信号 $\boldsymbol{x} \in \mathbb{R}^X (X \in \mathbb{N})$ の瞬時位相修正短時間フーリエ変換 $\mathcal{F}_{\mathrm{iPC}}^{(\boldsymbol{w},a,D)}\boldsymbol{x} \in \mathbb{C}^{M\times N} (M, N \in \mathbb{N})$ は，次のように定義される。

$$\left(\mathcal{F}_{\mathrm{iPC}}^{(\boldsymbol{w},a,D)}\boldsymbol{x}\right)[m,n] := \boldsymbol{A}[m,n]\left(\mathcal{F}^{(\boldsymbol{w},a,D)}\boldsymbol{x}\right)[m,n] \tag{5.5}$$

ここで，$\boldsymbol{A} \in \mathbb{C}^{M\times N}$ は，次のように定義される瞬時位相修正行列である。

$$\boldsymbol{A}[m,n] = \prod_{l=0}^{n-1} \exp\left(-\frac{2\pi i a \boldsymbol{I}[m,l]}{D}\right) \tag{5.6}$$

但し，すべての m に対して $\boldsymbol{A}[m,0] = 1$ であり，$\boldsymbol{I} \in \mathbb{R}^{M\times N}$ は相対瞬時周波数である。本節の冒頭でも述べたが，本章では，瞬時位相修正短時間フーリエ変換 $\mathcal{F}_{\mathrm{iPC}}^{(\boldsymbol{w},a,D)}\boldsymbol{x}$ に基づいて計算される時間周波数表現をスペクトログラムと呼ぶ。

図 5.1 は，背景雑音を含む 3 つの正弦波から構成される混合信号に対するスペクトログラムのヒートマップを表す。正弦波成分を含む周波数ビンでは，短時間フーリエ変換の位相は時間方向に回転しているのに対し，瞬時位相修正短時間フーリエ変換の位相は時間方向にほぼ一定になっている。換言すれば，瞬時位相修正短時間フーリエ変換に関しては，正弦波が支配的な周波数ビンにおいては，隣接フレーム間の位相変化がほぼ 0 になっている。その一方で，正弦波が支配的ではない周波数ビンにおいては，隣接フレーム間の位相変化はランダムに変化している様子が見て取れる。つまり，瞬時位相修正には，正弦波成分の存在を際立たせる効果がある。この効果のために，瞬時位相修正は，音声強調 [221] や調波打撃音分離 [125] のような，正弦波成分が重要となる課題に応用されている。

病的音声は，健常音声と比較して，非定常性や雑音成分の強さなどによって特徴づけられる。瞬時位相修正短時間フーリエ変換は，正弦波成分と雑音成分の違いを際立たせるような手法であり，その位相の時間変動が，非定常性等の声質に関する特徴を含んでいることが期待される。これが，本実験において，通常の短時間フーリエ変換ではなく，瞬時位相修正短時間フーリエ変換を採用した理由である。

入力特徴量

本章の実験では，対数振幅スペクトログラム（POW），位相スペクトログラムの時間差分の絶対値（DiffPhase），複素スペクトログラムの時間差分の絶対値（DiffComplex）という 3 種類のスペクトログラムを入力特徴量として検討する。複素スペクトログラム $\boldsymbol{X} = \mathcal{F}_{\mathrm{iPC}}^{(\boldsymbol{w},a,D)} \in \mathbb{C}^{M\times N} (M, N \in \mathbb{N})$ が与えられ，$\boldsymbol{X}[m,n] = \boldsymbol{A}[m,n]e^{i\phi[m,n]}$ のような

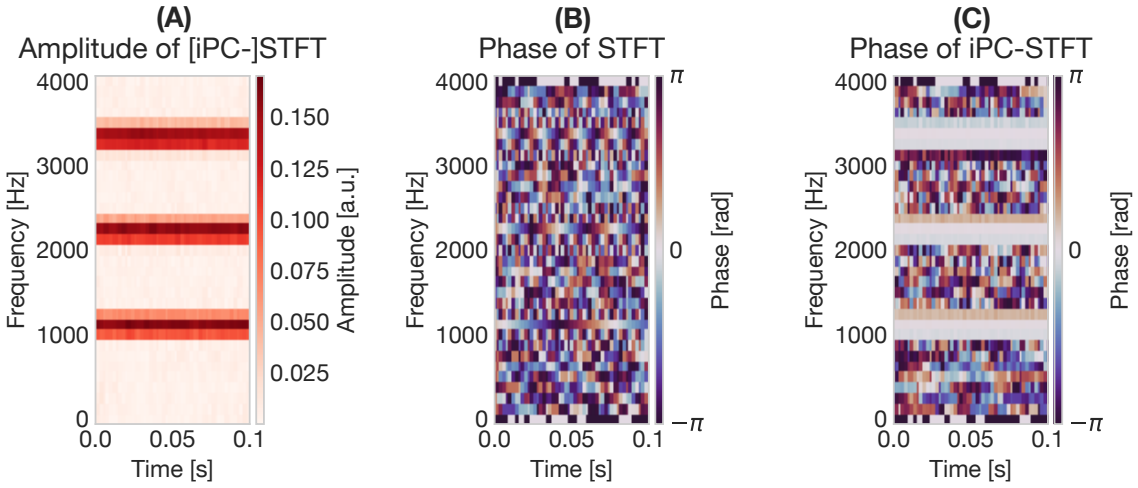


図 5.1 背景雑音を含む，3 つの正弦波から構成される混合信号に対する (A)（瞬時位相修正）短時間フーリエ変換の振幅スペクトログラム，(B) 短時間フーリエ変換の位相，(C) 瞬時位相修正短時間フーリエ変換の位相のヒートマップ。短時間フーリエ変換の振幅と，瞬時位相修正短時間フーリエ変換の振幅は同一であることに注意。瞬時位相修正短時間フーリエ変換に関して，瞬時位相修正によって，正弦波の影響が強い周波数ビンにおける位相の時間変化が小さくなっていることが観察される。標本化周波数は 8 kHz，窓長 W と DFT 点数 D はどちらも 8 ms (64 sample)，シフト長 a は 2 ms (16 sample) とし，窓関数としては Hann 窓を用いた。

極形式で振幅 $\boldsymbol{A} \in \mathbb{R}_{\geq 0}^{M \times N}$ と位相 $\phi \in (-\pi, \pi]^{M \times N}$ が表されるとき，各特徴量は，次のように定義される。

$$\mathrm{POW}^{(\boldsymbol{w},a,D)}(\boldsymbol{x}) = 10 \log_{10} |\boldsymbol{A}|^2 \tag{5.7}$$

$$\mathrm{DiffPhase}^{(\boldsymbol{w},a,D)}(\boldsymbol{x}) = |\mathcal{D}\mathcal{U}\phi| \tag{5.8}$$

$$\mathrm{DiffComplex}^{(\boldsymbol{w},a,D)}(\boldsymbol{x}) = 10 \log_{10} |\mathcal{D}\boldsymbol{X}|^2 \tag{5.9}$$

ここで，$\big(\mathcal{D}\boldsymbol{X}\big)[m,n] = \boldsymbol{X}[m,n] - \boldsymbol{X}[m,n-1]$ は時間方向の差分計算，$\mathcal{U}$ は位相の時間方向のアンラップ処理を表し，対数関数 $\log_{10}(\cdot)$ や絶対値関数 $|\cdot|$ は，行列の各要素に対して作用すると考える。POW（対数振幅スペクトログラム）は値の大きさがエネルギーの強さに対応しており，直観的に解釈をすることができる。DiffPhase は，正弦波が支配的な領域ではほとんど 0 に近い値をとる一方で，そうでない領域ではランダムなパターンを呈する。DiffPhase には，非定常性のような病的音声の声質に関わる特徴が含まれていることが期待される。DiffComplex は，他の二つの特徴量とは異なり，振幅と位相の両方を同時に考慮に入れる。DiffComplex は，DiffComplex の l^1 ノルムとして定義される位相修正総変動（phase corrected total variation） [220, 221] と呼ばれる特徴量の計算過程で表れる。DiffComplex ないし位相修正総変動は，雑音や変動成分の存在を示すこと

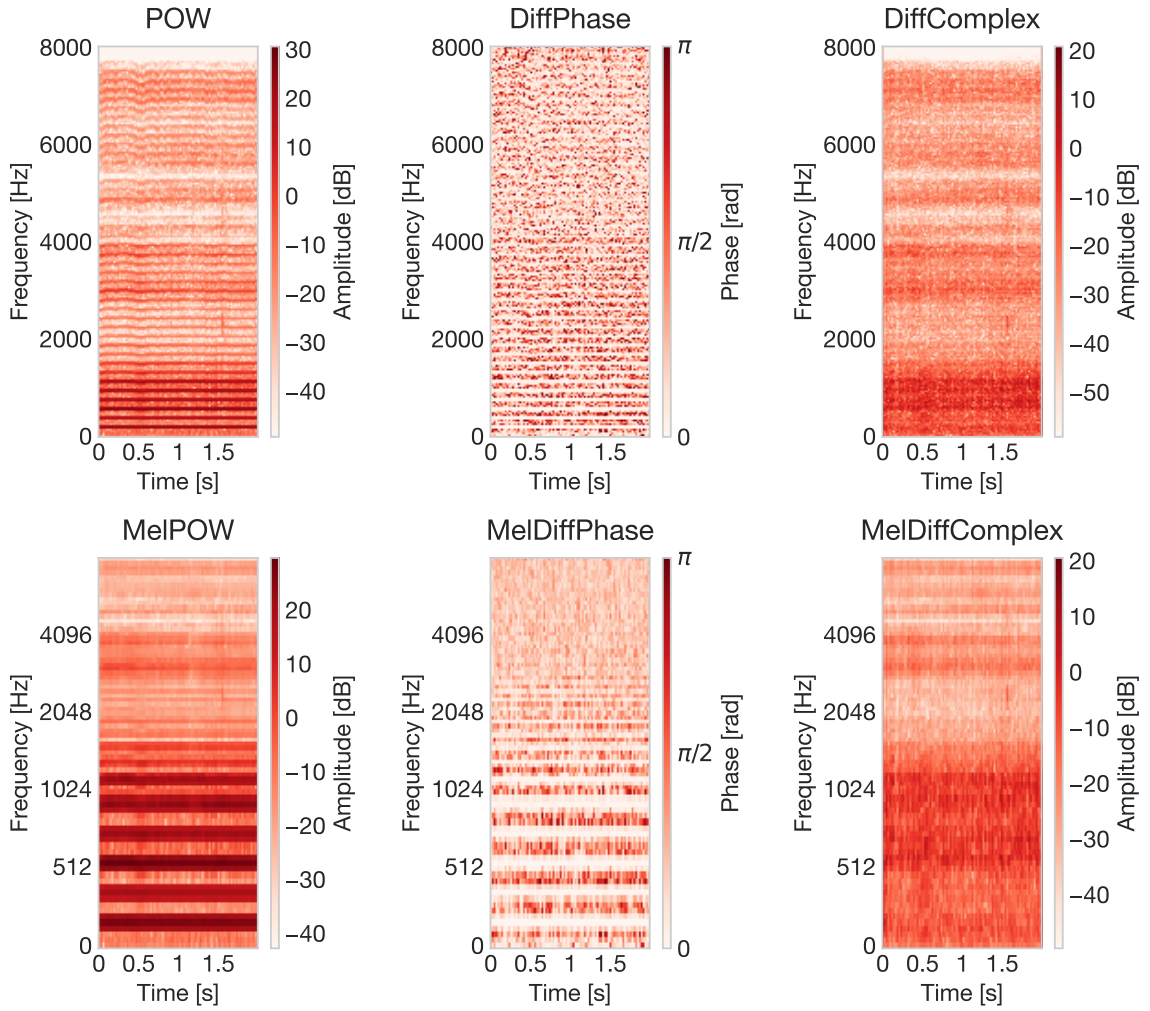


図 5.2 持続母音/a/の音声波形から計算される入力特徴量のヒートマップ。ヒートマップの計算には，実験と同じパラメータを用いた。

が期待される。DiffPhase は正弦波成分に対する非正弦波成分の「相対的強度」として，DiffComplex は非正弦波成分の「絶対的強度」として解釈することができる。

上述の特徴量に加えて，周波数軸をメル尺度に変換した特徴量も検討に入れた。病的音声の聴覚心理的評価自体がヒトの聴覚に基づくものであるため，ヒトの聴覚特性を反映したメル尺度は効率的な次元圧縮に有用であると期待される。メルフィルタバンク $\boldsymbol{M} \in \mathbb{R}_{\geq 0}^{K \times M} (K \in \mathbb{N})$ が与えられたとき，周波数軸をメル尺度に変換した特徴量 MelPOW，MelDiffPhase，MelDiffComplex は次のように定義される。

$$\mathrm{MelPOW}^{(\boldsymbol{w},a,D)}(\boldsymbol{x}) = 10 \log_{10} \left(\boldsymbol{M} \left(|\boldsymbol{A}|^2 \right) \right) \tag{5.10}$$

$$\mathrm{MelDiffPhase}^{(\boldsymbol{w},a,D)}(\boldsymbol{x}) = \boldsymbol{M} |\mathcal{D}\mathcal{U}\phi| \tag{5.11}$$

$$\mathrm{MelDiffComplex}^{(\boldsymbol{w},a,D)}(\boldsymbol{x}) = 10 \log_{10} \left(\boldsymbol{M} \left(|\mathcal{D}\boldsymbol{X}|^2 \right) \right) \tag{5.12}$$

表 5.1 DNN の構造。(A) は全体の構成，(B) は (A) の中に表れる「隠れブロック」の構成を表す。データは，順伝播では上から下へ，逆伝播では下から上へと流れる。括弧内の値は出力次元を表す。$F \in \mathbb{N}$ は入力特徴量の次元，$P, Q \in \mathbb{N}$ は調整されるハイパーパラメータである。なお，双方向 GRU の出力系列は，時間方向に平均化された後で後続層に伝播される。DNN の最後の出力次元は，4 クラス分類問題として GRBAS 尺度の評点 0，1，2，3 の確率を出力するために 4 次元となっている。

(A) DNN
1 次元 Batch Normalization (F)
全結合層 (P)
双方向 GRU ($2P$)
Dropout ($2P$)
Q 層の隠れブロック ($2P$)
全結合層 (4)
softmax 関数 (4)

(B) 隠れブロック
全結合層 ($2P$)
ReLU ($2P$)
Dropout ($2P$)

第 2.2.6 節で説明したように，位相の「微分」は極端に大きな値を取る場合があるが，位相の「差分」を用いた DiffPhase の値が極端に大きくなることはないため，DiffPhase から MelDiffPhase への変換は問題なく行えると考えられる。

図 5.2 は，持続母音/a/の音声波形から計算された入力特徴量のヒートマップを表す。これらの特徴量は，次節で説明する DNN への入力として使用した。

5.2.2　DNN の構造

本章の実験では，GRBAS 尺度の各項目につき，4 クラス分類問題として 0，1，2，3 の評点の確率を出力する DNN を作成した。各 DNN は表 5.1 に示すように，双方向 GRU [191, 194] と全結合層（多層パーセプトロン）を用いて構成した。入力特徴量は時間長が可変であり，ミニバッチは $\{\boldsymbol{X}^{(b)} \in \mathbb{R}^{F \times N_b}\}_{b=0}^{B-1} (B, N_0, N_1, \ldots, N_{B-1} \in \mathbb{N})$ のように表された。ミニバッチは，$\boldsymbol{X} = [(\boldsymbol{X}^{(0)})(\boldsymbol{X}^{(1)}) \cdots (\boldsymbol{X}^{(B-1)})]$ のように時間方向に結合された 1 つの行列として DNN に入力し，双方向 GRU に入力する直前で $\{\boldsymbol{Y}^{(b)} \in \mathbb{R}^{H \times N_b}\}_{b=0}^{B-1} (H \in \mathbb{N})$ のように可変長行列のバッチ表現に戻した。なお，最初の 1 次元 Batch Normalization は，入力特徴量を各周波数ビンに関して正規化するために導入し，学習パラメータは初期値で固定した。また，過剰適合を抑制するために Dropout [173] と l^2 正則化を，勾配爆発を防ぐために勾配ノルムクリッピング [172] を導入した。損失関数として交差エントロピー誤差を使用し，DNN のパラメータの最適化アルゴリズムとして Adam を使用した。

5.3 実験

本章の実験では，耳鼻咽喉科医 1 名によって GRBAS 尺度の評点が付与された持続母音/a/の音声データを用いた。2,245 サンプル（女性：798 サンプル，393 名，男性：1,447 サンプル，701 名）の音声データを訓練・検証データセットとして，300 サンプル（女性：133 サンプル，133 名，男性：167 サンプル，167 名）の音声データをテストデータセットとして用いた。図 5.3 に，GRBAS 尺度の評点の分布を示す。なお，GRBAS 尺度の評点の付与や音声データの選定方法等の詳細に関しては第 4 章で述べた。

病的音声データベースの各音声に対して，第 5.2.1 節で説明した入力特徴量を計算した。耳鼻咽喉科医による評価時は 48 kHz または 50 kHz であった標本化周波数は，特徴量計算の前に 16 kHz にダウンサンプリングした。瞬時位相修正短時間フーリエ変換の窓長 W と DFT 点数 D はどちらも 32 ms (512 sample)，シフト長 a は 8 ms (128 sample) とし，窓関数としては Hann 窓を用いた。メルフィルタバンクのチャンネル数は 80 とし，メル尺度の定義としては Slaney による実装 [146] を採用した。各入力特徴量の値は，0 から 1 になるように正規化した。

表 5.1 の中の隠れ表現の次元に関するハイパーパラメータ P（128，256，512），隠れブロックの数 Q（0，1，2）に関しては，グリッドサーチを行った。学習率は 0.001，l^2 正則化の正則化定数は 0.001，勾配ノルムクリッピングの閾値は 1.0，Dropout 率は 0.5，バッチサイズ B は 128 に設定した。DNN の訓練は，15 エポック連続で最小損失が更新されなかった時点で終了した。なお，DNN の構築と評価には機械学習フレームワークである PyTorch [222] を用いた。

GRBAS 尺度の各項目について，全 6 種類の入力特徴量（POW，DiffPhase，DiffComplex，MelPOW，MelDiffPhase，MelDiffComplex）と，POW と DiffPhase の組み合わせ（POW+DiffPhase），MelPOW と MelDiffPhase の組み合わせ（MelPOW+MelDiffPhase）を比較した。

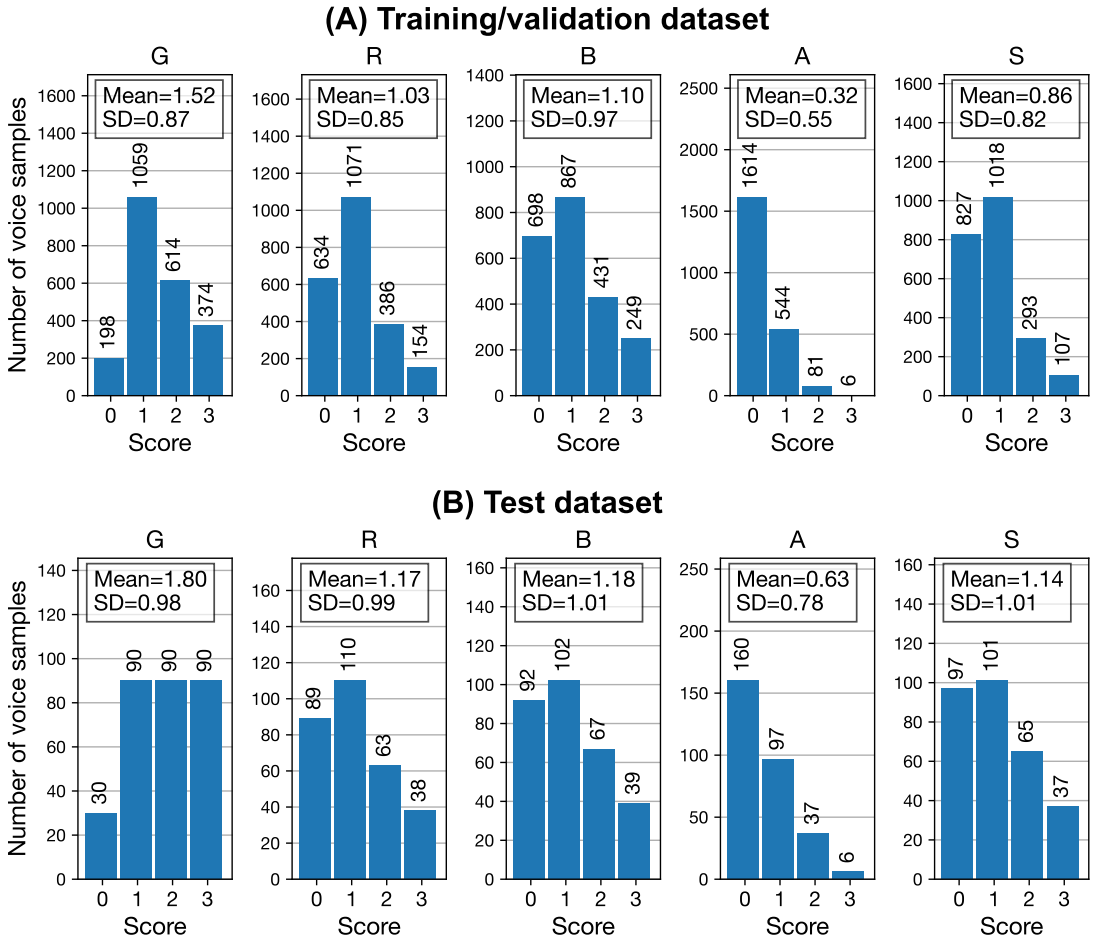


図 5.3 データセットの GRBAS 尺度の評点の分布。(A) は訓練・検証データセット，(B) はテストデータセットの分布。

表 5.2 グリッドサーチで κ が最高値を示したハイパーパラメータ。GRBAS 尺度の項目，特徴量ごとの結果を示している。P は表 5.1 の中の隠れ表現の次元に関するハイパーパラメータ，Q は表 5.1 の中の隠れブロックの数を表す。

	G		R		B		A		S	
特徴量	P	Q	P	Q	P	Q	P	Q	P	Q
POW	256	2	512	2	128	2	256	2	512	2
DiffPhase	512	2	512	2	128	0	256	0	512	1
POW+DiffPhase	512	2	512	2	512	2	128	2	512	1
DiffComplex	256	2	256	2	256	0	256	1	128	0
MelPOW	512	2	256	2	128	2	512	1	256	2
MelDiffPhase	512	1	128	2	256	2	512	1	256	2
MelPOW+MelDiffPhase	256	2	128	2	256	2	128	2	256	2
MelDiffComplex	256	2	128	2	128	2	512	1	128	0

5.4 結果と考察

5 分割交差検証を行い，偶然の一致（データセットの偏り）を考慮に入れた一致の強度を示す指標である二次重み付け Cohen のカッパ係数 κ [215–217] を用いて性能評価を行った。κ は最大値が 1 であり，値の大きさが一致の強さに対応する。表 5.2 に，グリッドサーチで得られた最適なハイパーパラメータを，図 5.4 と表 5.3 に，最適なハイパーパラメータを用いた時に得られた実験結果を示す。GRBAS 尺度のすべての項目に対して，一貫して周波数軸のメル尺度への変換は性能向上をもたらした。また，GRBAS 尺度のすべての項目に対して，MelPOW+MelDiffPhase を用いた時に κ は最大となった。その一方で，メル尺度への変換が行われた特徴量に関して，項目 G，R，B，S に対しては MelDiffComplex を用いた時に，項目 A に対しては MelDiffPhase を用いた時に κ は最小となった。

メル尺度への変換は行列による低次元空間への線形写像であるため，DNN の最初の全結合層でも学習可能な変換である。それにも関わらず，振幅情報のみならず，位相情報に対してもメル尺度への変換は有効に働いた。この結果は，メル尺度への変換は，ヒトの聴覚印象に関わる重要な情報を保持したまま効率的に次元を削減できることを示唆する。なお，周波数軸が線形尺度の特徴量の中では，項目 G，R，B に対して POW を用いた時に κ が最小となったのに対し，周波数軸がメル尺度の特徴量の中では，MelPOW を用いた時に κ が最小となることはなかった。このことから，メル尺度への変換は，特に振幅情報に関して有効に働いたと考えられる。

本章の実験で，短時間フーリエ変換の位相には，病的音声の声質評価に有用な情報が含まれていることが示された。加えて，項目 G，R，B，S に対しては，位相情報の方が振幅情報よりも効果的であった。これは，雑音や非定常性，揺らぎなどの嗄声に特有の情報を，位相の時間変動が含んでいることを示唆している。その一方で，項目 A に対してのみは，振幅情報の方がより高い性能をもたらした。このことは，無力性の特徴は振幅情報によく顕在化し，位相情報ではあまり表出しないことを示唆している。また，他の特徴量と比較して，複素数値の時間変動（DiffComplex，MelDiffComplex）は有効性が低かった。DiffComplex や MelDiffComplex は，図 5.4 に示すように，音声の重要な情報である調波構造が目立たなくなってしまうことが，有効性が低かった原因の一つではないかと考えられる。

表 5.3 5 分割交差検証で得られた二次重み付け Cohen のカッパ係数 κ の結果（図 5.4 と同じデータ）。$\pm$ の後の値は標準偏差を表す。テストデータセットは，検証損失が最小となったエポックのモデルを使って評価された。この図で示している GRBAS 尺度の各項目，各特徴量に対する κ の値は，表 5.2 のハイパーパラメータを使用した時に得られたものである。周波数軸が線形尺度の特徴量と，メル尺度の特徴量のそれぞれに関して，GRBAS 尺度の各項目に対する κ の最大値を**太字**で，κ の最小値を下線で示す。

特徴量	G	R	B	A	S
POW	$\underline{0.502 \pm 0.070}$	$\underline{0.273 \pm 0.079}$	$\underline{0.626 \pm 0.024}$	0.177 ± 0.051	0.399 ± 0.058
DiffPhase	$\mathbf{0.693 \pm 0.022}$	0.451 ± 0.070	0.683 ± 0.023	$\underline{0.041 \pm 0.053}$	0.424 ± 0.083
POW+DiffPhase	0.641 ± 0.026	$\mathbf{0.508 \pm 0.081}$	$\mathbf{0.712 \pm 0.021}$	$\mathbf{0.259 \pm 0.048}$	$\mathbf{0.487 \pm 0.063}$
DiffComplex	0.526 ± 0.028	0.461 ± 0.020	0.652 ± 0.023	0.147 ± 0.047	$\underline{0.372 \pm 0.025}$
MelPOW	0.644 ± 0.026	0.532 ± 0.024	0.672 ± 0.030	0.229 ± 0.044	0.517 ± 0.017
MelDiffPhase	0.763 ± 0.015	0.611 ± 0.021	$\mathbf{0.764 \pm 0.007}$	$\underline{0.062 \pm 0.051}$	0.568 ± 0.045
MelPOW+MelDiffPhase	$\mathbf{0.773 \pm 0.019}$	$\mathbf{0.656 \pm 0.015}$	$\mathbf{0.764 \pm 0.028}$	$\mathbf{0.267 \pm 0.042}$	$\mathbf{0.648 \pm 0.021}$
MelDiffComplex	$\underline{0.569 \pm 0.014}$	$\underline{0.514 \pm 0.037}$	$\underline{0.658 \pm 0.011}$	0.149 ± 0.088	$\underline{0.382 \pm 0.043}$

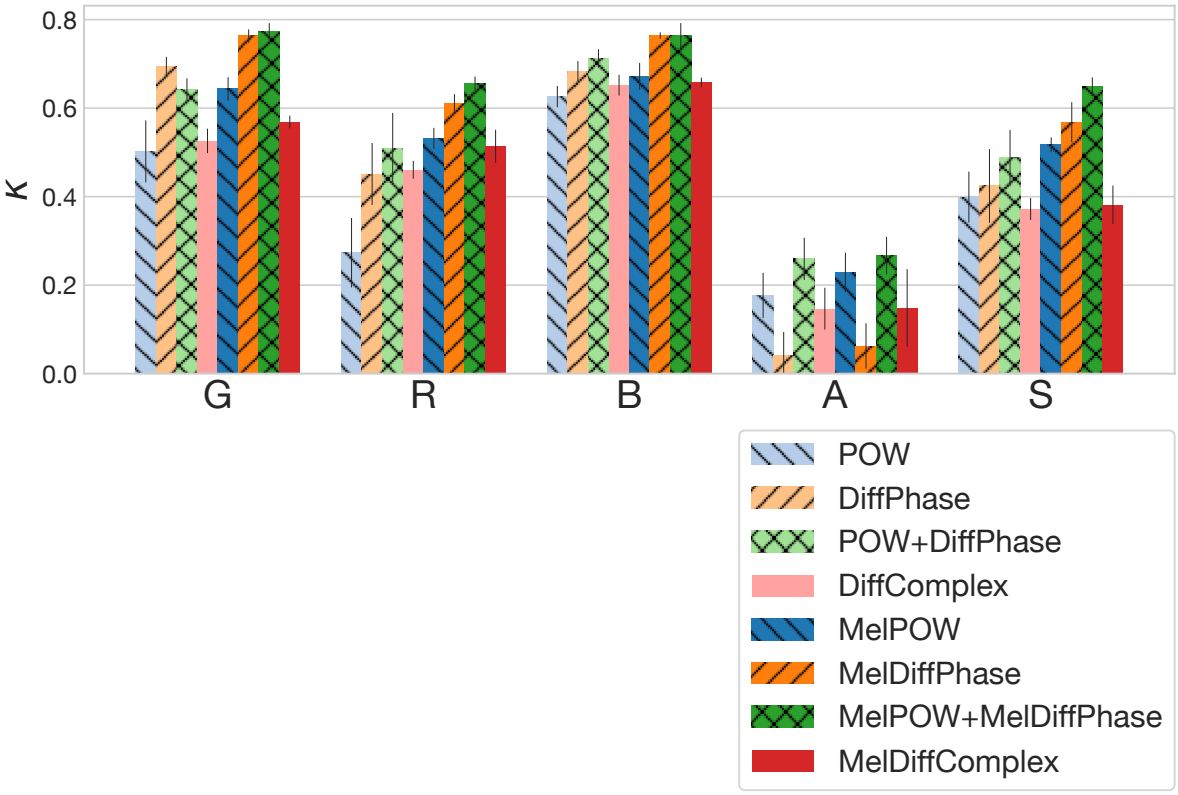


図 5.4 5 分割交差検証で得られた二次重み付け Cohen のカッパ係数 κ の結果（表 5.3 と同じデータ）。エラーバーは標準偏差を表す。テストデータセットは，検証損失が最小となったエポックのモデルを使って評価された。この図で示している GRBAS 尺度の各項目，各特徴量に対する κ の値は，表 5.2 のハイパーパラメータを使用した時に得られたものである。

5.4.1　本章の研究の限界

本章の実験では，DNN の主要な構成要素として RNN の一種である双方向 GRU を使用した。しかし，RNN は時間方向の依存関係を学習するのに特化した構造ではある一方で，どの周波数ビン同士が隣り合っているか等の，周波数方向の位置に関する情報については全く考慮されない。時間周波数表現のすべての位置情報を効果的に用いるためには，RNN よりも 2 次元 CNN を使うことが適切であると考えられる。実際，2 次元 CNN と時間周波数表現の組み合わせは，音声認識や感情音声分類，言語識別，音源分離，音声強調等の様々な応用において高い性能を発揮している [72, 75, 76, 198–202]。

本章の実験の結果，DiffPhase（MelDiffPhase）は DiffComplex（MelDiffComplex）よりも，病的音声の声質評価に効果的であることが示唆された。しかし，DiffComplex を計算する上では，瞬時位相修正がかなり重要な役割を果たしている一方で，DiffPhase に類似する時間周波数表現は，瞬時位相修正のような手法を用いずに求めることができる。DiffPhase は，瞬時位相修正を適用した位相スペクトログラムの時間差分の絶対値として定義したが（図 5.5 (C)），瞬時周波数の時間差分の絶対値は DiffPhase と非常に似た表現を与える（図 5.5 (B)）。仮に，瞬時周波数を 1 チャンネルの 2 次元画像として 2 次元 CNN に入力した場合，時間差分は 1 層の畳み込み層の 1 チャンネルの差分フィルタとして表現可能であり，絶対値は 2 つの ReLU で表現できる（$|x| = \mathrm{ReLU}(x) + \mathrm{ReLU}(-x)$）。つまり，DiffPhase に類する表現は，特徴表現学習によって，瞬時周波数から 2 次元 CNN の内部で容易に獲得されうるものである。また，時間差分や絶対値のような処理を行うと，瞬時周波数（図 5.5 (A)）に元々含まれていた情報の一部が失われてしまう。そのため，2 次元 CNN への入力特徴量としては，DiffPhase のように事前に時間差分や絶対値を適用した表現ではなく，瞬時周波数を直接用いることがより適切であると考えられる（但し，瞬時周波数の外れ値の影響を抑制する工夫は必要である）。

5.5　まとめ

本章の実験では，2,545 サンプルの病的音声データを元にした GRBAS 尺度の評点の自動推定実験を行い，時間周波数表現の位相情報として，特に瞬時周波数に関連する情報の有効性について検討した。実験の結果，項目 G，R，B，S に対しては，振幅情報よりも位相情報を用いたときに高い推定精度が得られた。この結果は，深層学習を用いた病的音声の声質評価において，位相情報が有効であることを示唆するものである。また，周波数軸のメル尺度への変換は，振幅情報だけでなく，位相情報に関しても有効であることが示された。次の課題としては，時間周波数表現と相性が良いことが知られている 2 次元

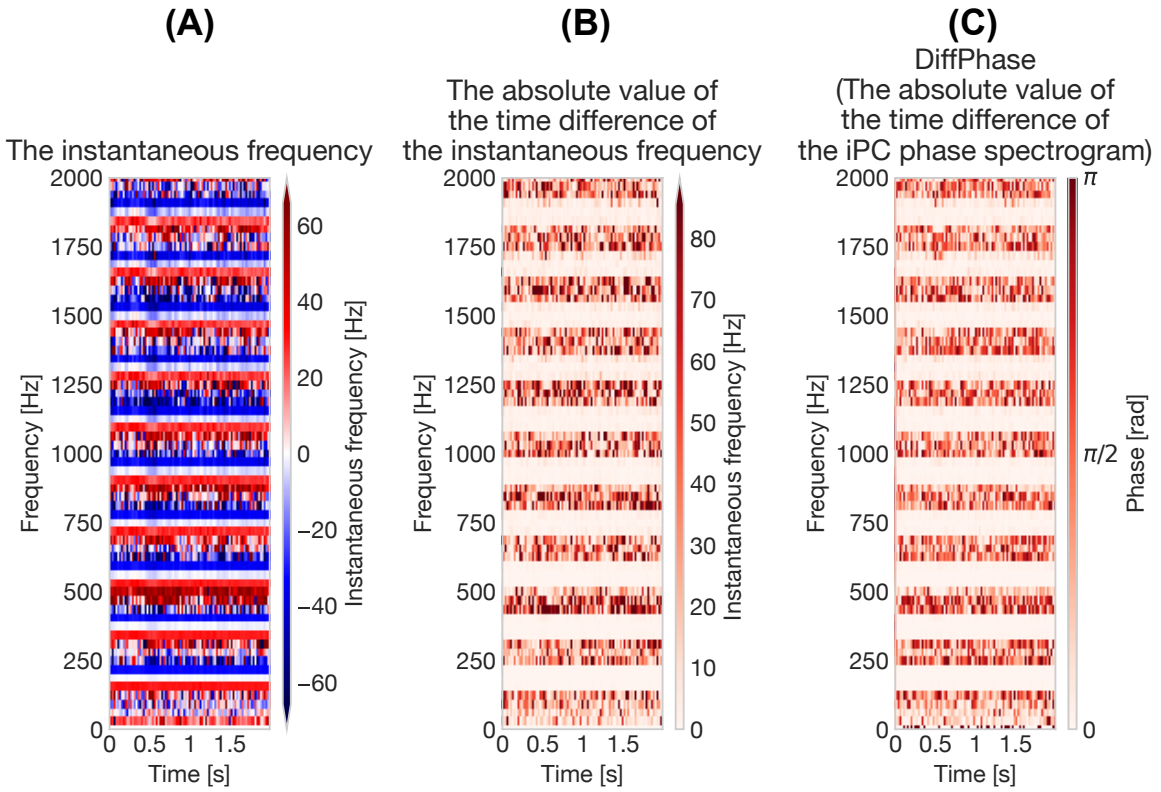


図 5.5 持続母音/a/の音声波形から計算される (A) 瞬時周波数，(B) 瞬時周波数の時間差分の絶対値，(C) DiffPhase（瞬時位相修正を適用した位相スペクトログラムの時間差分の絶対値）のヒートマップ。ヒートマップの計算には，実験と同じパラメータを用いた。外れ値の影響を除くため，時間周波数表現の絶対値の 95 パーセンタイル値を $\theta_{95\%}$ と記すとき，(A) では色の表示範囲を $[-\theta_{95\%}, \theta_{95\%}]$ に，(B) では $[0, \theta_{95\%}]$ に制限した。

CNN を DNN の構造に組み入れることに加えて，非可逆処理によって情報の欠損が生じる DiffPhase のような表現よりも，情報の欠損が少ない表現を入力特徴量として使うことが必要になると考えられた。

第 6 章

様々な音分類課題に対する位相情報の有効性の検討

6.1 はじめに

前章では，深層学習を用いた GRBAS 尺度の評点の自動推定における，位相情報の有効性と，位相情報の周波数軸をメル尺度に変換することの有効性を示した。その一方で，音響信号を対象とした深層学習に関する研究全般に関して，位相情報の有効性について検討した研究は多くない。音声強調 [71, 72] やなりすまし検出（spoofing detection）[223, 224] 等の一部の課題を除き，依然として振幅情報のみが用いられる場合が多数を占める [66, 68, 69, 75, 77–79]。音声強調は音声波形から音声波形を再構成する課題であり，DNN によって出力される音声波形の聴覚的音声品質を高めるためにも位相情報は重要である [113]。また，なりすましの検出で位相情報が有効な理由に対する定説はないが，音声合成では振幅情報に比べると位相情報の優先度が低い場合が多いため，位相情報に合成音声らしさが現れるのではないかと推察される。しかし，なりすまし検出のような一部の例外を除いた分類や回帰を目的とした課題では，出力が音声波形でなく，位相を積極的に使用する動機付けが少ない。このことが，位相情報を用いた研究が少ない理由であると考えられる。

そこで，本章では，GRBAS 尺度の評点の自動推定という課題から一度離れ，様々な音分類課題に対して，位相情報の有効性を検討した結果を報告する。前章では深層学習を用いた時の，瞬時周波数に強く関連する位相情報の有効性を示した。その一方で，深層学習を用いていない研究も含めると，オンセット検出 [127–130] 等の様々な課題で，群遅延の有効性が示されている。また，瞬時周波数や群遅延といった情報は，そもそも位相から算出されるものであるため，位相自体の有効性は検討されるべきであると考えられる。位相そのものは 2π で値が循環するため取り扱いが難しいが，位相 ϕ を $e^{i\phi}$ のようにフェーザ

表示にすれば，アンラップの必要がなくなって取り扱いやすくなる。以上より，本章の実験では，位相情報として，瞬時周波数と群遅延，および位相のフェーザ表示を検討した。なお，前章の実験では，周波数軸のメル尺度への変換は位相情報についても効果的であることが示された。そこで，本章では，上記の位相情報を，メル尺度のような周波数軸を持つ時間周波数表現で表す手法を提案した。その他に，前章の実験では DNN の主要な構造として RNN を用いたが，本章の実験では 2 次元 CNN を採用した。

本章の実験で作成したプログラムコードは GitHub のリポジトリ[*1]に公開している。

6.2　手法

第 2.3.2 節で述べたように，振幅情報の周波数軸は，メルフィルタバンクを使ってメル尺度に変換されてきた。しかし，この計算方法では，第 2.3.3 節で言及したように，位相情報の周波数軸はどのようにメル尺度に変換されるかが自明ではない。

対数振幅メル周波数スペクトログラムが現在も主流である一方で，近年では微分可能な演算機構を用いて，対数振幅スペクトログラムに匹敵するような最適な時間周波数表現を学習するアプローチ [149, 167, 168] も研究されている。それらの研究の一つで提案されている LEarnable Audio Frontend（LEAF）と呼ばれる手法は，様々な音分類課題において，対数振幅メル周波数スペクトログラムに匹敵する性能となることが実験的に示されている [168]。また，LEAF はその計算過程で，メル尺度のような周波数軸を持つ時間周波数表現を複素数領域で直接計算する。

そこで，本章の実験では，メル尺度のような周波数軸を持つ位相情報の時間周波数表現について検討するために，対数振幅メル周波数スペクトログラムではなく，LEAF を改変した上で使用した。本章では，振幅情報だけでなく位相情報も出力する LEAF を拡張 LEAF と呼び，位相情報の有無を区別しない場合は単に LEAF と呼ぶ。なお，LEAF の周波数軸はメル尺度に対応するように初期化されるが，学習に伴う変化によって完全にはメル尺度でなくなる。

6.2.1　オリジナルの LEAF：振幅情報の計算パス

図 6.1 に拡張 LEAF のブロックダイアグラムを示す。本節では，Zeghidour ら [168] によって提案されたオリジナルの LEAF に含まれている構成要素のみを説明する。オリジナルの LEAF には，位相情報の計算パスはなく，圧縮パワーのみを計算して出力する。オリジナルの LEAF は，原波形から時間周波数表現への計算と周波数軸の変換を行う Gabor フィルタバンク，振幅二乗値の計算，時間方向のダウンサンプリングを行う

[*1] https://github.com/onkyo14taro/investigation-phase

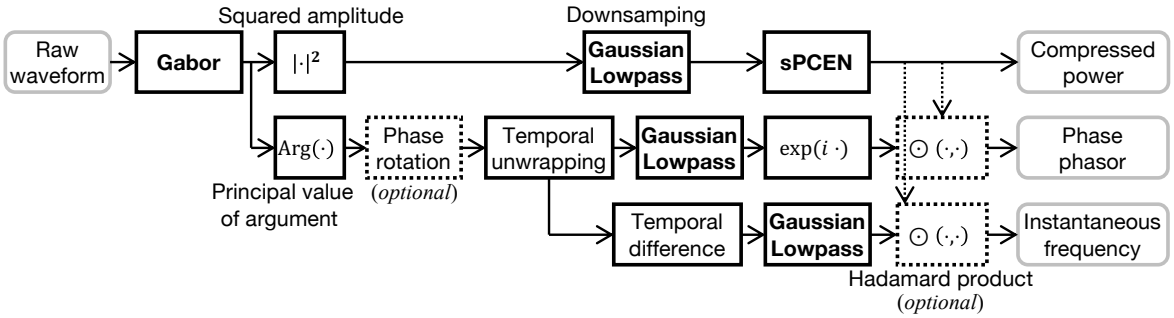


図 6.1 拡張 LEAF のブロックダイアグラム。圧縮パワー（compressed power）に繋がる計算パス以外を取り除くと，オリジナルの LEAF のブロックダイアグラムとなる。冗長性を排するため，この図では群遅延の計算パスは省略されている。群遅延の計算パスは，瞬時周波数の計算パスの時間方向のアンラップと差分を，それぞれ周波数方向のアンラップと差分に置換したものになる。

Gaussian ローパスフィルタバンク，振幅の非線形圧縮を行う sPCEN コンプレッサの縦続接続である。振幅二乗値の計算を除き，各構成要素は周波数ビンごとに学習パラメータを持つ。本節では，LEAF の周波数ビンの総数を $M \in \mathbb{N}$，窓幅を $W \in \mathbb{N}$ とし，窓関数の添字集合を $\mathfrak{W} := \{-\lfloor W/2 \rfloor, -\lfloor W/2 \rfloor + 1, \ldots, -1, 0, 1, \ldots, \lfloor (W-1)/2 \rfloor\}$ と表わすことにする。

Gabor フィルタバンク $\{\varphi_m : \mathfrak{W} \to \mathbb{C}\}_{m=0}^{M-1}$ は，中心正規化角周波数 $\{\omega_m : \mathfrak{W} \to [0, \pi]\}_{m=0}^{M-1}$（値域は取りうる値の制限値）と逆帯域幅パラメータ $\{\sigma_{\varphi,m} : \mathfrak{W} \to [4\sqrt{2\ln(2)}/\pi, W\sqrt{2\ln(2)}/\pi]\}_{m=0}^{M-1}$ という 2 種類の学習パラメータを持ち，次式で表現される[*2]。

$$\varphi_m[n] = \frac{\hat{\varphi}_m[n]}{\|\hat{\varphi}_m\|_{l^1}}, \qquad \hat{\varphi}_m[n] = \exp\left(-\frac{n^2}{2\sigma_{\varphi,m}^2} + i\omega n\right) \tag{6.1}$$

ここで，$i = \sqrt{-1}$，$m, n \in \mathbb{N}$ はそれぞれ周波数ビンと時間の添字を表わす。第 m 番目の Gabor フィルタ φ_m の周波数特性は，中心正規化角周波数が ω_m，正規化角周波数の単位で半値全幅が $2\sqrt{2\ln 2}/\sigma_{\varphi,m}$ であるようなガウス分布となる。Gabor フィルタバンクは，その周波数特性が M チャンネルのメルフィルタバンクに類似する形状に，具体的には各フィルタの中心周波数と半値全幅が同じになるように初期化される。このフィルタバンクは各フィルタがチャンネルであるカーネルと見做され，原信号と 1 次元畳み込みを行うこ

[*2] Zeghidour らの論文 [168] では，中心正規化「周波数」が学習パラメータとされていたが，その論文の公開実装 [225] では，中心正規化「角周波数」が学習パラメータとなっていた。また，論文 [168] では逆窓幅パラメータの値域の上限は $2W\sqrt{2\ln(2)}/\pi$ とされていたが，公開実装 [225] では，$W\sqrt{2\ln 2}/\pi$ となっていた。本章の実験では，基本的に公開実装 [225] に準拠した。その他に，論文 [168] では各フィルタが $\varphi_m[n] = \hat{\varphi}_m[n]/(\sqrt{2\pi}\sigma_{\varphi,m})$ のように l^1 正規化されていたが，この正規化では Gabor フィルタが無限長でない場合に正確に l^1 ノルムが 1 にならないため，本章の実験では正確な l^1 正規化に修正した。

とで，時間周波数表現が計算される。仮に，中心正規化角周波数が $\omega_m = m\omega$ $(\omega \in \mathbb{R})$ のように m に比例し，逆帯域幅パラメータ $\{\sigma_{\varphi,m}\}$ がすべて同じ値ならば，この変換はガウス窓を用いた（位相の符号は反転した）短時間フーリエ変換に相当する。なお，ここでの一次元畳み込みでのストライドは 1 である他，学習によって $\{\omega_m\}_{m=0}^{M-1}$ が単調増加でなくなった場合，単調増加になるように強制的に $\{\omega_m\}_{m=0}^{M-1}$ を並び替える。

Gaussian ローパスフィルタバンク $\{\phi_m : \mathfrak{W} \to \mathbb{C}\}_{m=0}^{M-1}$ は，窓幅パラメータ $\{\sigma_{\phi,m} : \mathfrak{W} \to [2/W, 0.5]\}_{m=0}^{M-1}$ を学習パラメータとして持ち，次式で表現される[*3]。

$$\phi_m[n] = \frac{\hat{\phi}_m[n]}{\|\hat{\phi}_m\|_{l^1}}, \qquad \hat{\phi}_m[n] = \exp\left(-\frac{n^2}{2(0.5(W-1)\sigma_{\phi,m})^2}\right) \tag{6.2}$$

Gabor フィルタバンクの出力の二乗振幅と Gaussian ローパスフィルタバンクに対して，1 より大きなストライドで 1 次元深さ別畳み込みを行うことによって，時間周波数表現を時間方向にダウンサンプリングする。

sPCEN コンプレッサ $\mathrm{PCEN} : \mathbb{R}_{\geq 0}^{M \times N} \to \mathbb{R}_{\geq 0}^{M \times N}$ は，周波数チャンネルごとにパワーを非線形圧縮する手法である Per-Channel Energy Normalization（PCEN） [226] を，微分可能で誤差逆伝播による学習を可能にしたものであり，次式で表される。

$$\begin{aligned} \Big(\mathrm{PCEN}(\boldsymbol{F})\Big)[m,n] &= \left(\frac{\boldsymbol{F}[m,n]}{(\epsilon + \boldsymbol{M}[m,n])^{\alpha_m}} + \delta_m\right)^{1/r_m} - \delta_m^{1/r_m}, \\ \boldsymbol{M}[m,n] &= (1-s_m)\boldsymbol{M}[m,n-1] + s_m \boldsymbol{F}[m,n] \end{aligned} \tag{6.3}$$

ここで，$N \in \mathbb{N}$ は時間フレーム長，$\boldsymbol{F} \in \mathbb{R}_{\geq 0}^{M \times N}$ は Gaussian ローパスフィルタバンクによって出力される時間周波数表現，ϵ はゼロ除算を防ぐ微小値，$\{\alpha_m \in [0,1]\}_{m=0}^{M-1}$，$\{\delta_m \in [0,\infty)\}_{m=0}^{M-1}$，$\{r_m \in [1,\infty)\}_{m=0}^{M-1}$，$\{s_m \in [0,1]\}_{m=0}^{M-1}$ は sPCEN コンプレッサの学習パラメータを表す。これ以降，sPCEN によって圧縮されたパワーを POW と記すことにする。

6.2.2 位相の 2 つの定義

ここで，拡張 LEAF の説明を一時中断して，第 2.2 節で説明した，短時間フーリエ変換の 2 つの定義について復習する。短時間フーリエ変換には，窓によって切り出された信号のフーリエ変換と見做す定義 $\mathcal{F}_1^{(\boldsymbol{w},a,D)}$ と，窓と変調された信号の畳み込みと見做す定

[*3] Zeghidour らの論文 [168] では，窓幅パラメータの値域は言及されていなかったが，その論文の公開された実装 [225] では値域が設定されていた。本章の実験では公開実装 [225] に準拠して値域を設定した。また，Zeghidour らの論文 [168] では各フィルタが $\phi_m[n] = \hat{\phi}_m[n]/(\sqrt{2\pi} \cdot (0.5(W-1)\sigma_{\varphi,m}))$ のように l^1 正規化されていた。しかし，その論文の公開された実装 [225] では，Gaussian ローパスフィルタバンクにはそもそも l^1 正規化が施されていなかった。

義 $\mathcal{F}_2^{(\boldsymbol{w},a,D)}$ が存在した。

$$\left(\mathcal{F}_1^{(\boldsymbol{w},a,D)}\boldsymbol{x}\right)[m,n] := \sum_{l=-\lfloor W/2\rfloor}^{\lfloor (W-1)/2\rfloor} \boldsymbol{x}[an+l]\boldsymbol{w}[l]\exp\left(-\frac{2\pi iml}{D}\right) \tag{6.4}$$

$$\left(\mathcal{F}_2^{(\boldsymbol{w},a,D)}\boldsymbol{x}\right)[m,n] := \sum_{l=an-\lfloor W/2\rfloor}^{an+\lfloor (W-1)/2\rfloor} \boldsymbol{x}[l]\boldsymbol{w}[l-an]\exp\left(-\frac{2\pi iml}{D}\right) \tag{6.5}$$

ここで，$\boldsymbol{x} \in \mathbb{R}^X (X \in \mathbb{N})$ は入力信号，$\boldsymbol{w} \in \mathbb{R}^W (W \in \mathbb{N})$ は窓関数，$i = \sqrt{-1}$，$m \in \{0, 1, \ldots, M-1\}$ は周波数の添字，$n \in \{0, 1, \ldots, N-1\}$ は時間の添字，$a \in \mathbb{N}$ はシフト長，$D \in \mathbb{N}\ (D \geq W)$ は DFT 点数である。

$\mathcal{F}_1^{(\boldsymbol{w},a,D)}\boldsymbol{x}, \mathcal{F}_2^{(\boldsymbol{w},a,D)}\boldsymbol{x} \in \mathbb{C}^{M\times N}$ の位相を $\angle\mathcal{F}_1^{(\boldsymbol{w},a,D)}\boldsymbol{x}, \angle\mathcal{F}_1^{(\boldsymbol{w},a,D)}\boldsymbol{x} \in \mathbb{R}^{M\times N}$，瞬時周波数を $\boldsymbol{I}_1, \boldsymbol{I}_2 \in \mathbb{R}^{M\times N}$，群遅延を $\boldsymbol{G}_1, \boldsymbol{G}_2 \in \mathbb{R}^{M\times N}$ と表すとき，以下の関係式が成立する。

$$\left(\angle\mathcal{F}_1^{(\boldsymbol{w},a,D)}\boldsymbol{x}\right)[m,n] = \left(\angle\mathcal{F}_2^{(\boldsymbol{w},a,D)}\boldsymbol{x}\right)[m,n] + 2\pi amn/D \tag{6.6}$$

$$\boldsymbol{I}_1[m,n] = \boldsymbol{I}_2[m,n] + 2\pi m/D \tag{6.7}$$

$$\boldsymbol{G}_1[m,n] = \boldsymbol{G}_2[m,n] - an \tag{6.8}$$

但し，瞬時周波数と群遅延の単位は，それぞれ rad/sample，sample とした。$\boldsymbol{I}_1$ は絶対瞬時周波数，$\boldsymbol{I}_2$ は相対瞬時周波数である。$\boldsymbol{G}_2$ は時間方向のバイアス（時間の絶対位置への依存）があるため，通常，群遅延と言えば $\boldsymbol{G}_1$ のみを指すが，便宜上，$\boldsymbol{G}_2$ のことも群遅延と呼ぶことにする。拡張 LEAF は，定義の異なる位相，瞬時周波数，群遅延を出力できるように設計した。

6.2.3 拡張 LEAF：位相情報の計算パス

本節では，拡張 LEAF（図 6.1）の位相情報の計算パスについて説明する。既に述べたが，図 6.1 が拡張 LEAF の全体図であり，拡張 LEAF は，2 つの位相の定義に基づいて，位相のフェーザ表示と瞬時周波数，群遅延を出力できる。なお，本章では，位相，瞬時周波数，群遅延という用語を，その符号に注意を払わずに用いる。実際，拡張 LEAF が出力する位相と瞬時周波数は，短時間フーリエ変換で定義されるものとは符号が逆である。しかし，短時間フーリエ変換で定義される本来の符号から反転していても，それは拡張 LEAF の出力を最初に受け取る線形層の重みのパラメータの符号が反転するのと変わらない。そのため，位相情報の符号が元の定義と同じかどうかは，DNN の学習には影響を与えず，位相情報の符号に特別拘る必要はない。

まず初めに，位相のフェーザ表示の計算パスを説明する。これ以降，簡単のため，位相のフェーザ表示を位相フェーザと呼ぶことにする。位相の計算パスでは，初めに，Gabor

フィルタバンクが出力する複素時間周波数表現の偏角の主値 $\theta_1 \in (-\pi, \pi]$ が計算される。この偏角 θ_1 は，短時間フーリエ変換の位相 $\angle \mathcal{F}_1^{(\boldsymbol{w},a,D)}\boldsymbol{x}$ に対応する。オプションで，この偏角 θ_1 は，異なる定義の短時間フーリエ変換の位相 $\angle \mathcal{F}_2^{(\boldsymbol{w},a,D)}\boldsymbol{x}$ に対応する $\theta_2 \in (-\pi, \pi]$ に変換される。

$$\theta_2[m,n] = p\big(\theta_1[m,n] + \eta_m n\big) \tag{6.9}$$

ここで，η_m は Gabor フィルタバンクの第 m 番目のフィルタの中心正規化角周波数，$p(\theta) := \theta_{\mathrm{principal}}$ は偏角 $\theta \in \mathbb{R}$ を主値 $\theta_{\mathrm{principal}} \in (\pi, \pi]$ に変換する関数である。その後，位相 θ_1, θ_2 を時間方向にアンラップし，Gaussian ローパスフィルタバンクでダウンサンプリングする。最後に，ダウンサンプリングした位相を $f(\theta) := e^{i\theta}$ によってフェーザ表示に変換する。位相フェーザは，実部と虚部の 2 チャンネル画像として見做される。これ以降，最終的に得られる位相フェーザを $\mathrm{PHS}_1, \mathrm{PHS}_2$ と記す（下付き添字の 1 と 2 はそれぞれ，位相の定義変更を行わなかったものと，定義変更を行ったものを表す)。

続いて，瞬時周波数の計算パスを説明する。瞬時周波数の計算パスでは，初めに偏角 θ_1 または θ_2 を計算する。そして，偏角を時間方向にアンラップした後に，時間方向の差分で瞬時周波数を近似的に計算する[*4]。Gabor フィルタバンクのストライドが 1 であることから，差分によって十分に微分を近似できていると考えられる。最後に，Gaussian ローパスフィルタバンクでダウンサンプリングを行うことによって，最終的な瞬時周波数の出力が得られる。第 2.2.6 節で説明したように，位相の微分を解析的に求めた場合，パワーが小さい要素で，位相の微分の絶対値が極端に大きくなることがある。しかし，拡張 LEAF では，解析的な微分の代わりに差分計算を用いていることから，値の絶対値が極端には大きくなりにくい。また，Gaussian ローパスフィルタバンクによって，パルスのように瞬間的に生じる絶対値が大きな値は抑制される。これ以降，拡張 LEAF によって出力される瞬時周波数を $\mathrm{IF}_1, \mathrm{IF}_2$ と記す。

群遅延は，瞬時周波数の計算パスの時間方向のアンラップと差分を，それぞれ周波数方向のアンラップと差分に置換することによって計算することができる。これ以降，拡張 LEAF によって出力される群遅延を $\mathrm{GD}_1, \mathrm{GD}_2$ と記す。

オプションで，位相特徴量に対して，圧縮パワー POW を要素ごとに乗算した。この要素ごとの乗算は，パワーが弱い要素での位相特徴量の影響力を弱めることが期待される。

図 6.2 に拡張 LEAF が出力する特徴量のヒートマップを示す。正弦波成分が存在する領域に関して，PHS_1 では値が周波数軸方向に揃って，時間方向に回転しているパターンが観察できる一方で，PHS_2 ではそのようなパターンは見えない。IF_1 は絶対瞬時周波数に対応するため周波数軸に沿ったバイアスが存在する一方で，IF_2 は相対瞬時周波数

[*4] 実装では，時間方向にアンラップした後に差分を計算する代わりに，時間方向に差分を計算した後に主値 $(-\pi, \pi]$ へラップした。計算誤差を無視すれば，2 つの処理は同じ結果を出力する。

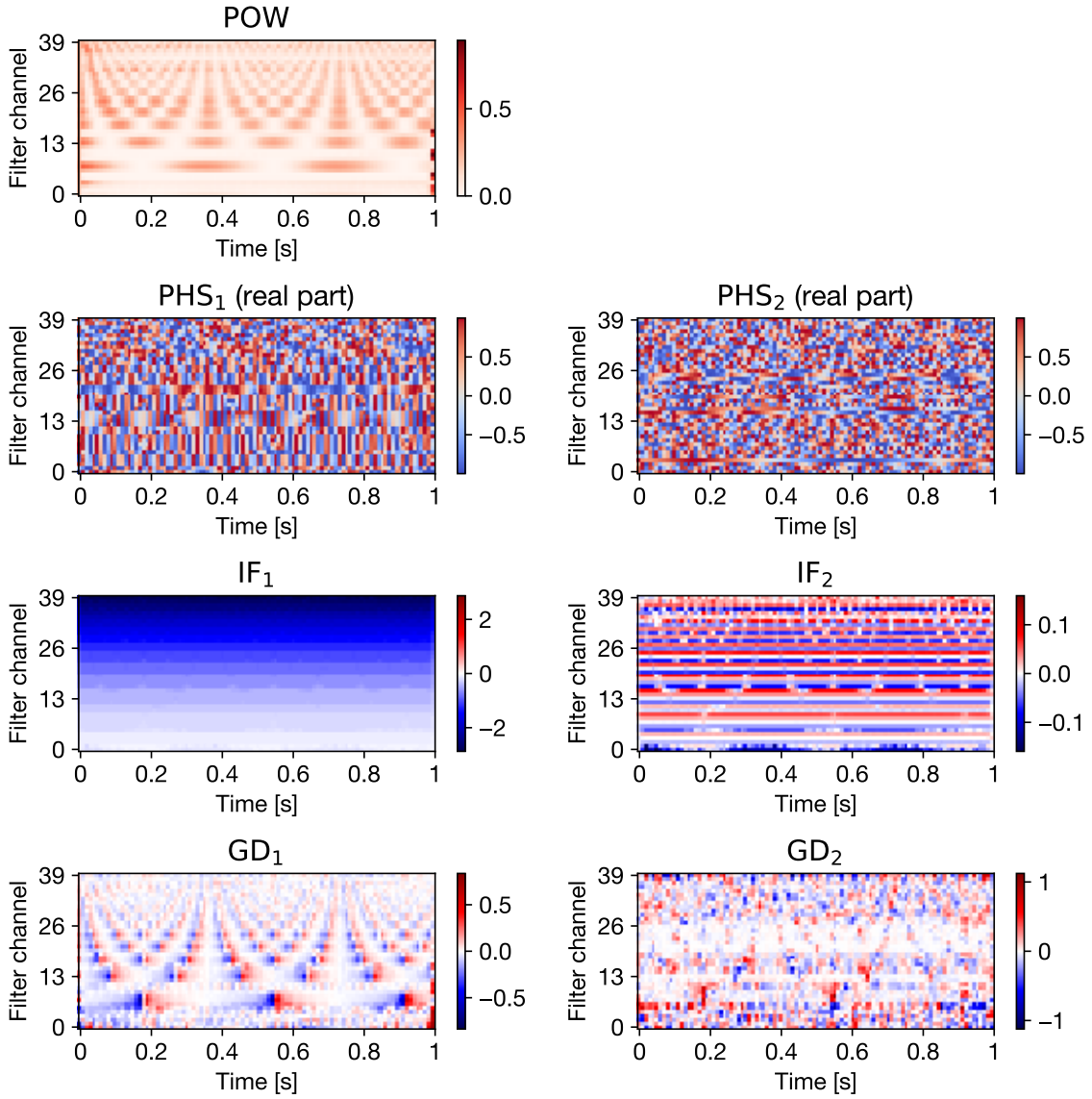


図 6.2 未学習の拡張 LEAF から出力された特徴量のヒートマップ。入力としては音高が C5 の電子キーボードの音（データセット NSynth [227] の “keyboard_electronic_012-072-127.wav”）を用い，拡張 LEAF のパラメータは実験と同じ値に設定した。

に対応しておりバイアスがない。GD_1 ではオンセットで値が大きく，オフセットで値が小さくなるパターンが観察できる一方で，GD_2 では不可解で乱雑なパターンが表れる。式 (6.8) によると，GD_2 には時間軸に沿ったバイアスが表れるように期待される。しかし，実際には GD_2 は周波数方向の微分ではなく差分であり，位相が 2π で循環する性質によって，差分で微分を近似できなくなってしまう。このため，本実験で定義した GD_2 には，図 6.2 に示したように非本質的なノイズが強く表れる。

表 6.1 本章の実験で調査した音分類課題とデータセット。

課題	データセット	クラス数	サンプル数 訓練	検証・テスト
楽器の音高識別	NSynth [227]	112	289,205	16,774
楽器の種類識別	NSynth [227]	11	289,205	16,774
言語識別	VoxForge [201]	6	148,654	27,764
話者識別	VoxCeleb [75]	1,251	128,086	25,430
鳥の存在判定	DCASE2018 [228]	2	35,690	12,620
音環境分類	TUT [229]	10	6,122	2,518
単語認識	SpeechCommands [230]	35	84,843	20,986
感情音声分類	CREMA-D [231]	6	5,146	2,296

6.2.4　振幅が 0 の要素における位相の取り扱い

時間周波数表現の振幅が 0 となる要素では，位相は定義されない。実装では，このように不定である位相の要素を使用して計算された瞬時周波数および群遅延の要素も，同様に不定であると見做した。Gaussian ローパスフィルタバンクによるダウンサンプリングの際は，不定の値を線形補間によって置換した。ダウンサンプリングの中心要素の値が不定である場合，ダウンサンプリング後の該当する要素の値も不定とした。最終的に，不定の値は，瞬時周波数と群遅延に関しては 0 で，位相に関しては位相フェーザを $0 + 0i$ で置換した。また，振幅が 0 となる要素における偏角の誤差逆伝播は 0 とした。

6.3　実験

本章の実験では，表 6.1 に示す 8 種類の音分類課題に対して，位相情報の付与が性能向上に寄与するかどうかを調査した。各課題について，Zeghidour らによる LEAF の提案論文 [168] で使用されていたのと同じデータセットを用いた。各データセットの標本化周波数は 16 kHz に統一した。複数の録音条件を含むデータセットに関しては，訓練用と検証用の間で録音条件が共有されないように分割を行った。Zeghidour らの論文 [168] では，データセットがどのように訓練用・検証用に分割されていたかが不明であったため，本実験では独自にデータセットを分割した。そのため，Zeghidour らの論文 [168] の結果と本実験の結果は，直接比較できないことに留意する必要がある。なお，Zeghidour らの論文 [168] では検証データセットとテストデータセットが区別されていなかったため，本実験でも，検証データセットとテストデータセットには同じものを用いた。

音分類課題のための DNN として，拡張 LEAF，2 次元 Batch Normalization [180] と 2 次元 CNN の一種である EfficientNetB0 [232] を縦続接続したモデルを用いた（EfficientNetB0 の詳細に関しては付録 A.1 を参照）。拡張 LEAF の圧縮パワーと位相情報は，2 次元画像の意味でのチャンネル方向に重ねた。また，位相フェーザ $\mathrm{PHS}_1, \mathrm{PHS}_2$ に関しては，1 チャンネルの複素数の画像と見做して複素 Batch Normalization [233] を適用した後，実部と虚部を分解して 2 チャンネルからなる実数の画像と見做して圧縮パワーとチャンネル方向に重ねた。通常の Batch Normalization は実直線上での平均と標準偏差に関して正規化するのに対し，複素 Batch Normalization は複素平面上での平均と共分散に関して正規化（白色化）を行う。拡張 LEAF の周波数ビンの総数 M は 40，窓幅 W は 401 とした。Gabor フィルタバンクは，HTK [147] で定義されたメル尺度上のメルフィルタバンクと類似する周波数特性を持つように初期化した。Gaussian ローパスフィルタバンクの $\sigma_{\phi,m}$ は 0.4 に初期化した。sPCEN コンプレッサの $\alpha_m, \delta_m, r_m, s_m$ はそれぞれ 0.96，2.0，2.0，0.04 に初期化し，ゼロ除算を防ぐ ϵ は 5×10^{-8} に設定した。

訓練時は，サンプルからランダムに 1 秒間の区間を切り出し入力に用いた。検証・テスト時は，サンプルを重複なく 1 秒ずつ切り出し入力に用い，それぞれに対して DNN が出力するロジット（logit; softmax 関数を除いた出力）の平均値を用いて評価を行った。鳥の存在判定，言語識別，楽器の種類識別に関しては，50,000 ステップ連続して，他の課題に関しては，20,000 ステップ連続して検証損失が改善しなかった時点で学習を終了した。損失関数としては交差エントロピー誤差，最適化アルゴリズムとしては Adam [164] を使用し，学習率は 0.0001 に設定した。バッチサイズは 256 とした。過剰適合を抑制するために，EfficientNetB0 への入力に，SpecAugment [234]（論文 [234] の中のパラメータ W: 0，F: 5，T: 11，時間マスキングと周波数マスキングの回数: それぞれ 2 回ずつ），EfficientNetB0 の畳み込み層に l^2 正則化（正則化定数: 0.00001），EfficientNetB0 の最後の全結合層に Dropout [173]（率: 0.2），正解ラベルに label smoothing [235]（率: 0.1）を適用した。なお，SpecAugment は，時間周波数表現の一部領域の値を 0 に置換することで入力情報をマスクする正則化手法，label smoothing は，正解ではないクラスにも値を割り当てて one-hot 表現にノイズを加える正則化手法である。なお，DNN の構築と評価には機械学習フレームワークである PyTorch [222] を用いた。

表 6.2 音分類課題に対する正解率（%）。± の後の値は 95% 信頼区間を表わす。正解率（%）を $100p$，標本サイズを N とするとき，95% 信頼区間は $1.96((p(1-p))/N)$ という式で計算した。$\odot$ は要素ごとの乗算（Hadamard 積）を表す。POW を有意に（95% 信頼区間を超えて）上回る正解率を**太字**で，POW を有意に下回る正解率を下線で示す。

特徴量	楽器の音高識別	楽器の種類識別	言語識別	話者識別	鳥の存在判定	音環境分類	単語認識	感情音声分類
POW	91.5 ± 0.4	69.2 ± 0.7	74.7 ± 0.5	35.5 ± 0.6	76.4 ± 0.9	66.2 ± 1.8	94.1 ± 0.3	60.2 ± 2.0
$+\mathrm{PHS}_1$	$\mathbf{92.9 \pm 0.4}$	$\mathbf{71.5 \pm 0.7}$	$\mathbf{79.3 \pm 0.5}$	36.1 ± 0.6	75.1 ± 0.9	66.7 ± 1.8	94.4 ± 0.3	60.7 ± 2.0
$+\mathrm{PHS}_1 \odot \mathrm{POW}$	91.9 ± 0.4	69.6 ± 0.7	74.5 ± 0.5	35.6 ± 0.6	75.6 ± 0.9	68.8 ± 1.8	94.1 ± 0.3	62.0 ± 2.0
$+\mathrm{PHS}_2$	$\mathbf{92.7 \pm 0.4}$	68.0 ± 0.7	$\underline{72.8 \pm 0.5}$	$\mathbf{37.4 \pm 0.6}$	74.9 ± 1.0	65.3 ± 1.9	94.4 ± 0.3	60.8 ± 2.0
$+\mathrm{PHS}_2 \odot \mathrm{POW}$	91.9 ± 0.4	69.0 ± 0.7	74.0 ± 0.5	$\mathbf{37.1 \pm 0.6}$	74.8 ± 1.0	65.5 ± 1.9	94.2 ± 0.3	61.1 ± 2.0
$+\mathrm{IF}_1$	$\mathbf{93.2 \pm 0.4}$	69.9 ± 0.7	$\underline{71.3 \pm 0.5}$	34.6 ± 0.6	76.0 ± 0.9	69.4 ± 1.8	94.2 ± 0.3	57.3 ± 2.0
$+\mathrm{IF}_1 \odot \mathrm{POW}$	92.1 ± 0.4	69.5 ± 0.7	$\mathbf{78.1 \pm 0.5}$	35.0 ± 0.6	76.2 ± 0.9	65.9 ± 1.9	94.0 ± 0.3	61.0 ± 2.0
$+\mathrm{IF}_2$	$\mathbf{93.2 \pm 0.4}$	70.2 ± 0.7	$\underline{49.2 \pm 0.6}$	$\mathbf{37.2 \pm 0.6}$	$\underline{73.7 \pm 1.0}$	68.2 ± 1.8	94.5 ± 0.3	57.5 ± 2.0
$+\mathrm{IF}_2 \odot \mathrm{POW}$	$\mathbf{93.0 \pm 0.4}$	69.9 ± 0.7	$\underline{68.3 \pm 0.5}$	36.1 ± 0.6	$\underline{73.5 \pm 1.0}$	66.8 ± 1.8	94.4 ± 0.3	60.9 ± 2.0
$+\mathrm{GD}_1$	91.3 ± 0.4	$\mathbf{72.6 \pm 0.7}$	$\mathbf{83.9 \pm 0.4}$	35.4 ± 0.6	$\mathbf{79.8 \pm 0.9}$	68.2 ± 1.8	94.4 ± 0.3	58.7 ± 2.0
$+\mathrm{GD}_1 \odot \mathrm{POW}$	$\mathbf{92.5 \pm 0.4}$	$\underline{67.7 \pm 0.7}$	$\mathbf{79.4 \pm 0.5}$	$\mathbf{37.6 \pm 0.6}$	77.0 ± 0.9	67.2 ± 1.8	94.4 ± 0.3	60.5 ± 2.0
$+\mathrm{GD}_2$	$\mathbf{92.8 \pm 0.4}$	69.2 ± 0.7	$\underline{70.5 \pm 0.5}$	$\underline{32.8 \pm 0.6}$	77.4 ± 0.9	66.4 ± 1.8	$\underline{92.8 \pm 0.3}$	60.6 ± 2.0
$+\mathrm{GD}_2 \odot \mathrm{POW}$	92.3 ± 0.4	68.3 ± 0.7	$\mathbf{76.1 \pm 0.5}$	$\underline{32.0 \pm 0.6}$	77.5 ± 0.9	68.5 ± 1.8	93.6 ± 0.3	57.9 ± 2.0

6.4 結果と考察

表 6.2 に実験結果を示す。各課題，各特徴量に関して，独立して DNN を訓練した。本節では，全体的な傾向について考察した後に，個別の課題ごとに説明する。

6.4.1 全体的な傾向とその考察

ある特定の課題に対して，位相フェーザの付与が有意な性能改善をもたらした場合，位相の微分も同様に有意な性能改善をもたらした。音声強調 [72, 114, 115] では位相の値それ自体が重要であるが，本実験の結果からは，音分類課題のように出力目標が音声波形ではない課題では，微分で表現されるような位相の隣接関係の方が重要であることが示唆された。

POW と位相情報の要素ごとの乗算は，有意な性能改善をもたらす場合もあれば，有意な性能低下をもたらす場合もあった。要素ごとの乗算は，不要な情報を抑制する一方で，位相情報の物理的意味を破壊するものでもある。この物理的意味の破壊は，要素ごとの乗算を用いるのではなく，情報の取捨選択を自動的に行うゲート機構 [236] を導入することによって，抑制することができる可能性がある。

圧縮パワーと要素ごとに乗算した位相特徴量を含めない場合，有意に性能が向上した課題の数は，位相フェーザの中では PHS_1（3 つの課題），瞬時周波数の中では IF_2（2 つの課題），群遅延の中では GD_1（3 つの課題）を付与した時に最大になった。但し，PHS_1 と GD_1 の付与は有意な性能低下を引き起こさなかった一方で，IF_2 を付与した時に 2 つの課題で有意な性能低下が生じた。IF_2 の付与による性能低下の原因に関しては，個別の課題ごとに考察する。また，GD_2 を付与した場合，有意な性能向上よりも，有意な性能低下に繋がった課題の数の方が多かった。これは，第 6.2.3 節で説明したように，GD_2 には本質的でないパターンが表れることが性能低下に繋がったと考えられる。なお，GD_2 は PHS_2 の周波数方向の差分として計算されるため，GD_2 が GD_1 より劣っているということは，PHS_2 からは群遅延に関する有益な情報が抽出されにくいことを暗に示している。このことが，PHS_2 が PHS_1 よりは有効でなかった原因の一つになっている可能性がある。

図 6.3 は，Gabor フィルタバンクの中心周波数の学習による変化を表している。全体的な傾向としては，初期値のメル尺度から大きくは変動しなかった。群遅延を含む条件 $\mathrm{POW} + \mathrm{GD}_1, \mathrm{POW} + \mathrm{GD}_2$ では，他の条件と比べて最低周波数が低くなった。拡張 LEAF では，群遅延は位相の周波数方向の差分として計算されるため，隣接するフィルタ同士の中心周波数が近い場合，それらのフィルタにおける $\mathrm{GD}_1, \mathrm{GD}_2$ の値は小さくなる。

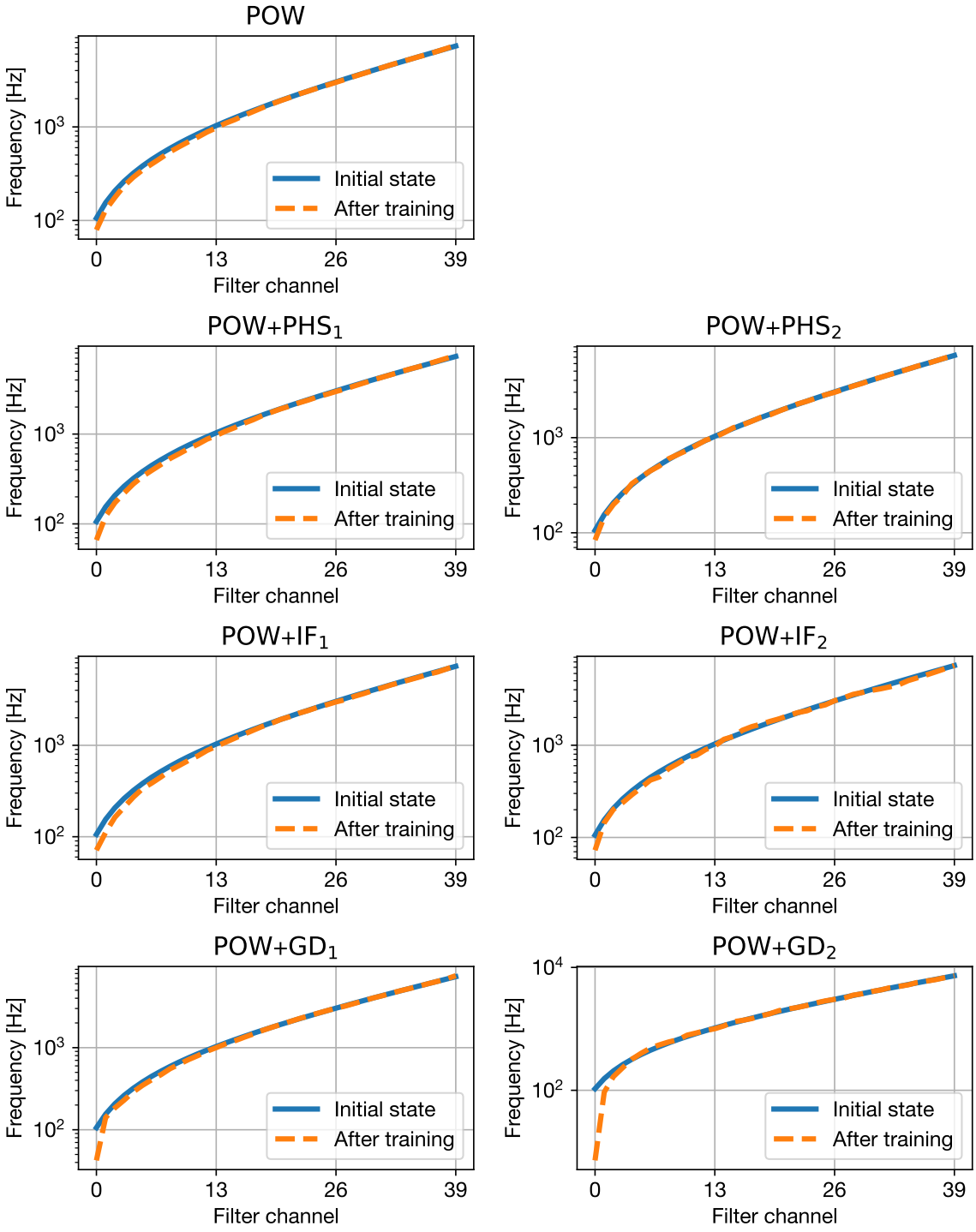


図 6.3 学習前後の Gabor フィルタバンクの中心周波数。学習後の中心周波数としては，各分類課題に対して最適化された中心周波数の，全課題における平均値を表示している。

そのため，条件 $\mathrm{POW}+\mathrm{GD}_1, \mathrm{POW}+\mathrm{GD}_2$ で中心周波数の最低値が小さくなった理由としては，低域の群遅延情報を利用するために，隣接するフィルタ同士の中心周波数が学習によって離された可能性があると考えられる。

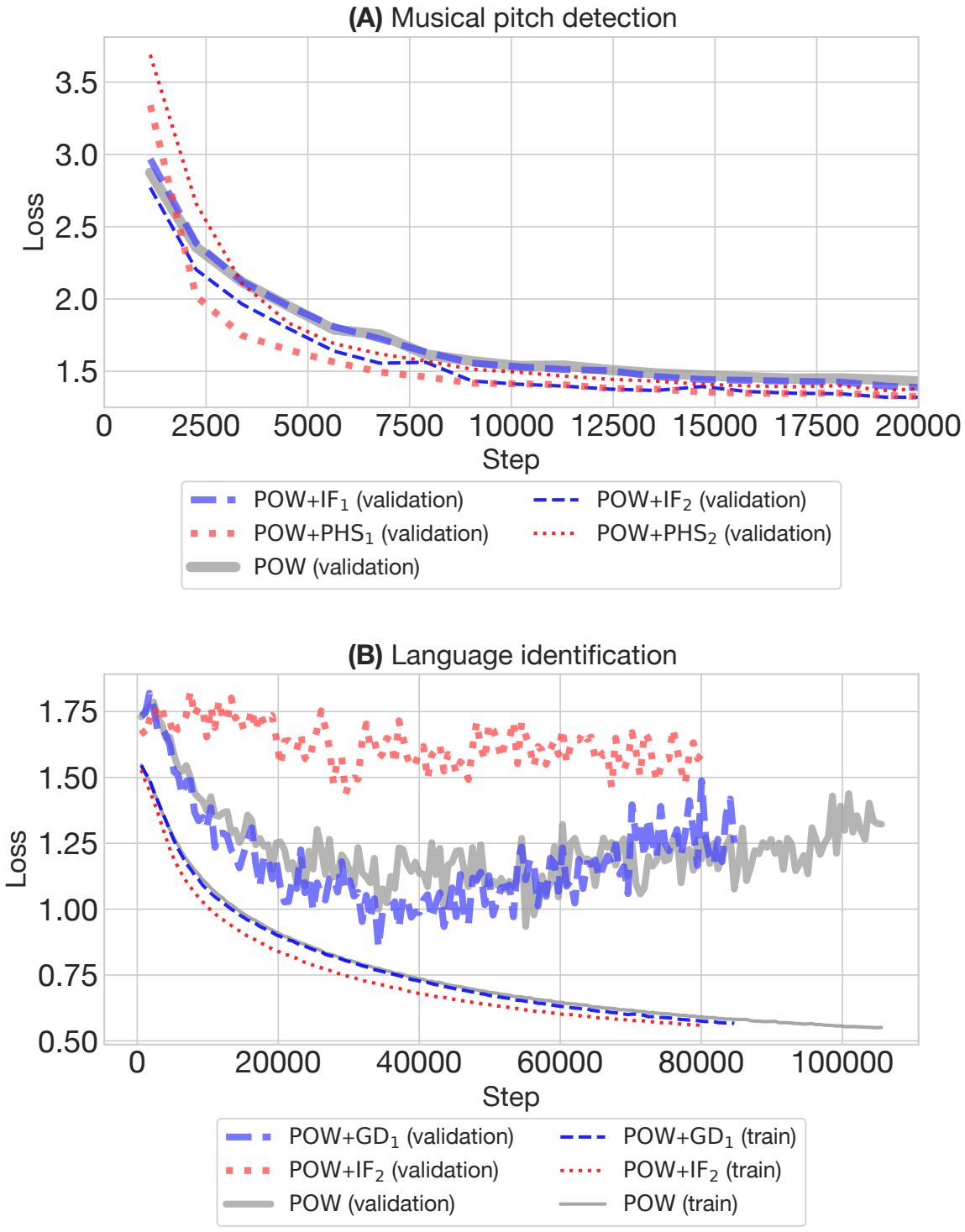


図 6.4 (A) 楽器の種類識別と，(B) 言語識別における学習曲線

6.4.2　個々の課題に対する結果と考察

楽器の音高識別課題では，位相特徴量のほとんどが性能向上に寄与し，特に瞬時周波数を付与した場合に性能が大きく向上した。瞬時周波数は基本周波数推定 [117, 118] に応用されていることからも，楽器の音高識別課題において瞬時周波数は特に有効であったのだと考えられる。図 6.4 (A) は，楽器の音高識別課題における学習曲線である。この図より，$\mathrm{POW} + \mathrm{IF}_1$ を用いた時よりも，$\mathrm{POW} + \mathrm{IF}_2$ を用いた時の方が早く収束していることがわかる。図 6.2 に示すように，IF_2 には周波数軸方向のバイアスがないのに対し，IF_1 には周波数軸方向のバイアスが存在する。CNN は相対的な位置関係に基づいて現れるパターンの抽出に特化した構造であり，IF_1 のように周波数ビンによって値の分布が異なるようなデータは，2 次元 CNN に対してあまり適当ではない可能性がある。実際，2 次元 CNN に対する入力として振幅スペクトログラムを用いた際に，周波数ビンごとに正規化を行うことで話者識別の性能が向上したことを報告した研究 [75] もある。このことから，2 次元 CNN の入力としては，絶対瞬時周波数（IF_1）よりも相対瞬時周波数（IF_2）の方が優れていることが示唆される。但し，瞬時周波数に関しては IF_1 より IF_2 の方が良かったのとは対照的に，位相フェーザに関しては PHS_2 より PHS_1 を用いた時の方が早く収束した。おそらく，図 6.2 に示すように，PHS_1 では，正弦波成分が存在する周波数帯域で値が縦に揃って横に回転するパターンが現れ，このパターンが 2 次元 CNN によって抽出されやすかったのではないかと考えられる。

楽器の種類識別課題では，GD_1 と PHS_1 の付与が有意な性能改善をもたらした。図 6.2 に示すように，群遅延はオンセットやオフセットに強く反応する。この性質からオンセット検出に応用されており [127–130]，擦弦楽器のように緩やかなオンセット（soft onset）の検出にも有効であることが示唆されている [129, 130]。このことから，楽器の種類識別課題において，群遅延は楽音の繊細な情報を捉えるのに寄与したのではないかと考えられる。

言語識別課題では，いくつかの位相特徴量の付与が有意な性能改善をもたらし，特に GD_1 の付与による性能改善の程度が大きかった。対照的に，他のいくつかの位相特徴量の付与は有意な性能低下に繋がり，特に IF_2 を付与した時の性能低下が著しかった。図 6.4 (B) は，POW，$\mathrm{POW} + \mathrm{GD}_1$，$\mathrm{POW} + \mathrm{IF}_2$ を用いた時の学習曲線である。$\mathrm{POW} + \mathrm{GD}_1$ に関しては，訓練損失と検証損失の両方が，POW よりも早く減少している様子が観察できる。その一方で，$\mathrm{POW} + \mathrm{IF}_2$ に関しては，訓練損失は最も早く減少する一方で，検証損失はほとんど減少していない様子が見て取れる。Voxforge は，多数の発話者による発話データから構成されるデータセットであり，録音環境は発話データごとに異なっていた。今回の実験では，発話者の重複がないように，訓練データセットと検証

データセットを分割した。このことから，IF_2 には，録音環境に関する情報が多く含まれており，IF_2 を用いた学習では訓練データセットに対する過剰適合が生じてしまったのだと考えられる。

鳥の存在判定課題では，GD_1 の付与が有意な性能改善をもたらした。楽器の種類識別課題に関する考察で言及したように，GD_1 が，鳥の鳴き声のオンセット検出に役立った可能性があると考えられる。その一方で，IF_2 の付与は有意な性能低下をもたらした。DCASE2018 は，3 つの独立したデータセットを組み合わせて構築されたデータセットであり，本実験では，その内 2 つを訓練データセットに，残りの 1 つを検証データセットに割り当てた。このことから，言語識別課題の時と同じように，IF_2 の付与による有意な性能低下は，録音条件に対する過剰適合によって引き起こされた可能性がある。実際，訓練データセットに割り当てたデータセットの一つである BirdVox は，その音データの多くがハムノイズを含んでおり，他のデータセットとは録音環境が異なることが確認できた。

話者識別課題では，PHS_2 と IF_2，GD_1 の付与が有意な性能改善をもたらした。但し，話者識別課題に対する正解率は 30% 台とかなり低く，データセットのサイズが大きいために 95% 信頼区間が狭いため，この有意な性能向上は　ある程度偶発的にもたらされた可能性がある。

音環境分類課題，単語認識課題，感情音声分類課題では，位相特徴量の付与は有意な性能改善をもたらさなかった。これらの課題に対しては，位相情報があまり有効ではなかったか，データセットのサイズが小さいために有意差が出なかった可能性が考えられる。

6.4.3 本章の研究の限界

本章の実験では，メル尺度のような周波数軸を持つ位相情報の時間周波数表現について検討するために，拡張 LEAF を提案したが，拡張 LEAF には下記に示す課題がある。

- 拡張 LEAF では，位相微分を計算する際に近似として差分を用いるが，差分は必ずしも微分の近似にはならない。特に，位相の周波数方向の差分は，フィルタバンクの中心周波数の間隔に依存するため，群遅延の近似としては正確性に欠ける。
- 周波数ビンの総数 M を大きくする場合，隣接するフィルタ同士の通過帯域の重なりを減らすためには，フィルタ長 W も大きくする必要がある。また，Gabor フィルタバンクを使った畳み込みのシフト長（ストライド）は 1 sample であるため，周波数ビンの総数 M を増やすと計算コストが大幅に増加する。なお，この課題はオリジナルの LEAF と拡張 LEAF の両方に共通する。
- 位相情報に関する処理の多くをダウンサンプリングの前に行うため，位相情報のための計算コストが大きい。

また，Gabor フィルタバンクの中心周波数は，学習後も値が大きくは変化しなかった（図 6.3）。以上を踏まえると，フィルタバンクの中心周波数が学習可能であることは大きな長所ではなく，それよりも，差分計算が微分の正確な近似にはならない問題や，計算コストが膨大になる問題を解消することが重要であると考えられた。

6.5　まとめ

本章の実験では，8 種類の音分類課題に対して位相情報の有効性を検証した。実験の結果，5 種類の音分類課題に対して，圧縮パワーのみを用いた場合と比較して，位相フェーザや瞬時周波数，群遅延を用いることで性能向上がもたらされた。また，位相フェーザが効果的であった課題に対しては，瞬時周波数や群遅延も効果的であった。この結果は，出力目標が音声波形ではない回帰や分類のような課題では，位相の値それ自体よりも，微分で表現されるような位相の隣接関係の方が重要であることを示唆する。その他に，実験結果は，2 次元 CNN への入力として使う位相微分の表現としては，時間方向や周波数方向にバイアスが存在しない表現が効果的であることを示唆した。その一方で，提案した拡張 LEAF には，計算コストや，位相微分の計算の正確性に関する問題があった。そのため，次の課題としては，周波数軸がメル尺度である位相情報の時間周波数表現を計算するための，拡張 LEAF 以外のアプローチを考える必要があると考えられた。

第 7 章

位相微分を用いた小規模なデータセットに基づくGRBAS 尺度の自動評価

7.1 はじめに

前章では，8 種類の音分類課題に対して位相情報の有効性を検証し，5 種類の課題に対して，瞬時周波数と群遅延の両方，あるいはどちらか一方が有効であることを示した。また，実験結果から，2 次元 CNN への入力として使う位相微分の表現としては，時間方向や周波数方向にバイアスが存在しない表現が効果的であることが示唆された。

本章では，前章の実験で得られた上記の知見を活かしつつ，再び，聴覚心理的評価の評点の自動推定という課題に立ち返って実施した実験について述べる。本章の実験では，瞬時周波数と群遅延の有効性をさらに検討することに加えて，病的音声データセットが小規模な場合における，推定精度向上のための手段に関して検討することを目的とした。具体的には, 8 名の専門家によって評価された 300 サンプルの病的音声データを元に, 振幅, 瞬時周波数，および群遅延から得られる時間周波数表現を用いて，GRBAS 尺度の評点の自動推定実験を行った。本章では，周波数軸がメル尺度である位相情報の時間周波数表現を計算するために，計算効率と正確性の両方に関して拡張 LEAF よりも優れた手法を提案する。また，この実験では，サンプル数の少なさを補うデータ拡張（data augmentation）に関する調査も行った。データ拡張手法としては，音声の再生速度をランダムに変化させるランダム速度摂動（random speed perturbation） [237]，音声波形をランダムな位置で切り出すランダムクロップ（random crop） [49]，時間周波数表現の一部の周波数帯域の値を隠して無効にする周波数マスキング（frequency masking） [234] の 3 つの手法を検討した。

本章の実験で作成したプログラムコードは GitHub のリポジトリ[*1]に公開している。

7.2　手法

本節では，GRBAS 尺度の評点が付与された病的音声データセットと，評点の自動推定を行う DNN の構築について説明する。

7.2.1　データセット

本章の実験では，音声を専門とした臨床の経験年数が 2 年以上である専門家 8 名（耳鼻咽喉科医 6 名，言語聴覚士 2 名）によって GRBAS 尺度の評点が付与された，300 サンプルの持続母音/a/の音声データを用いた。評者内信頼性を調べるために 2 回評価された 50 サンプル（女性：26 サンプル，26 名，男性：24 サンプル，24 名）の音声データをテストデータセットに，残りの 250 サンプル（女性：107 サンプル，107 名，男性：143 サンプル，143 名）の音声データを訓練・検証データセットに割り当てた。なお，GRBAS 尺度の評点の付与や音声データの選定方法等の詳細に関しては第 4 章で述べた。本章では，これ以降，8 名の評者によって付与された GRBAS 尺度の評点の平均値を与える架空の評者を「平均評者（the average rater）」と呼ぶことにする。特に，小数点以下を保持したまま平均値を実数で与える評者を「連続平均評者（the continuous average rater）」，平均値を四捨五入して整数で与える評者を「離散平均評者（the discrete average rater）」と呼ぶことにする。図 7.1 に，「平均評者」が与える GRBAS 尺度の評点の分布を示す。

また，専門家 8 名の評者間信頼性と評者内信頼性は表 7.1 のようになった。信頼性の大きさは，信頼性指標である二次重み付け Cohen のカッパ係数で定量化した。カッパ係数の値は最大で 1 であり，値が大きいほど信頼性が高いことを表す。ある評者の評者間信頼性は，その評者と「離散平均評者」の間の信頼性として計算し，各評者の評者内信頼性は，各評者によって 2 回ずつ評価されたテストデータセットの音声データに対してのみ計算した。テストデータセットの 50 サンプルの音声データに関して，GRBAS 尺度のすべての項目で一貫して，評者内信頼性のカッパ係数の値が，評者間信頼性のカッパ係数の値より大きかった。

DNN は，連続値で評点を与えるように設計し，「連続平均評者」を模倣するよう回帰的に訓練した（詳細は第 7.2.2 節）。そのため，基本的に DNN の性能評価には，DNN が与える連続値の推定評点と「連続平均評者」の評点から算出される評者間信頼性の指標値を用いた。しかし，実際には，人間の専門家は評点を離散尺度（0，1，2，3）で与える。そのため，DNN の推定精度と専門家の信頼性を比較する際は，DNN の性能として，整数

[*1] https://github.com/onkyo14taro/auto-grbas

表 7.1 専門家 8 名の評者間信頼性と評者内信頼性に対応する二次重み付け Cohen のカッパ係数。カッパ係数の値は「平均 ± 標準偏差」で表している。ある評者の評者間信頼性は，その評者と「離散平均評者」の間の信頼性として計算した。テストデータセットの音声データは各評者によって 2 回ずつ評価されたが，テストデータセットの評者間信頼性を計算する際は，各評者の評点として，2 回の評価の平均を整数値に丸めたものを用いた。

	訓練・検証 (250 サンプル)	テスト (50 サンプル)	
項目	評者間	評者間	評者内
G	0.789 ± 0.048	0.744 ± 0.068	0.803 ± 0.086
R	0.599 ± 0.113	0.547 ± 0.109	0.727 ± 0.113
B	0.792 ± 0.088	0.730 ± 0.085	0.780 ± 0.055
A	0.472 ± 0.172	0.465 ± 0.157	0.528 ± 0.203
S	0.613 ± 0.096	0.674 ± 0.058	0.689 ± 0.080

値に丸めた DNN の推定評点と「離散平均評者」の評点から算出される評者間信頼性の指標値を用いた。なお，この比較の際は，専門家の評者間・評者内信頼性を表す指標値として，テストデータセットの 50 サンプルに対する二次重み付け Cohen のカッパ係数の値の結果（表 7.1 の右側 2 列）を用いた。

7.2.2 DNN の構造

図 7.2 に，本実験で用いた DNN の全体図を示す。DNN は，音声波形/a/のオンセットからオフセットまでを入力として受け取り，「連続平均評者」と同じ GRBAS 尺度の評点を推定するように設計した。なお，1 つの DNN は，GRBAS 尺度の 1 つの項目の評点を推定するため，GRBAS 尺度の各項目につき 1 つずつ DNN を作成した。オーディオフロントエンドでは，データ拡張と，持続母音/a/の可変長の音声波形から時間周波数表現への変換を行った。続く CRNN エンコーダでは，可変長の時間周波数表現を固定長の隠れ表現に変換した。そして，最後の MLP ヘッドでは，GRBAS 尺度の評点の離散確率分布を出力した。本節では，DNN の各構成要素に対して詳細な説明を行う。なお，DNN の構築と評価には機械学習フレームワークである PyTorch [222] を用いた。

オーディオフロントエンド

図 7.3 にオーディオフロントエンドの構造を示す。オーディオフロントエンドは，持続母音/a/の可変長の波形を入力として受け取って，時間周波数表現の計算を行うことに

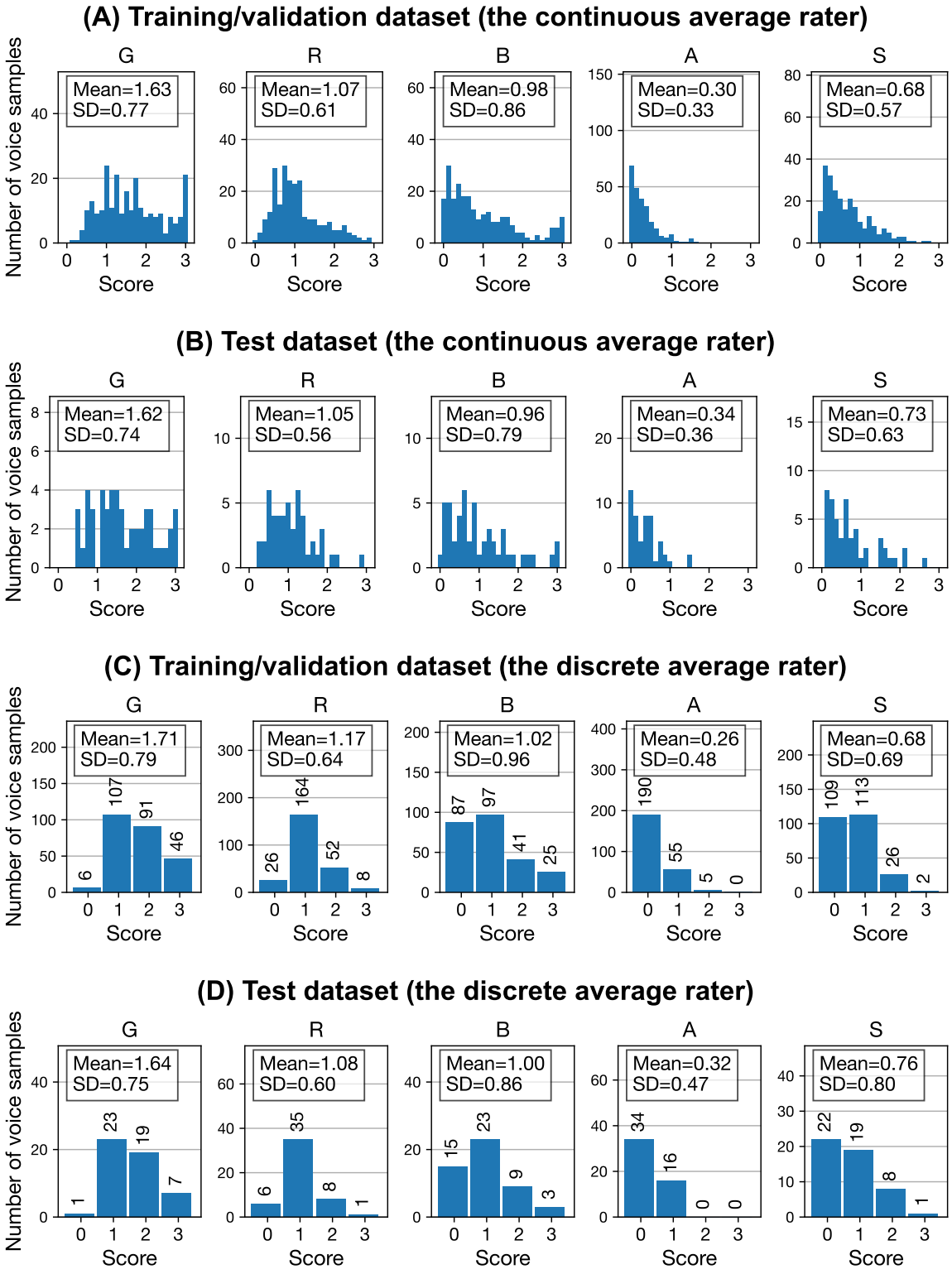


図 7.1 データセットの GRBAS 尺度の評点の分布。(A) は訓練・検証データセット（「連続平均評者」），(B) はテストデータセット（「連続平均評者」），(C) は訓練・検証データセット（「離散平均評者」），(D) はテストデータセット（「離散平均評者」）の分布。

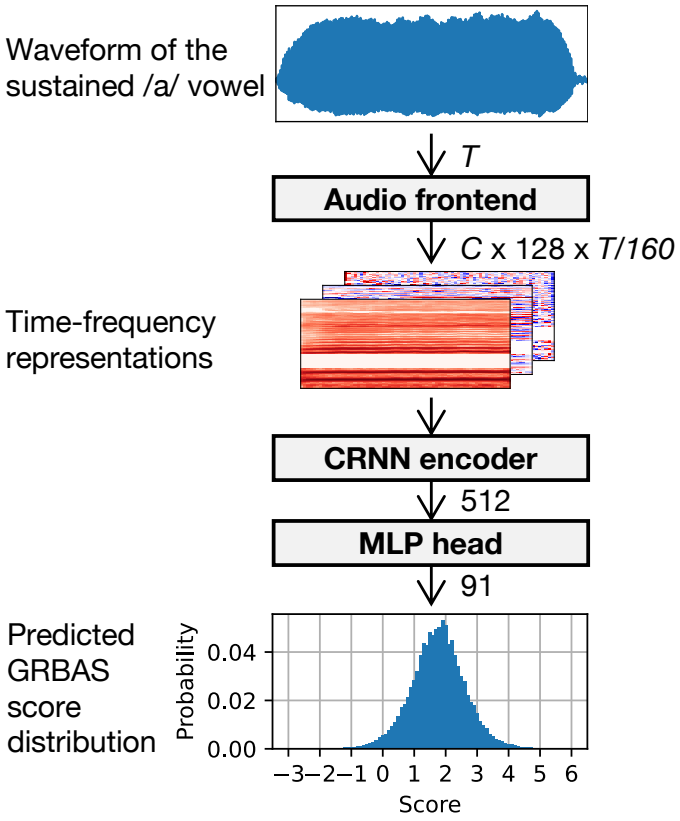


図 7.2 DNN の全体図。矢印の隣にある数値は出力テンソルの次元を表す。T は入力波形の時間長 [sample]，C は時間周波数表現の数を表す。

加えて，データ拡張も行った。オーディオフロントエンドが計算した時間周波数表現は，パワースペクトログラム，瞬時周波数，群遅延の 3 種類である。第 2.2.6 節でも説明したように，パワーが 0 の要素で位相の微分は特異値になるという問題がある [132, 143]。たとえパワーが 0 ではなくとも，その値がかなり小さい要素では，位相の微分は発散する。オーディオフロントエンドでは，後述する工夫によって，特異値の影響を取り除いた。

まず初めに，オーディオフロントエンドでは音声波形に対する前処理を行った。前処理は，振幅正規化，ランダム速度摂動 [237]，ランダムクロップ [238]，フェード処理という順で行った。なお，ランダム速度摂動とランダムクロップはデータ拡張手法であり，DNN の訓練時にのみ適用した。振幅は，平均が 0，標準偏差が 1 になるように正規化を行った。速度は，ランダムな比率でリサンプリングによって時間方向に一様な摂動を与えた。ランダムクロップは，元々の長さに対してランダムな比率になるように，波形の始めと終わりを切り取ることによって行った。切り取りの開始位置と終了位置もまたランダムに選択した。フェード処理は，波形の最初の 10 ms と最後の 10 ms がそれぞれ線形増加・減少するように適用し，振幅が急激に変化するのを抑制した。

音声波形は前処理した後で，周波数軸がメル尺度の時間周波数表現に変換した。ここで，第 2.2 節で説明したが，$\boldsymbol{w} \in \mathbb{R}^W (W \in \mathbb{N})$ を窓関数とする離散信号 $\boldsymbol{x} \in \mathbb{R}^X (X \in \mathbb{N})$

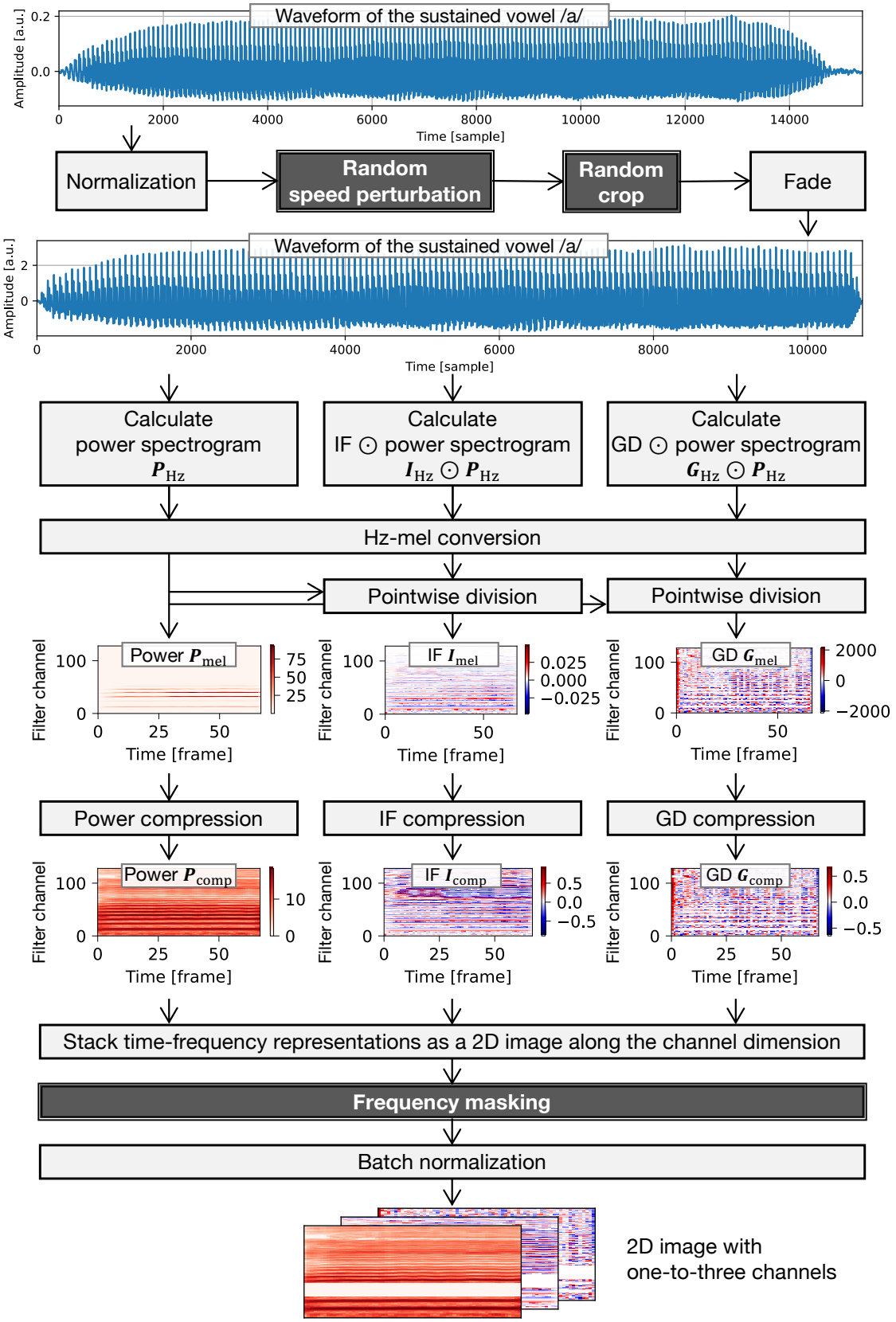


図 7.3 オーディオフロントエンドの全体図。この図では，パワー，瞬時周波数（IF），群遅延（GD）のすべての時間周波数表現の計算パスが描かれているが，いずれか 1，2 種類の時間周波数表現を出力することも可能である。

の短時間フーリエ変換 $\mathcal{F}^{(\boldsymbol{w},a,D)}\boldsymbol{x} \in \mathbb{C}^{M\times N}(M,N\in\mathbb{N})$ は，以下のように定義される（式 (2.11) の再掲）。

$$\left(\mathcal{F}^{(\boldsymbol{w},a,D)}\boldsymbol{x}\right)[m,n] := \sum_{l=an-\lfloor W/2\rfloor}^{an+\lfloor (W-1)/2\rfloor} \boldsymbol{x}[l]\boldsymbol{w}[l-an]\exp\left(-\frac{2\pi i m l}{D}\right) \tag{7.1}$$

ここで，窓関数の添字集合は $\{-\lfloor W/2\rfloor, -\lfloor W/2\rfloor+1,\ldots,-1,0,1,\ldots,\lfloor (W-1)/2\rfloor\}$ で与えられ，$i=\sqrt{-1}$，$m\in\{0,1,\ldots,M-1\}$ は周波数の添字，$n\in\{0,1,\ldots,N-1\}$ は時間の添字，$a\in\mathbb{N}$ はシフト長，$D\in\mathbb{N}$ $(D\geq W)$ は DFT 点数を表す。この短時間フーリエ変換の定義を用いて，パワースペクトログラム $\boldsymbol{P}_{\mathrm{Hz}}\in\mathbb{R}_{\geq 0}^{M\times N}$，瞬時周波数 $\boldsymbol{I}_{\mathrm{Hz}}\in\mathbb{R}^{M\times N}$（式 (2.35) の相対瞬時周波数 $\boldsymbol{I}_2$ に対応），群遅延 $\boldsymbol{G}_{\mathrm{Hz}}\in\mathbb{R}^{M\times N}$（式 (2.36) の群遅延 $\boldsymbol{G}_1$ に対応）を計算した。

$$\boldsymbol{P}_{\mathrm{Hz}}[m,n] = \left|(\mathcal{F}^{(\boldsymbol{w},a,D)}\boldsymbol{x})[m,n]\right| \tag{7.2}$$

$$(\boldsymbol{I}_{\mathrm{Hz}}\odot\boldsymbol{P}_{\mathrm{Hz}})[m,n] = -\operatorname{Im}\left((\mathcal{F}^{(\mathcal{D}\boldsymbol{w},a,D)}\boldsymbol{x})[m,n]\overline{(\mathcal{F}^{(\boldsymbol{w},a,D)}\boldsymbol{x})[m,n]}\right) \tag{7.3}$$

$$(\boldsymbol{G}_{\mathrm{Hz}}\odot\boldsymbol{P}_{\mathrm{Hz}})[m,n] = \operatorname{Re}\left((\mathcal{F}^{(\mathcal{T}\boldsymbol{w},a,D)}\boldsymbol{x})[m,n]\overline{(\mathcal{F}^{(\boldsymbol{w},a,D)}\boldsymbol{x})[m,n]}\right) \tag{7.4}$$

ここで，$\odot$ は要素ごとの積（Hadamard 積），$\mathcal{D}\boldsymbol{w}$ は $\boldsymbol{w}$ を微分したもの，$\left(\mathcal{T}\boldsymbol{w}\right)[n] := n\boldsymbol{w}[n]$ である。なお，ここでは，瞬時周波数と群遅延は，敢えてパワースペクトログラムと乗算した表現で求めた。

$K\in\mathbb{N}$ チャンネルのメルフィルタバンク $\boldsymbol{F}\in\mathbb{R}_{\geq 0}^{K\times M}$ が与えられたとき，メル周波数パワースペクトログラム $\boldsymbol{P}_{\mathrm{mel}}\in\mathbb{R}_{>0}^{K\times N}$ は次のように計算した。

$$\boldsymbol{P}_{\mathrm{mel}}[m,n] = \max((\boldsymbol{F}\boldsymbol{P}_{\mathrm{Hz}})[m,n],\theta) \tag{7.5}$$

メル周波数パワースペクトログラムの計算の際，閾値 $\theta>0$ を設定した。閾値 θ は，対数パワーが負の無限大に発散するのを抑制するのに加え，後述する位相微分が発散するのも抑制する。メル周波数瞬時周波数 $\boldsymbol{I}_{\mathrm{mel}}\in\mathbb{R}^{K\times N}$ は，次のように計算した（メル周波数群遅延 $\boldsymbol{G}_{\mathrm{mel}}$ に関しても同様に計算した）。

$$\boldsymbol{I}_{\mathrm{mel}}[m,n] = \frac{(\boldsymbol{F}(\boldsymbol{I}_{\mathrm{Hz}}\odot\boldsymbol{P}_{\mathrm{Hz}}))[m,n]}{\boldsymbol{P}_{\mathrm{mel}}[m,n]} \tag{7.6}$$

$\boldsymbol{P}_{\mathrm{mel}}[m,n]$ は閾値 θ が設定されているために必ず正であるため，$\boldsymbol{I}_{\mathrm{mel}}[m,n], \boldsymbol{G}_{\mathrm{mel}}[m,n]$ は特異値にはならない。また，式 (7.6) は，メルフィルタバンクを平滑化カーネルと見做すことによって，位相微分のパワー加重平均として解釈することができる。

$$\boldsymbol{I}_{\mathrm{mel}}[m,n] = \frac{\sum_{l=0}^{M-1}(\boldsymbol{F}[m,l]\boldsymbol{P}_{\mathrm{Hz}}[l,n]\boldsymbol{I}_{\mathrm{Hz}}[l,n])}{\sum_{l=0}^{M-1}(\boldsymbol{F}[m,l]\boldsymbol{P}_{\mathrm{Hz}}[l,n])} \tag{7.7}$$

見易さのため，上式では閾値 θ を省いた表現を用いた。パワー加重平均は，位相微分の値の絶対値が過剰に大きくなるのを抑制する効果がある [132]。

周波数軸をメル尺度に変換した後で，値の圧縮を行った。圧縮パワー $\boldsymbol{P}_{\mathrm{comp}} \in \mathbb{R}_{\geq 0}^{K \times N}$ は，$\boldsymbol{P}_{\mathrm{mel}}$ の自然対数をとることによって計算した。

$$\boldsymbol{P}_{\mathrm{comp}}[m,n] = \ln(\boldsymbol{P}_{\mathrm{mel}}[m,n]) - \ln\theta \tag{7.8}$$

式 (7.5) と式 (7.8) より明らかなように，$\boldsymbol{P}_{\mathrm{comp}}$ は非負である。位相微分の圧縮には，Batch Normalization [180] に触発された手法を用いた。様々な時間長を持つメル周波数瞬時周波数のミニバッチ $\{\boldsymbol{I}_{\mathrm{mel},b} \in \mathbb{R}^{K \times N_b}\}_{b=0}^{B-1}$ が与えられたとき，第 b 番目の圧縮瞬時周波数 $\boldsymbol{I}_{\mathrm{comp,b}} \in \mathbb{R}^{K \times N_b}$ は次のように計算した（圧縮群遅延 $\boldsymbol{G}_{\mathrm{comp,b}}$ に関しても同様に計算した)。

$$\boldsymbol{I}_{\mathrm{comp},b}[m,n] = \tanh\left(\frac{\boldsymbol{I}_{\mathrm{mel},b}[m,n]}{\rho[m]+\epsilon} \cdot 0.1\gamma[m]\right) \tag{7.9}$$

ここで，$B \in \mathbb{N}$ はバッチサイズ，$\gamma \in \mathbb{R}^K$ は学習パラメータ，$\epsilon > 0$ はゼロ除算を防ぐ微小値であり，$\rho \in \mathbb{R}^K$ は次のように定義される平均絶対値である。

$$\rho[m] = \frac{1}{B}\sum_{b=0}^{B-1}\left(\frac{1}{N_b}\sum_{n=0}^{N_b-1}\left|\boldsymbol{I}_{\mathrm{mel},b}[m,n]\right|\right) \tag{7.10}$$

Batch Normalization と比較して，位相微分の圧縮は以下の点で異なる。

- Batch Normalization では平均と標準偏差に関して正規化を行うが，位相微分の圧縮では平均絶対値に関して正規化を行った。平均に関する正規化を行わなかったのは，位相微分では 0 からの偏差が重要な意味を持つためである。このため，位相微分の圧縮においては，平行移動を行うバイアスの学習パラメータを省いた。また，標準偏差は外れ値の影響を強く受けるため，大きさを調整するための統計量として，標準偏差の代わりに平均絶対値を用いた。
- Batch Normalization はミニバッチサイズ内のテンソルの形状がすべて同じであることを仮定しているが，位相微分の圧縮では可変長入力を取り扱えるようにした。
- 位相微分の外れ値の影響を抑制するために，tanh 関数を導入した。

パワーと比較して，位相微分は外れ値を除いてダイナミックレンジが狭い。そのため，位相微分の圧縮では，tanh 関数を除いて対数関数のような非線形変換を用いなかった。

圧縮した時間周波数表現は，各表現を 1 チャンネルの 2 次元画像と見做してチャンネル方向に結合した。最終的に結合した時間周波数表現は，周波数マスキング [234] と Batch Normalization を適用した後で，後続の CRNN エンコーダに入力した。周波数マスキングはデータ拡張技術の一つであり，訓練時のみ適用した。周波数マスキングでは，ランダ

ムに選択した連続する周波数ビンに存在する要素の値を 0 にした。チャンネル方向に結合された時間周波数表現のミニバッチに関して，各サンプルごとに異なるマスクを用いたが，一つのサンプル内に含まれる複数の時間周波数表現の間では同じマスクを共有した。なお，Batch Normalization の定義は，可変幅の入力を取り扱うことができるように修正した。可変幅テンソルのバッチ $\{\boldsymbol{Z}_b \in \mathbb{R}^{C\times H\times W_b}\}_{b=0}^{B-1}$ が与えられたとき，第 b 番目の正規化テンソル $\tilde{\boldsymbol{Z}}_b \in \mathbb{R}^{C\times H\times W_b}$ は次のように計算した。

$$\tilde{\boldsymbol{Z}}_b[c,h,w] = \frac{\boldsymbol{Z}_b[c,h,w] - \mu[c]}{\sqrt{(\sigma[c])^2 + \epsilon}} \cdot \gamma[c] + \beta[c] \tag{7.11}$$

ここで，$C \in \mathbb{N}$ はチャンネル数，$H \in \mathbb{N}$ は高さ，$W_b \in \mathbb{N}$ は幅，$\gamma, \beta \in \mathbb{R}^C$ は学習パラメータ，$\epsilon > 0$ はゼロ除算を防ぐ微小値であり，$\mu, \sigma \in \mathbb{R}^C$ は次のように定義される修正平均と修正標準偏差である。

$$\mu[c] = \frac{1}{B}\sum_{b=0}^{B-1}\left(\frac{1}{HW_b}\sum_{h=0}^{H-1}\sum_{w=0}^{W_b-1}\boldsymbol{Z}_b[c,h,w]\right) \tag{7.12}$$

$$\sigma[c] = \sqrt{\frac{1}{B}\sum_{b=0}^{B-1}\left(\frac{1}{HW_b}\sum_{h=0}^{H-1}\sum_{w=0}^{W_b-1}(\boldsymbol{Z}_b[c,h,w] - \mu[c])^2\right)} \tag{7.13}$$

CRNN エンコーダ

CRNN エンコーダでは，可変長の時間周波数表現を固定長の隠れ表現テンソルに変換した。図 7.4 に CRNN エンコーダの全体図を示す。CRNN エンコーダに入力された時間周波数表現は，2 次元 CNN である EfficientNetV2-B0 特徴量抽出器によって局所的情報を抽出した後，周波数次元に関して平坦化した上で，RNN の一種である双方向 LSTM [191–193] によって固定長テンソルに変換した。EfficientNetV2 [239] は性能と計算効率を両立させた 2 次元 CNN ファミリーである。EfficientNetV2-B0 は，EfficientNetV2 の中で最も小規模な 2 次元 CNN である。本実験では，PyTorch Image Models version 0.5.4 [240] の EfficientNetV2-B0 の実装を参照した。但し，本実験で作成した EfficientNetV2-B0 特徴量抽出器は，オリジナルの EfficientNetV2-B0 と 2 つの点で異なる。一つは，オリジナルの EfficientNetV2-B0 の最後の畳み込み層，Batch Normalization，活性化関数，大域平均プーリング，全結合層を省略したことである。もう一つは，Batch Normalization の定義を，第 7.2.2 節で説明した可変幅入力を扱えるものに修正したことである。EfficientNetV2-B0 特徴量抽出器の詳細に関しては，付録 A.2 で説明した。

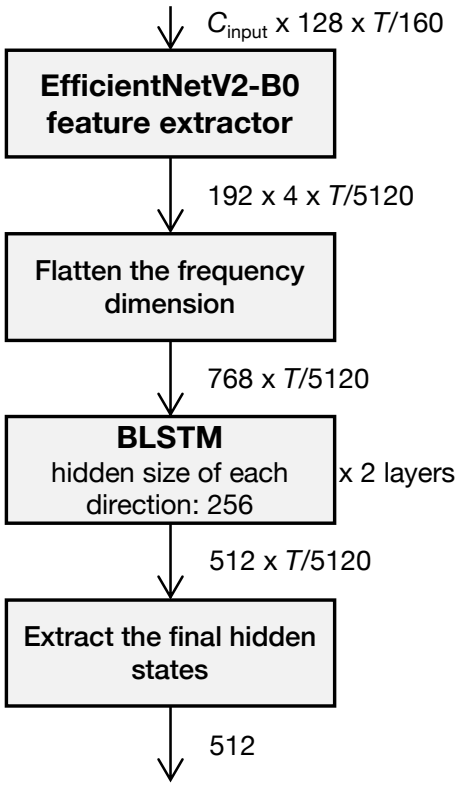


図 7.4 CRNN エンコーダの構造。矢印の隣にある数値は出力テンソルの次元を表す。T は入力波形の時間長 [sample]，C_{input} は入力チャンネル数を表す。

MLP ヘッド

図 7.5 に MLP ヘッドの構造を示す。MLP ヘッドでは，CRNN エンコーダによって出力された隠れ表現テンソルを使って，GRBAS 尺度の評点の離散確率分布を計算した。離散確率分布の値の範囲は $[-3, 6]$ とし，その範囲を 91 個の等幅なビンに分割した（1 つのビンの幅は 0.1）。なお，評点それ自体を出力するのではなく，評点の確率分布を出力するように設計したのは，Deep Label Distribution Learning（DLDL） [241] と呼ばれる手法を用いるためである。DLDL は，ラベルの曖昧性を利用することによって，少量の曖昧なラベルが付与されたデータセットを使ったときの，モデルの性能を向上させることを意図した手法である。

ここで，ラベルの種類の総数が $C \in \mathbb{N}$ である順序付きラベル集合 $S = \{s_c\}_{c=0}^{C-1}$ を考える。本実験では，順序付きラベル集合は GRBAS 尺度の評点 $S = \{-3.0 + 0.1c\}_{c=0}^{90}$ が対応する。ある音声サンプルのある GRBAS 尺度の項目について，複数の評者によって与えられた評点の平均と標準偏差がそれぞれ μ, σ であった場合，DLDL の正解の確率分布

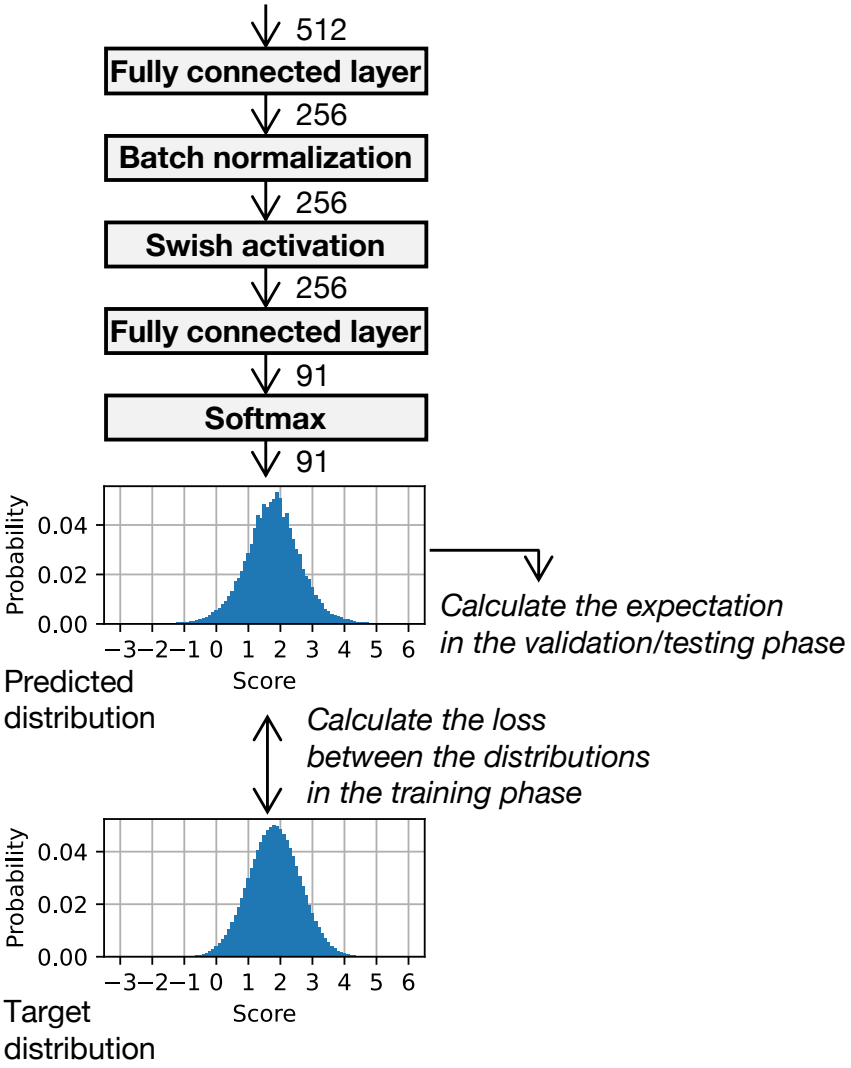


図 7.5 MLP ヘッドの構造。矢印の隣にある数値は出力テンソルの次元を表す。

$\boldsymbol{y} = \{y_c\}_{c=0}^{C-1}$ は，次式のような正規化ガウス分布として与えた。

$$y_c = \frac{p(s_c; \mu, \sigma)}{\sum_{k=0}^{C-1} p(s_k; \mu, \sigma)} \tag{7.14}$$

ここで，y_c はラベル s_c の確率質量，$p(s_c; \mu, \sigma)$ は，正規化前の s_c の確率質量を表す。

$$p(s_c; \mu, \sigma) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left(\frac{(s_c - \mu)^2}{2\sigma^2}\right) \tag{7.15}$$

本実験において，確率分布の値の範囲を $[0, 3]$ ではなく $[-3, 6]$ に設定したのは，正解となるガウス分布の裾が $[0, 3]$ をはみ出すことがあるのを考慮したためである（例えば，図 7.5 を参照）。

DNN は，予測分布と確率分布の間の差異を表す指標である Kullback–Leibler 情報量（Kullback–Leibler divergence） [242] を最小化するように学習を行った。DNN の出力

が $\hat{\boldsymbol{y}} = \{\hat{y}_c\}_{c=0}^{C-1}$ で与えられたとき，損失関数 $T(\boldsymbol{y}, \hat{\boldsymbol{y}})$ は次のように定義した。

$$T(\boldsymbol{y}, \hat{\boldsymbol{y}}) := -\sum_{c=0}^{C-1} y_c \ln \hat{y}_c \tag{7.16}$$

なお，正解となる確率分布の平均は，「連続平均評者」が与える評点に対応するため，DNN の学習目標は「連続平均評者」の近似に対応している。また，検証とテストの際は，DNN の予測評点 $\hat{s}$ は，出力確率分布の期待値として求めた。

$$\hat{s} = \sum_{c=0}^{C-1} \hat{y}_c s_c \tag{7.17}$$

7.3　実験

第 7.2 節で説明したデータセットと DNN を用いて，持続母音/a/の音声波形から GRBAS 尺度の評点を自動推定する実験を行った。実験では，パワー，瞬時周波数，群遅延という 3 種類の特徴量と，ランダム速度摂動，ランダムクロップ，周波数マスキングという 3 種類のデータ拡張手法に関して性能比較を行った。GRBAS 尺度の各項目，各実験条件につき独立した DNN を構築した。また，DNN の条件間の比較を行う前に，入力特徴量としてパワーを用いて，GRBAS 尺度の各項目についてデータ拡張手法の最適なパラメータを探索した。ランダム速度摂動の探索空間は 95〜105%，85〜115%，75〜125%，65〜135%，55〜145%，45〜155%，35〜165%，25〜175%，ランダムクロップの探索空間は 80〜100%，60〜100%，40〜100%，20〜100%，周波数マスキングの探索空間は 0〜25 ビン，0〜51 ビン，0〜77 ビン，0〜103 ビンとした。その後，パワーに対して最適化したパラメータを，全ての特徴量の組み合わせパターンに対して用いて比較実験を行った。DNN の性能評価は，5 分割交差検証によって行った。5 分割交差検証は，乱数シードを変えて 4 回行ったため，各実験条件の試行回数は 20 ($= 5 \times 4$) 回となった。訓練・検証用の持続母音/a/の 250 サンプルはランダムに 5 等分した。5 分割交差検証の各分割では，5 等分されたサブセットの内 4 つを訓練データセット，残り 1 つのサブセットを検証データセットとして用いた。

DNN の学習は，250 エポック連続して平均検証誤差が 0.01 以上改善しなかった時点で終了した。テスト時には，平均検証損失が最小となった時点のモデルのパラメータを用いた。DNN は Adam [164] を使用し，学習率は 0.0001 に設定した。バッチサイズは 50 とした。また，過剰適合を抑制するために，全結合層と畳み込み層，RNN の重みパラメータに l^2 正則化（正則化定数: 0.00004）を適用した。

オーディオフロントエンドに関して，入力音声波形の標本化周波数は 16 kHz に設定した。窓関数としては，l^2 ノルムが 1 に正規化された 50 ms（800 sample）の Hann 窓を用

いた。DFT 点数 D は 250 ms（4,000 sample），シフト長 a は 10 ms（160 sample）とした。メルフィルタバンクのチャンネル数 K は 128 とし，メル尺度の定義には HTK [147] の実装と同じものを用いた。$\boldsymbol{P}_{\mathrm{mel}}$ を計算する際の閾値 θ は 1×10^{-6} とした。位相微分の圧縮に関して，学習パラメータ γ の初期値は 1 に，微小値 ϵ は 1×10^{-5} に設定した。

7.4 結果

本節では，実験結果を説明する。これ以降，パワー，瞬時周波数，群遅延を用いた実験条件をそれぞれ POW，IF，GD と記す。同様に，ランダム速度摂動，ランダムクロップ，周波数マスキングを用いた実験条件をそれぞれ RS，RC，FM と記す。なお，特徴量の組み合わせやデータ拡張手法の組み合わせは，例えば POW+IF のように記す。特徴量に関しては，POW，IF，GD，POW+IF，POW+GD，IF+GD，POW+IF+GD の 7 条件，データ拡張手法に関しては，データ拡張なし，RS，RC，FM，RS+RC+FM の 5 条件，合わせて 35($= 7 \times 5$) 通りの条件で性能比較を行った。

7.4.1 性能評価のアプローチ

従来研究では，自動推定の性能を評価するために，二次重み付け Cohen のカッパ係数 [49]，Pearson の相関係数（PCC） [62]，平均絶対誤差（mean average error; MAE） [63]，一致率（accuracy; ACC） [49, 63] 等の指標が用いられてきた。これらすべての指標の結果を表示すると，膨大な量になってしまう。そのため，詳細な結果を提示する前に，性能評価を行うアプローチに関して議論する。

本実験では，DNN は「連続平均評者」を近似するように学習を行った。これは回帰問題に対応する。回帰性能の評価には，DNN によって予測された連続値の評点と，「連続平均評者」による評点を用いた。一方，分類性能の評価には，DNN によって予測された丸め評点と，「離散平均評者」による整数値の評点を用いた。図 7.6 に，テストデータセットに対する予測結果の散布図の例を示す。この図からわかるように，座標 $(0.5, 0.5)$ の近傍のような分類境界では，たとえ回帰誤差がかなり小さい場合にも，分類問題としては誤分類となってしまう場合がある。そのため，丸め誤差が分類性能の結果に与える影響を避けるために，DNN の条件間比較をより正確に行うには，回帰性能を基準にした方が良いと考えられる。他方，人間の専門家は評点を整数値で与えていたため，専門家の信頼性と DNN の性能を比較する際には，分類性能を基準にするべきであると考えられる。

本実験では，二次重み付け Cohen のカッパ係数 κ [215–217] を DNN の性能評価指標として用いた。二次重み付け Cohen のカッパ係数は通常，順序尺度で表されるデータの分類問題に使われるが，次式のように表現することで，直ちに回帰問題に適用することが

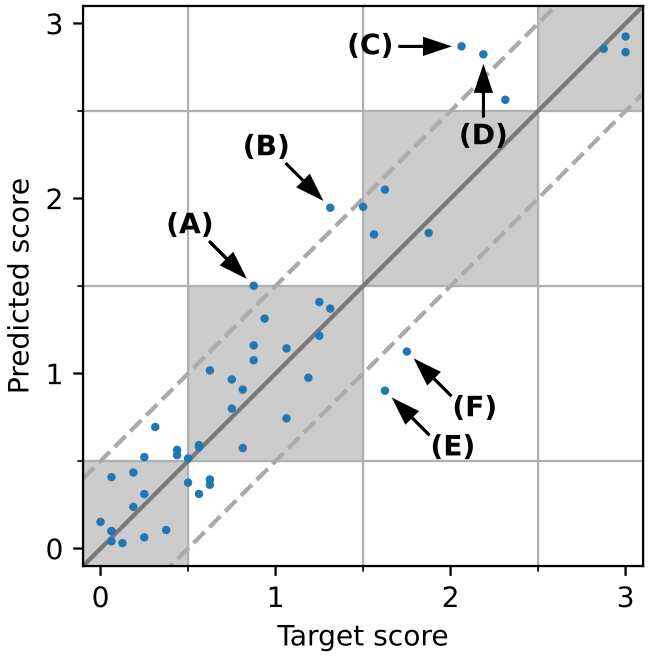


図 7.6 テストデータセットに対して得られた項目 B の予測評点の散布図。この結果は，特徴量条件を POW+IF+GD，データ拡張条件を RS+RC+FM とした時に，一つの乱数シード，5 分割交差検証の一つの分割に対して得られた。散布図の各点は音声サンプルに対応する。2 本のダッシュ線に挟まれた領域は，絶対予測誤差が 0.5 以下であることを表している。絶対予測誤差が 0.5 を超えた音声サンプルには (A)–(F) のアノテーションを付与した。アノテーションを付与したこれらの音声サンプルは，図 7.13 で参照している。評点を整数値に丸めたとき，各音声サンプルはグリッドで分割された正方形に割り当てられる。音声サンプルを表す点が灰色の正方形内にあるということは，丸められた正解スコアと予測スコアが分類の意味で一致することを表す。

できる。

$$\kappa = 1 - \frac{N \sum_{j=0}^{N-1} (x_j - y_j)^2}{\sum_{k=0}^{N-1} \sum_{l=0}^{N-1} (x_k - y_l)^2} \tag{7.18}$$

ここで，$N \in \mathbb{N}$ は評価されたサンプル数，x_j, y_j はそれぞれ「平均評者」と DNN によって第 j 番目のサンプルに与えられた評点を表す。二次重み付け Cohen のカッパ係数を用いた主な理由は，同じ基準，同じ指標で人間の専門家の評者と DNN を比較するためである。他の理由としては，二次重み付け Cohen のカッパ係数が，評点の誤差の大きさと，偶然の一致の両方を考慮した指標であることが挙げられる。本研究で用いたデータセットのように，分布が偏ったデータセットを使う際は，偶然の一致の影響を補正することが重要である。先述したように，二次重み付け Cohen のカッパ係数以外の指標として，分類問

題では PCC や MAE，分類問題では ACC のような指標が使われることもある。しかし，PCC は 2 変数が線形関係にあるかどうかのみに着目した指標であり，2 変数間の誤差の大きさは考慮しない。また，MAE は偶然の一致を考慮していない他，ACC は誤差の大きさも偶然の一致も考慮していない指標である。

これ以降，回帰性能と分類性能に関する二次重み付け Cohen のカッパ係数を，それぞれ $\kappa_{\mathrm{reg}}, \kappa_{\mathrm{class}}$ と記すことにする。κ_{reg} は DNN の条件間の性能比較，κ_{class} は DNN と人間の専門家の間の比較を行うために用いる。ここで，指標の違いが性能のランキングに及ぼす影響を提示することによって，指標の選択が結果の解釈に影響を与えるかどうかを検証する。表 7.2 に，性能ランキングに関する Spearman の順位相関行列を示す。なお，Cohen のカッパ係数に類似する指標である Krippendorff のアルファ係数 α [217, 243] も含めている。α の下付き文字は，κ の下付き文字と同様の意味を表す。Krippendorff のアルファ係数にはいくつか種類があるが，ここでは間隔尺度で表されるデータに対する定義を用いた。Cohen のカッパ係数と Krippendorff のアルファ係数の間の順位相関係数が最も高かった（分類に関しては 1.000，回帰に関しては 0.999）。全体的な傾向として，指標の種類間の差異よりも，回帰性能と分類性能間での差異が大きくなった（回帰性能を表す指標と，分類性能を表す指標の間の順位相関係数が小さくなった）。その一方で，回帰性能を表す指標間，分類性能を表す指標間では順位相関係数が比較的大きく，指標の選択が結果の解釈に与える影響は小さいと考えられる。このことから，すべての指標の結果を示すのは冗長であり，κ_{reg} と κ_{class} の結果を示せば十分であることが示唆された。よって，本章では，κ_{reg} と κ_{class} の結果をを示し，その結果に基づいて考察を行う。

表 7.2 性能ランキングに関する Spearman の順位相関行列。Spearman の順位相関係数の値は「平均 ± 標準偏差」で表している。なお，ここでは第 7.5.5 節で行った線形回帰と LightGBM のモデルを使った時の結果も含めた性能ランキングを用いた。GRBAS 尺度の各項目に対する順位相関係数を計算するために，725 パターン（DNN について 700 パターン，LightGBM について 20 パターン，線形回帰について 5 パターン）の結果から得られた性能指標の値を用いた。DNN についての 700 $(= 7 \times 5 \times 5 \times 4)$ パターンは，特徴量条件（7），データ拡張条件（5），5 分割交差検証（5），異なる乱数シードによる 4 回の試行（4）の全組み合わせである。LightGBM についての 20 $(= 5 \times 4)$ パターンは，5 分割交差検証（5）と異なる乱数シードによる 4 回の試行（4）の全組み合わせである。線形回帰についての 5 パターンは，5 分割交差検証の分割数に対応する。

	κ_{reg}	α_{reg}	PCC	MAE	κ_{class}	α_{class}	ACC
κ_{reg}	1.000 ± 0.000						
α_{reg}	1.000 ± 0.000	1.000 ± 0.000					
PCC	0.978 ± 0.016	0.976 ± 0.018	1.000 ± 0.000				
MAE	-0.961 ± 0.029	-0.960 ± 0.030	-0.970 ± 0.016	1.000 ± 0.000			
κ_{class}	0.899 ± 0.053	0.900 ± 0.052	0.869 ± 0.081	-0.868 ± 0.091	1.000 ± 0.000		
α_{class}	0.901 ± 0.051	0.902 ± 0.049	0.869 ± 0.083	-0.867 ± 0.093	0.999 ± 0.001	1.000 ± 0.000	
ACC	0.788 ± 0.091	0.788 ± 0.091	0.780 ± 0.092	-0.797 ± 0.104	0.915 ± 0.076	0.911 ± 0.073	1.000 ± 0.000

7.4.2 回帰性能：DNN の条件間の比較

図 7.7 に入力特徴量の条件を POW とし，データ拡張のパラメータ探索を行った時の κ_{reg} の結果を示す。各データ拡張のパラメータ値の違いよりも，データ拡張の種類の違いの方が κ_{reg} に大きな影響を与えた。パラメータ探索で得られた最適なパラメータは，POW 以外の特徴量条件を用いた後続の実験でも使用した。

図 7.8 に，すべての条件に関する κ_{reg} の結果を示す。まず，データ拡張がない場合の結果について説明する。項目 G，R，B，A について，IF と POW+IF の結果は POW の結果よりも良かった。また，項目 A と S について，GD と POW+GD の結果は POW の結果よりも良かった。その一方で，IF+GD と POW+IF+GD は，GRBAS 尺度の全項目で POW の結果よりも悪かった。次に，データ拡張がある場合の結果について説明する。全体的に，データ拡張条件の違いは，特徴量条件の違いよりも大きな影響を及ぼした。表 7.3 に，GRBAS 尺度の各項目についての，κ_{reg} の高さに関する第 10 位までのランキングを示す。GRBAS 尺度の全項目で，データ拡張がない条件は第 10 位までに入らなかった。データ拡張がない条件では，POW は GRBAS 尺度の全項目で第 3 位以下であったのとは対照的に，データ拡張がある条件では，POW は項目 A について第 1 位，項目 G（第 1 位より 0.004 低い），R（第 1 位より 0.012 低い），S（第 1 位より 0.001 低い）について第 2 位となった。その一方で，項目 B については POW は第 9 位（第 1 位より 0.024 低い）となった。項目 B について POW より上位であった特徴量条件は，POW+IF+GD，POW+GD，IF+GD，GD であった。

特徴量条件ごとに，効果的なデータ拡張手法は異なった。表 7.4 に，各特徴量条件におけるデータ拡張の導入が κ_{reg} に与える影響を示す。まず，単一のデータ拡張を導入したときの結果について説明する。すべての特徴量条件において，RS が最も有効なデータ拡張条件となった。POW 以外の特徴量条件では RC が二番目に有効であり，RC は特に群遅延を含む特徴量条件（GD，POW+GD，IF+GD，POW+IF+GD）で効果的であった。FM は POW に対しては二番目に有効であった一方で，位相微分を含む他の特徴量条件に対しては最も有効性が低かった。続いて，すべてのデータ拡張を用いた条件 RS+RC+FM について説明する。単一のデータ拡張を用いたときの最大の改善量と比較して，4 つの特徴量条件（POW，IF，GD，POW+GD）に関しては 0.007〜0.027 改善した一方で，残りの特徴量条件（POW+IF，IF+GD，POW+IF+GD）に関しては 0.001〜0.019 悪化した。

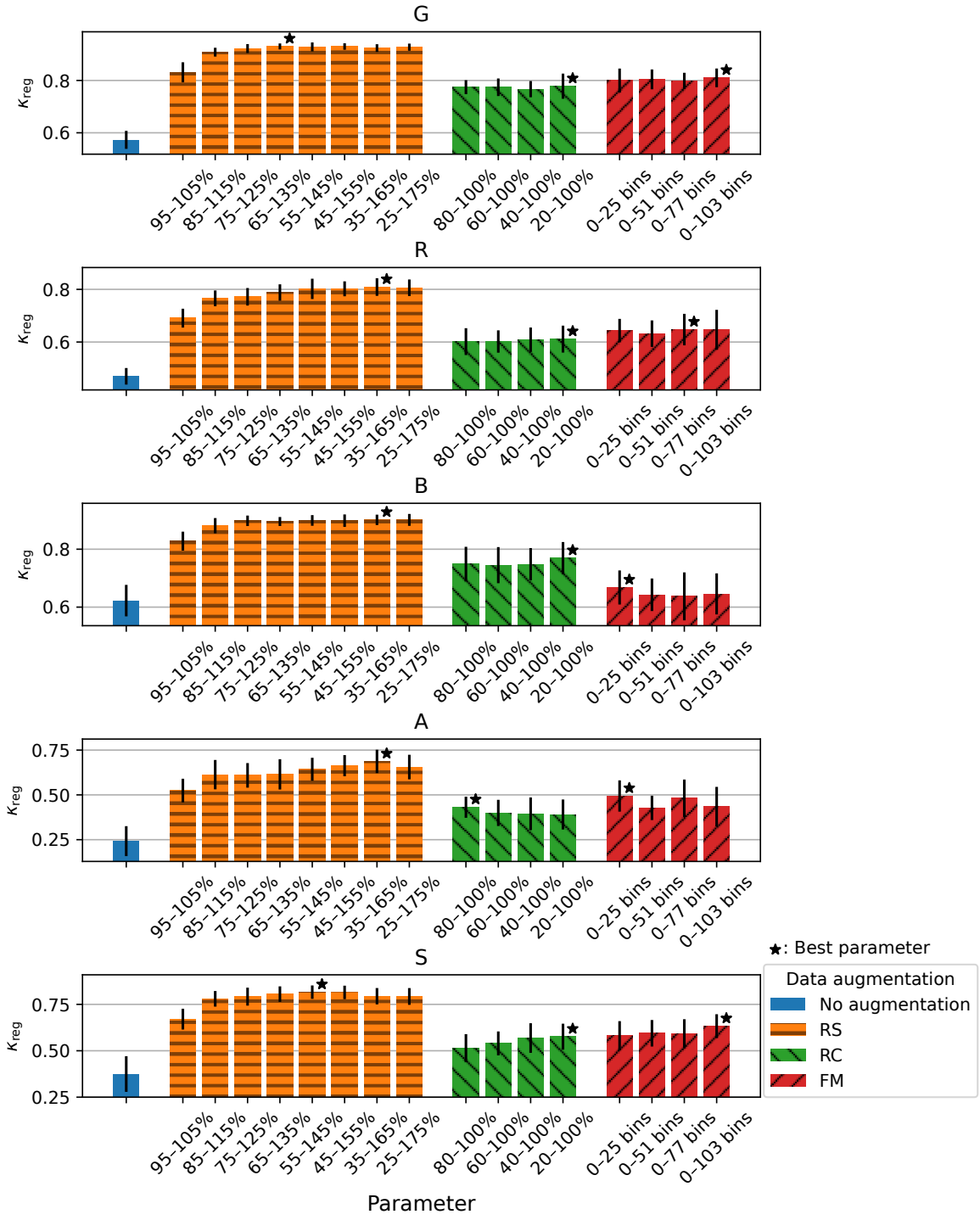


図 7.7 データ拡張のパラメータ探索の κ_{reg} の結果の棒グラフ。エラーバーは標準偏差を表す。この結果は，入力特徴量としてパワーを用い，4 つの異なる乱数シードを用いて 5 分割交差検証を行って得られたものである。GRBAS 尺度の各項目，データ拡張の各手法に対して，κ_{reg} が最大となったパラメータの結果を表す棒の傍に星印を付与した。

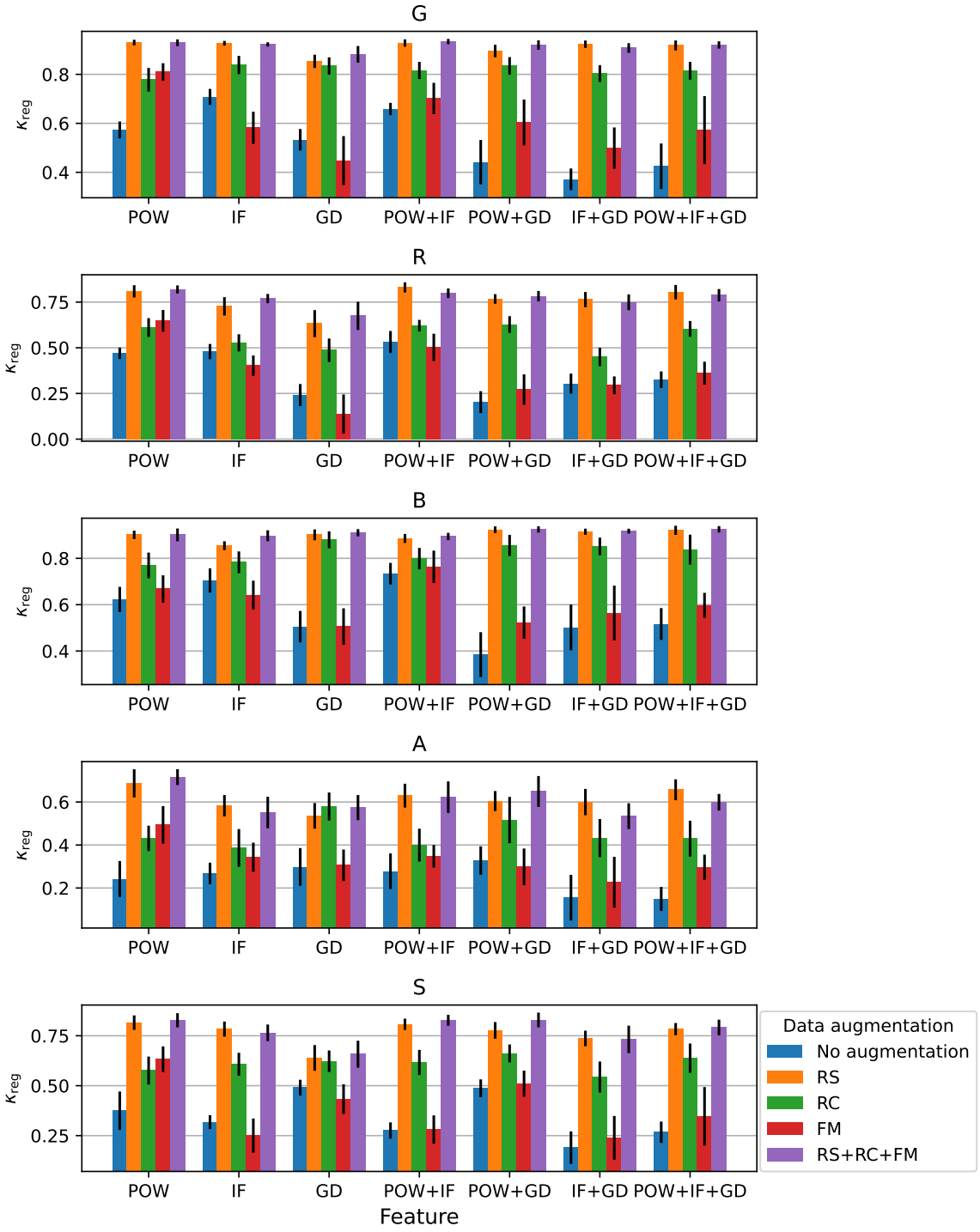


図 7.8 4 つの異なる乱数シードを用いて 5 分割交差検証を行って得られた κ_{reg} の結果の棒グラフ。エラーバーは標準偏差を表す。

表 7.3 κ_{reg} の高さに関する第 10 位までのランキング。κ_{reg} の値は「平均 ± 標準偏差」で表している。

	G			R			B		
順位	特徴量	データ拡張	κ_{reg}	特徴量	データ拡張	κ_{reg}	特徴量	データ拡張	κ_{reg}
1	POW+IF	RS+RC+FM	0.934 ± 0.011	POW+IF	RS	0.831 ± 0.028	POW+IF+GD	RS+RC+FM	0.925 ± 0.014
2	POW	RS	0.930 ± 0.011	POW	RS+RC+FM	0.819 ± 0.022	POW+GD	RS+RC+FM	0.925 ± 0.014
3	POW	RS+RC+FM	0.929 ± 0.014	POW	RS	0.809 ± 0.034	POW+GD	RS	0.923 ± 0.015
4	POW+IF	RS	0.928 ± 0.015	POW+IF+GD	RS	0.804 ± 0.039	POW+IF+GD	RS	0.921 ± 0.020
5	IF	RS	0.927 ± 0.010	POW+IF	RS+RC+FM	0.798 ± 0.026	IF+GD	RS+RC+FM	0.917 ± 0.010
6	IF+GD	RS	0.923 ± 0.015	POW+IF+GD	RS+RC+FM	0.787 ± 0.034	IF+GD	RS	0.915 ± 0.013
7	IF	RS+RC+FM	0.922 ± 0.009	POW+GD	RS+RC+FM	0.782 ± 0.029	GD	RS+RC+FM	0.910 ± 0.015
8	POW+IF+GD	RS+RC+FM	0.920 ± 0.015	IF	RS+RC+FM	0.769 ± 0.025	GD	RS	0.901 ± 0.023
9	POW+GD	RS+RC+FM	0.920 ± 0.019	POW+GD	RS	0.767 ± 0.026	POW	RS	0.901 ± 0.018
10	POW+IF+GD	RS	0.918 ± 0.020	IF+GD	RS	0.764 ± 0.042	POW	RS+RC+FM	0.901 ± 0.027

	A			S		
順位	特徴量	データ拡張	κ_{reg}	特徴量	データ拡張	κ_{reg}
1	POW	RS+RC+FM	0.716 ± 0.037	POW+GD	RS+RC+FM	0.829 ± 0.037
2	POW	RS	0.687 ± 0.065	POW	RS+RC+FM	0.828 ± 0.035
3	POW+IF+GD	RS	0.657 ± 0.049	POW+IF	RS+RC+FM	0.827 ± 0.027
4	POW+GD	RS+RC+FM	0.649 ± 0.072	POW	RS	0.816 ± 0.036
5	POW+IF	RS	0.630 ± 0.055	POW+IF	RS	0.807 ± 0.028
6	POW+IF	RS+RC+FM	0.622 ± 0.073	POW+IF+GD	RS+RC+FM	0.791 ± 0.039
7	POW+GD	RS	0.604 ± 0.047	IF	RS	0.783 ± 0.038
8	IF+GD	RS	0.600 ± 0.061	POW+IF+GD	RS	0.783 ± 0.031
9	POW+IF+GD	RS+RC+FM	0.599 ± 0.039	POW+GD	RS	0.777 ± 0.041
10	IF	RS	0.583 ± 0.050	IF	RS+RC+FM	0.765 ± 0.042

表 7.4 データ拡張の導入による κ_{reg} の改善量。κ_{reg} の改善量は「平均 ± 標準偏差」で表している。平均と標準偏差は，GRBAS 尺度の 5 項目と，5 分割交差検証の組み合わせである計 25 $(= 5 \times 5)$ パターンから計算した。各パターンの値としては，4 つの異なる乱数シードを用いた試行の平均値を用いた。

特徴量	RS	RC	FM	RS+RC+FM
POW	0.372 ± 0.071	0.177 ± 0.042	0.194 ± 0.085	0.382 ± 0.077
IF	0.279 ± 0.114	0.133 ± 0.089	-0.052 ± 0.077	0.286 ± 0.097
GD	0.299 ± 0.101	0.267 ± 0.088	-0.048 ± 0.054	0.326 ± 0.104
POW+IF	0.321 ± 0.129	0.155 ± 0.107	0.023 ± 0.042	0.320 ± 0.136
POW+GD	0.425 ± 0.128	0.330 ± 0.134	0.072 ± 0.083	0.452 ± 0.111
IF+GD	0.483 ± 0.063	0.312 ± 0.109	0.060 ± 0.058	0.464 ± 0.075
POW+IF+GD	0.480 ± 0.050	0.328 ± 0.065	0.098 ± 0.053	0.468 ± 0.052

7.4.3 分類性能：人間の専門家との比較

図 7.9 に，すべての条件に関する κ_{class} の結果を示す。なお，丸め込みの影響で，κ_{reg} と比較して，κ_{class} の平均値（棒の高さ）は低く，標準偏差（エラーバー）は大きくなり，順位が僅かに変動した。DNN の分類性能は，人間の専門家の評者内信頼性，および人間の専門家と「離散平均評者」の間の評者間信頼性と比較した。平均値に関して，人間の専門家と「離散平均評者」の間の評者間信頼性を超えた DNN の条件数は，項目 G について 14 条件，項目 R について 12 条件，項目 B について 16 条件，項目 A について 9 条件，項目 S について 10 条件となった。人間の専門家の評者内信頼性を超えた DNN の条件数は，項目 G について 13 条件，項目 R について 0 条件，項目 B について 13 条件，項目 A について 2 条件，項目 S について 8 条件となった。

7.5 考察

7.5.1 データ拡張

全体的に，ランダム速度摂動（RS）が最も有効なデータ拡張手法であった。速度摂動の有効性は音声認識において確認されており，健常音声の大規模データベース [237] のみならず，発語運動障害（dysarthria）の小規模データベース [244] に対しても有効性が実証されている。ランダムクロップ（RS）や周波数マスキング（FM）は音声データの情報の一部を欠損させる処理であるのに対し，ランダム速度摂動は音声データの多様性を水増しする処理である。再生速度の加速や減速（時間軸上の縮小と拡張）は，それぞれ周波数軸

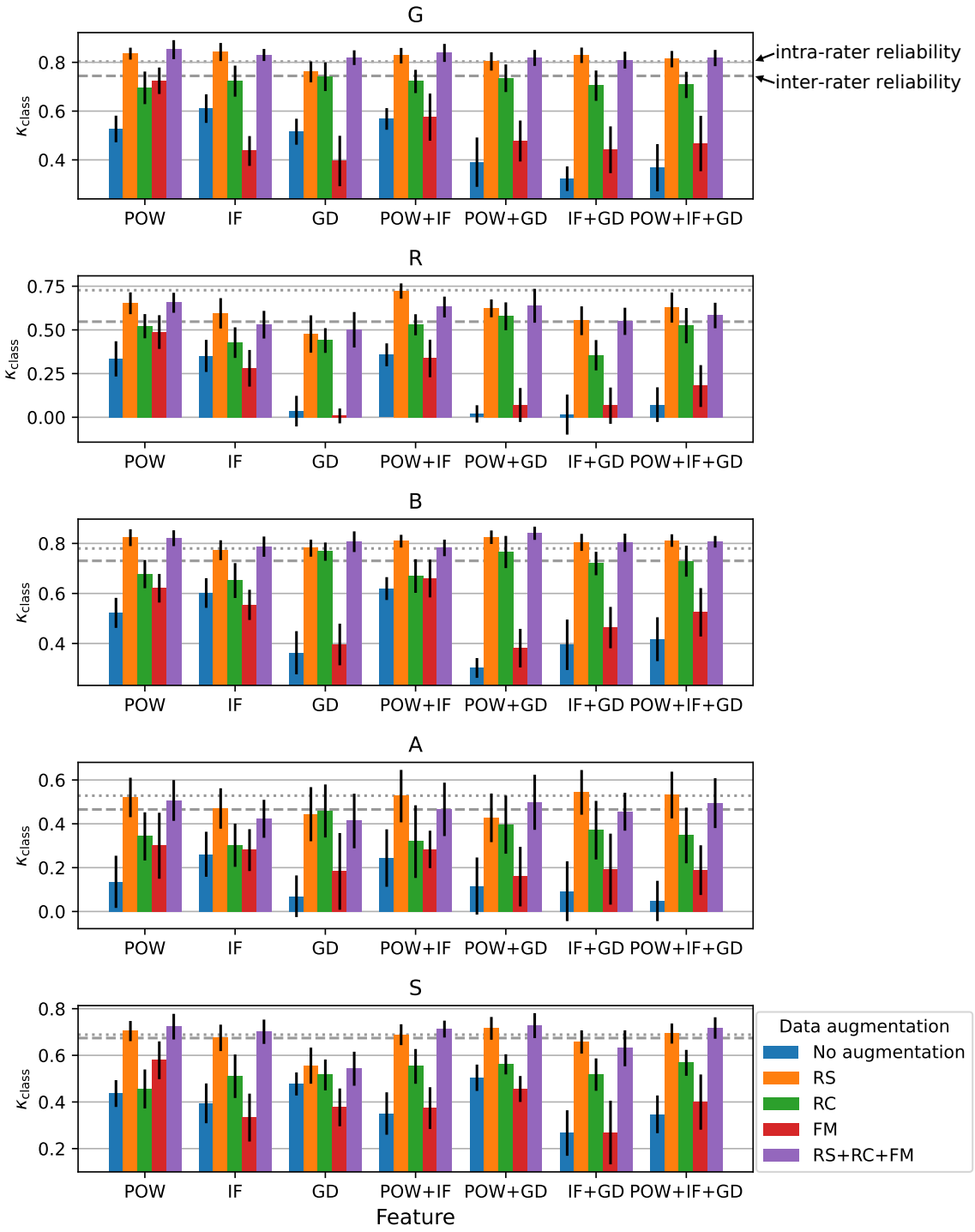


図 7.9 4 つの異なる乱数シードを用いて 5 分割交差検証を行って得られた κ_{class} の結果の棒グラフ。エラーバーは標準偏差を表す。点線は専門家の評者内信頼性の平均値，ダッシュ線は，専門家と「離散平均評者」の間の評者間信頼性の平均値を表す。

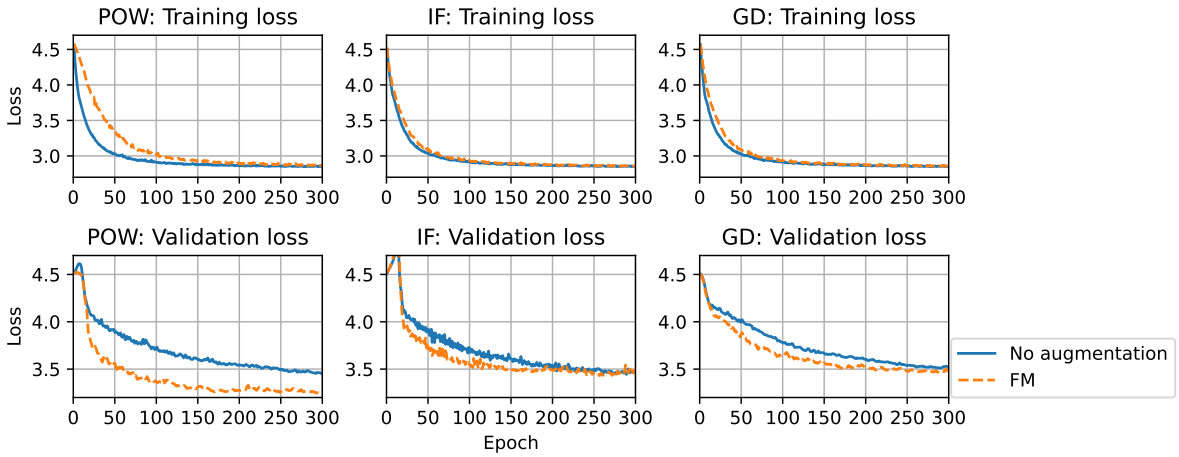


図 7.10 特徴量条件を POW，IF，GD，データ拡張条件をデータ拡張なし，あるいは FM とした時の学習曲線。この結果は，項目 B の学習時に，一つの乱数シード，5 分割交差検証の一つの分割に対して得られた。

上の一様な拡張と縮小に対応する。この拡張と縮小が，音声データの多様性の増加に貢献したと考えられる。

周波数マスキングは深層学習を用いた音響信号処理で頻用されるデータ拡張手法であるが，POW 以外の特徴量条件では性能向上に最も寄与しなかった。周波数マスキングは，DNN を情報の部分的欠損に対して頑健にするために提案された [234]。周波数マスキングは，訓練中に周波数情報を部分的に消すため，学習速度の低下を引き起こす。図 7.10 に，特徴量条件が POW，IF，GD であったときの学習曲線を，周波数マスキングの有無を比較して示す。訓練損失に関する図に着目すると，特徴量条件 IF と GD では，特徴量条件 POW よりも周波数マスキングによる学習速度の減速幅が小さかったことが観察される。加えて，特徴量条件 POW では，周波数マスキングが検証損失を低下させているのに対し，IF と GD では検証損失があまり低下していなかった。これらの結果は，位相微分を含む特徴量条件に関しては，特定の周波数帯域とは独立した情報が DNN の学習で用いられた可能性を示唆していると考えられる。

ランダムクロップ（データ拡張条件 RC）による性能向上幅は，群遅延を含まない特徴量条件（POW，IF，POW+IF）よりも，群遅延を含む特徴量条件（GD，POW+GD，IF+GD，POW+IF+GD）で大きかった。群遅延は，オンセットやオフセットで絶対値が大きくなるため，オンセット検出に応用されてきた [127–130]。図 7.11 に，同一波形からランダムに切り出された 3 つの群遅延 $\boldsymbol{G}_{\mathrm{comp}}$ を示す。この図から観察されるように，クロップの仕方によって，時間の始端と終端の様子が大きく異なることがわかる。この観察から，ランダムクロップは，群遅延を用いたときに DNN がオンセットやオフセットの

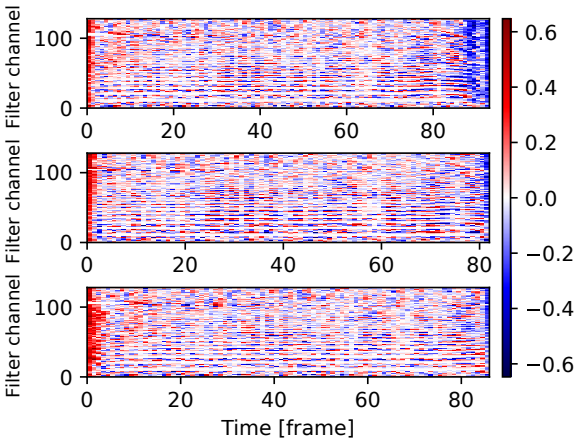


図 7.11 同一波形からランダムに切り出された 3 つの波形に対応する群遅延 $\boldsymbol{G}_{\mathrm{comp}}$ のヒートマップ

情報に過剰適合するのを抑制する効果をもたらしたのではないかと推察される。逆に言えば，データ拡張がない場合，群遅延を用いると，DNN がオンセットやオフセットの情報に過剰適合しやすくなる可能性があると考えられる。

データ拡張のパラメータ探索では，データ拡張条件 RC と FM よりも，データ拡張条件 RS でパラメータ値の違いが κ_{reg} の値に大きな影響を与えた。データ拡張条件 RS では，最も狭い場合で 65〜135%，最も広い場合で 35〜165% が最適な摂動幅となった。これらの摂動幅は，健常音声やディサースリアの音声認識 [234, 237] で有効な摂動幅 90〜110% よりもかなり広い。音声認識の分析対象である連続発声では発話テンポが重要であるため，過度な速度摂動（加速や減速）は音声を不自然にしてしまう。対照的に，持続母音では発話テンポが相対的に重要ではないため，速度摂動の程度が比較的大きくても，持続母音はあまり不自然にならない。

読者の中には，データ拡張手法が声質に与える影響が気になる方もいると思われる。速度が摂動された音声を筆者が実際に聴取した時に，たとえ速度摂動の程度が大きい場合でも，声質に対する影響はあまり大きくないように感じられた（例えば，重度の息漏れ声は，速度摂動後も重度の息漏れ声のままであった）。それゆえ，速度摂動が GRBAS 尺度の評点に与える影響はあまり大きくないと考えられる。加えて，データ拡張条件 RS では加速と減速の両方を行ったため，速度摂動がもたらす GRBAS 尺度の評点の増減は相殺されたのではないかと考えられる。他方，ランダムクロップ（データ拡張条件 RC）では持続母音のオンセットやオフセットが除外されたが，持続母音のオンセットやオフセットは僅かながら声質に関わるかもしれないと示唆する研究がある [245, 246]。実際にクロップされた音声を聴取した印象としては，クロップによる聴感的な声質の変化はほとんどないように感じられた。データ拡張条件 FM では，声質評価に関わる重要な周波数帯域の情報が

取り除かれてしまった可能性がある。例えば，声質評価のための音響特徴量の一つである離散フーリエ変換比 [43] は，低域のエネルギーと高域のエネルギーの比として定義される。データ拡張条件 RS+RC+FM は，必ずしも RS よりも良い結果を示さなかった。これは，すべてのデータ拡張手法を組み合わせることによって，本来の声質とデータ拡張手法的適用後の声質が過度に異なってしまった可能性があるのではないかと考えられる。

7.5.2 GRBAS 尺度の項目間の差異

特徴量との関係

データ拡張を用いなかった第 5 章の実験では，項目 G，R，B，S について，瞬時周波数に基づく特徴量を用いた時に，パワーに基づく特徴量を用いた時よりも高い推定精度を実現できた。本章の実験では，データ拡張を用いなかった場合，項目 G，R，B，A について，瞬時周波数を使った特徴量条件（IF と POW+IF）が，POW の推定精度を上回った。項目 A と S に関しては異なる結果となったものの，2 つの実験で概ね類似する傾向が得られた。しかし，データ拡張を伴う場合，特徴量条件が POW であった時に，κ_{reg} の値に関して，項目 A で第 1 位，項目 G，R，S で第 2 位となった。このことは，項目 G，R，A，S に関しては，データ拡張がある条件（RS または RS+RC+FM）では，入力特徴量としてはパワーのみが与えられたら十分であり，位相微分（瞬時周波数と群遅延）の追加は冗長で過剰適合に繋がったことを示唆している。

対照的に，項目 B では，特徴量条件 POW は第 9 位になった。項目 B の上位 8 位に入った条件は，すべて群遅延を含む条件（POW+IF+GD，POW+GD，IF+GD，GD）であった。気息性は，音声合成における非周期性指標（aperiodicity） [131] の概念と関連している。非周期性指標は，正弦波成分と非周期成分の比として定義される。気息性雑音成分は，非周期成分に対応する。非周期性指標の推定に関する研究では，瞬時周波数や群遅延の概念が応用されている [131–133]。特に，広く普及しているヴォコーダである WORLD [131, 247] では，非周期性指標の推定のために，瞬時周波数ではなく群遅延のみを使用している。このことは，非周期性指標に関連する項目 B の推定には，瞬時周波数よりも群遅延がさらに有用である可能性を示唆している。図 7.12 に，健常音声と息漏れ声から計算したパワー $\boldsymbol{P}_{\mathrm{comp}}$，瞬時周波数 $\boldsymbol{I}_{\mathrm{comp}}$，群遅延 $\boldsymbol{G}_{\mathrm{comp}}$ を示す。瞬時周波数では，健常音声に関しては調波成分に対して，息漏れ声に関してはフォルマントに対して特徴的なパターンが現れた。瞬時周波数と広帯域分析（短い窓関数）を組み合わせることによってフォルマント推定に応用した研究 [119] があるが，強い息漏れ声では調波成分が非周期成分に埋もれており，実質的に広帯域分析を行った時と近いパターンが現れたと考えられる。その一方で，瞬時周波数とは異なり，群遅延では調波成分に対してのみ特徴的なパターンが現れ，フォルマントを含む他の成分に対してはランダムなパターンが現れた。ま

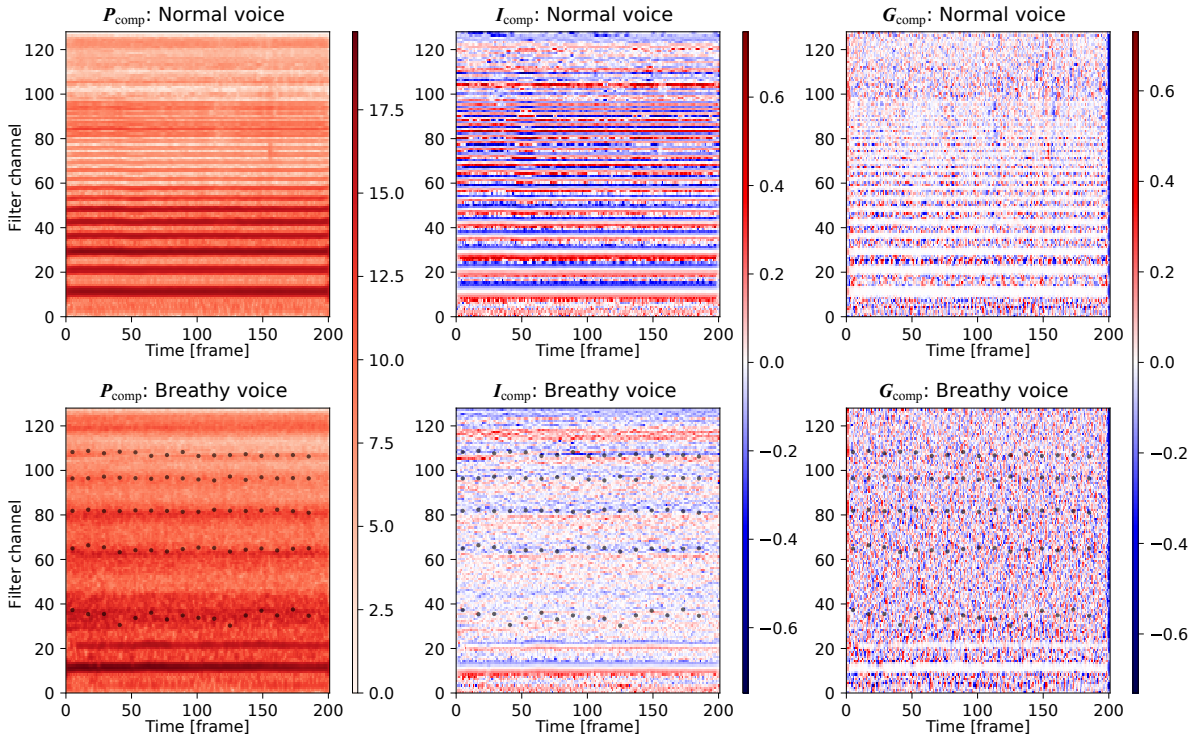


図 7.12 健常音声と息漏れ声に対して計算されたパワー $\boldsymbol{P}_{\mathrm{comp}}$，瞬時周波数 $\boldsymbol{I}_{\mathrm{comp}}$，群遅延 $\boldsymbol{G}_{\mathrm{comp}}$ のヒートマップ。息漏れ声のヒートマップ中には，Praat によって抽出されたフォルマントを点線で表示している。

とめると，パワーと瞬時周波数では，調波成分とフォルマントの両方に対して特徴的なパターンが現れたのに対し，群遅延では調波成分に対してのみ特徴的なパターンが現れた。群遅延のこの性質は，項目 B の評点を推定するのに効果的であると考えられる。群遅延と他の特徴量（パワーや瞬時周波数）を組み合わせることによって性能がさらに向上したのは，これらの特徴量が相補的な関係であったためであると推察される。

人間の専門家との比較

人間の専門家と「離散平均評者」の間の評者間信頼性を上回った DNN の条件の数は，項目 B で一番多く，項目 G で二番目に多かった。専門家の評者内信頼性を上回った DNN の条件の数についても，項目 G と B で一番多くなり，類似する傾向が得られた。GRBAS 尺度の他の項目と比較して，項目 G と B のデータセットの分布は偏りが小さかった（図 7.1）。分布が相対的に偏っていなかったことが，DNN の性能向上に寄与した可能性があると考えられる。

項目 R では，人間の評者内信頼性を上回った DNN の条件はなかった。項目 R は本質的に推定が難しい項目であった可能性もある一方で，別の可能性を考えることもできる。

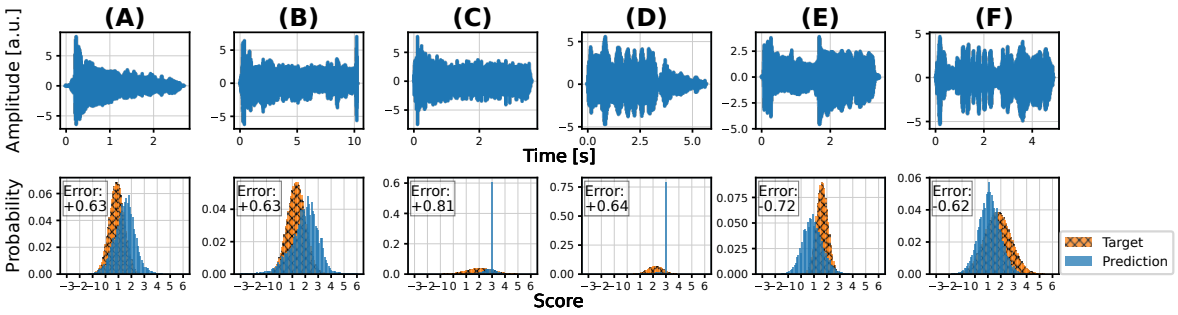


図 7.13 図 7.6 で絶対予測誤差が 0.5 以上であった音声サンプルの波形と確率分布。各サブ図の記号 (A)–(F) は図 7.6 の記号と対応する。各サブ図中の “Error” は，目標分布の期待値と予測分布の期待値の間の誤差を表す。

DNN の学習目標は，「平均評者」を近似することであったが，そもそも「平均評者」が与える評点が信用に足るものでない場合，すなわち正解ラベル自体が正確ではない場合，DNN の性能は制限されてしまう。そして，「平均評者」が与える評点の正確さは，人間の専門家の評者間変動に依存する。本章の実験で用いたテストデータセットでは，評者間信頼性と評者内信頼性の間の二次重み付け Cohen のカッパ係数の差は，項目 R で最も大きく 0.180 であり，二番目に大きかった項目 A で 0.063 であった（表 7.1）。そのため，項目 R で専門家の評者間信頼性を超える DNN の条件がなかった理由は，必ずしも DNN の問題ではなく，評者内信頼性と評者間信頼性の間の乖離があったデータセットの問題であった可能性があると考えられる。Fujimura ら [49] も同様に，データセットの質が DNN の性能に影響を与えることを報告している。

他方，人間の専門家と「離散平均評者」の間の評者間信頼性を上回った DNN の条件の数は，項目 A で最少，項目 S で二番目に少なかった。また，専門家の評者内信頼性を上回った DNN の条件数に関しても，項目 A で二番目，項目 S で三番目に少なかった（一番少なかったのは項目 R）。いくつかの研究によって，項目 A と S は他の項目と比べて信頼性が低くなることが報告されている [36, 39]。実際，GRBAS 尺度に影響を受けて考案された RBH 尺度（the RBH scale） [40] では項目 A と S が，CAPE-V [41] では項目 A が定量評価項目から除外されている。そのため，項目 A と S は，他の項目と比べて本質的に自動推定が難しかったのではないかと考えられる。

7.5.3　評点の推定が難しかった音声サンプル

図 7.6 で示したように，項目 B の推定は全体的にうまくいった（図 7.6 の結果に関しては，κ_{reg} は 0.927，κ_{class} は 0.815 であった）。しかし，図中に (A)–(F) でアノテーションを付与した 6 つの音声サンプルについては，回帰の推定誤差の絶対値が 0.5 以上となっ

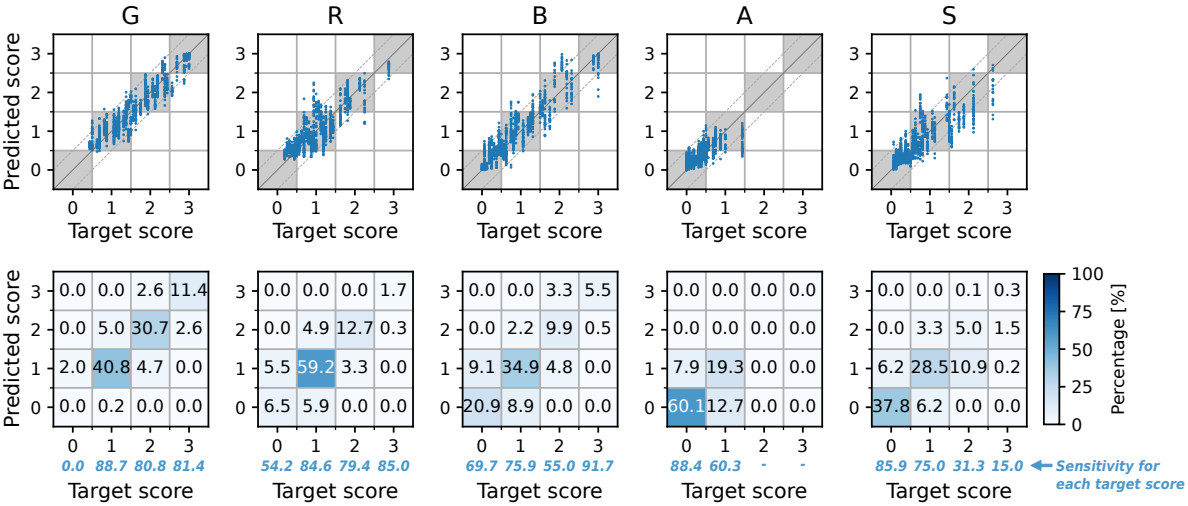


図 7.14 κ_{reg} が最高値を示したモデルを用いた時に得られた，GRBAS 尺度の評点の散布図と混同行列。各サブ図では，4 つの異なる乱数シードを用いて 5 分割交差検証を行った 20 ($= 4 \times 5$) 個の結果を同時に表示している。一つ一つの結果は 50 サンプルのテストデータセットに対するものであるため，各散布図には 1000 ($= 50 \times 20$) 個の点が含まれる。各混同行列の各数値は割合 [%] を表す。各混同行列の下部には，各評点（例えば，G0，G1，G2，G3）に対する感度（sensitivity）を記している。

た。図 7.13 に，アノテーションを付与した音声の波形と確率分布を示す。波形から観察できるように，これらの音声サンプルには，音量の揺らぎや声質の急激な変化，声の断絶などが含まれていた。本章の実験で用いた音声サンプルを選定する際は，A 特性重み付けレベルに関する SN 比が 30 dB 以上であることは課したが，その他には極端な問題がない限り選定対象とした。持続母音の定常性を仮定した先行研究では，図 7.13 に示すように複雑な音声サンプルの一部は，そもそも検討の対象外とされていた可能性がある。なお，音声サンプル (C), (D) について，DNN による予測分布は，ガウス分布に似た分布と，B3（項目 B の評点が 3 点）のところに存在するパルスが重なったような形状となった。音声サンプル (C) は正弦波成分がかなり弱い音声であった。気息性が強い音声は確かに正弦波成分が弱いことが多いため，DNN は B3 である可能性が高いと推定したと考えられるが，実際は，音声サンプル (C) は粗糙性（項目 R）の強さが際立つ音声であった。音声サンプル (D) は，有声区間と無声区間の両方が存在する複雑な音声であったため，DNN は無声区間に反応して B3 である可能性が高いと推定したと考えられる。

図 7.14 に，κ_{reg} が最大値を示した DNN の条件を用いた時に得られた，GRBAS 尺度の各項目についての散布図と混同行列を示す。全体的な傾向として，(0.5, 0.5) のような分類境界付近で誤分類が多く見られた。対照的に，S2 と S3 については，誤分類された多

くの音声サンプルは，分類境界から遠く離れた位置にあり，すなわち回帰誤差が大きかった。加えて，GRBAS 尺度の全項目，全評点の中で，S2 と S3 についての感度は，それぞれ 3 番目，2 番目に低かった（感度が最低だったのは G0 であるが，G0 の誤分類は分類境界付近で生じていた）。正解が S2 や S3 である音声サンプル数はそもそも少なかったこと（図 7.1）に加えて，S2 と S3 の音声サンプルの多くは，図 7.13 に示した音声サンプルのように非定常性がかなり強かった。以上より，データセットが小規模である場合，複雑な音声サンプルや正解ラベルが少ない評点の音声サンプルに対して，正確な推定を行えるよう学習することは難しいことが示唆された。

7.5.4 臨床応用アプリケーションとしての可能性

自動推定の主たる目標の一つは，人間による評価に内在する評者間変動や評者内変動を取り除くことであるため，自動推定の性能と専門家の信頼性を比較することが重要である。しかし，ほとんどの研究では，自動推定の性能と専門家の信頼性の比較が行われていない。Fujimura らの研究 [49] では，専門家の信頼性の指標（Krippendorff のアルファ係数）と自動推定の性能評価の指標（二次重み付け Cohen のカッパ係数）が異なっているため，両者を直接的に比較することができない。また，Kojima らの研究 [62] では，共通の指標（PCC）を用いて専門家の信頼性と自動推定の性能を比較しているものの，訓練データセットの正解ラベルを付与した評者と，テストデータセットの正解ラベルを付与した評者が異なっていた。訓練データセットとテストデータセットで評者が異なる場合，両者の間の評者間変動が存在するが，この評者間変動については示されていなかったため，結果の解釈は複雑であった。本章の実験では，同一の指標（二次重み付け Cohen のカッパ係数）を用いて専門家の信頼性と自動推定の性能を評価し，訓練・検証データセットの評者とテストデータセットの評者を同じにした。そのため，本章の実験では，専門家の信頼性と自動推定の性能を定量的に比較することができた。実験の結果，GRBAS 尺度の全項目で，人間の専門家と「離散平均評者」の間の評者間信頼性を DNN が上回った。加えて，項目 G，B，A，S で，DNN は専門家の評者内信頼性を上回った。これらの結果は，自動推定の性能は臨床応用に導入するのに十分な精度であったことを示唆している。

臨床応用では，専門家は，典型的には患者の音声の全体を聴取して声質評価を行う。しかし，先行研究で提案されている固定長波形を入力とする DNN では，切り出された固定長の音声セグメントごとに推定評点を与える [49,62,63]。複数の固定長音声セグメントに対する評点の平均値や中央値を使うことによって，音声全体に対する単一の評点を与えることは可能であるが，必ずしも平均値や中央値のような統計量が妥当かどうかは明らかではない。対照的に，本章で提案した DNN は，オンセットからオフセットまでを含む可変長の音声波形全体を入力とし，音声全体に対して一つの推定評点を直接与える。加えて，

オンセットからオフセットまでの切り出しを自動化すれば，声質評価を完全に自動化することができる。持続母音の定常部のみを抽出する操作と比較して，オンセットからオフセットまでの抽出はより明快でミスなく自動化しやすい操作である。

本章の実験で使用した 2 次元 DNN は，軽量かつ高速である EfficientNetV2-B0 に基づくものであった。GRBAS 尺度の 1 つの項目に対する DNN のパラメータ数は約 940 万，メモリサイズは 37.775 MB であった。テストモード（誤差逆伝播のための計算の途中経過保存や，データ拡張処理等が行われない状態）において，GRBAS 尺度の 1 つの項目に対する DNN は，5 秒間の音声サンプルから評点を推定するのに，ノートパソコン（Apple: MacBook Pro 13-inchi M1 2020）の CPU による計算で 112 ms ± 4.47 ms（平均 ± 標準偏差）要した。GRBAS 尺度の全項目を推定する場合，パラメータ数や実行時間は 5 倍となるが，これは実用に十分耐えうる規模である。

本実験では，GRBAS 尺度に焦点を当てたが，これは，データセットの正解ラベルを付与した専門家は，日常的に CAPE-V ではなく GRBAS 尺度を用いていたためである。その一方で，本研究で提案した DNN は，出力する確率分布を設計し直すことによって，直ちに CAPE-V の推定に適用することも可能である。そもそもの評価尺度が連続である CAPE-V の評点の推定では，回帰と分類の区別は不要となり，単に回帰問題と考えれば良くなる。

7.5.5　慣習的な音響特徴量に基づく手法との比較

声質の自動推定に関する多くの先行研究では，CPPS [44] 等の慣習的な音響特徴量に基づく手法が用いられている。本節では，慣習的な音響特徴量に基づく機械学習手法の性能と，本章で提案した DNN の性能を比較した結果を説明する。慣習的な音響特徴量に基づく機械学習手法としては，線形回帰と LightGBM [248] を採用した。線形回帰は声質の自動推定のために最も頻用される手法である [53, 54]。LightGBM は，勾配ブースティング決定木（gradient-boosting decision tree; GBDT）のフレームワークである。GBDT は，誤差が小さくなるように反復的に複数の決定木を構築し，各決定木の出力の和を最終的な出力とする手法であり，音声障害 [249] やパーキンソン病 [250] の自動検出に応用されている。LightGBM では分類モデルと回帰モデルの両方を作成できるが，ここでは回帰モデルを用いた。線形回帰モデルと LightGBM による回帰モデルは，慣習的な音響特徴量を入力とし，「連続平均評者」によって与えられた評点を予測するように設計した。また，線形回帰モデルと LightGBM のハイパーパラメータは，最適化ライブラリ Optuna [251] に実装されている tree-structured Parzen estimator（TPE）アルゴリズム [252] によって最適化した。TPE アルゴリズムは，ハイパーパラメータと損失の間の確率モデルを考えることによって，効率的にハイパーパラメータを最適化する手法である [252]。なお，

表 7.5 機械学習手法間の κ_{reg} による性能比較。κ_{reg} の値は「平均 ± 標準偏差」で表している。GRBAS 尺度の各項目につき，最高値を**太字**で示している。この結果は，各機械学習手法，GRBAS 尺度の各項目につき，κ_{reg} が最高値を示したモデルを用いた時に得られたものである。

手法	G	R	B	A	S
DNN	$\mathbf{0.934 \pm 0.011}$	0.831 ± 0.028	$\mathbf{0.925 \pm 0.014}$	$\mathbf{0.716 \pm 0.037}$	$\mathbf{0.829 \pm 0.037}$
線形回帰	0.914 ± 0.004	$\mathbf{0.837 \pm 0.016}$	0.888 ± 0.002	0.619 ± 0.019	0.781 ± 0.007
LightGBM	0.920 ± 0.008	0.773 ± 0.042	0.901 ± 0.011	0.608 ± 0.024	0.752 ± 0.050

音響特徴量と機械学習手法の設定の詳細については，付録 B で説明した。

表 7.5 に，最適な条件を用いた時の DNN と，最適化された線形回帰と LightGBM のモデルの κ_{reg} の結果を示す。DNN は，項目 R では第 2 位（第 1 位より 0.006 低い）であったが，項目 G（第 1 位より 0.014 高い），B（第 1 位より 0.024 高い），A（第 1 位より 0.097 高い），S（第 1 位より 0.048 高い）では線形回帰と LightGBM の両方のモデルより高い推定精度を示した。この結果は，たとえデータセットのサンプル数が少ない（本章の実験の訓練データセットのサンプル数は 200）場合でも，DNN は，慣習的な音響特徴量に基づく機械学習手法を上回る推定精度を実現できることを示唆している。

7.5.6 異なる母音やデータベースに対する頑健性

本章の実験では，ほとんどの先行研究で使われている持続母音/a/を用いた [49, 53–58, 60–63, 253]。その一方で，CAPE-V では母音/a/以外にも母音/i/が評価対象である他，いくつかの研究 [61, 253] では母音/a/，/i/，/u/の 3 種類が用いられている。また，本研究では，単一施設で収録された音声データから構成されるデータセットを用いたが，実際の臨床応用では，施設や録音環境が異なる状況下でも頑健に高い推定精度を実現する必要がある。医療機関によって，録音機器や防音室の有無，音声障害の分布等は異なるのが実情である。いくつかの研究によって，訓練データセットとテストデータセットの間の差異が，自動推定精度に影響を与えることが示唆されている [58, 62]。

そこで，本節では，九州大学病院で収録された 200 サンプルの持続母音/a/の音声データを使って学習した機械学習モデルが，他の母音やデータベースに対してどの程度頑健であるかを調査した結果を報告する。頑健性の調査のために，具体的には，九州大学病院で収録された 50 サンプルの持続母音/e/の音声データと，Perceptual Voice Qualities Database（PVQD） [254] に含まれる 134 サンプルの持続母音/a/と 134 サンプルの/i/の音声データを用いた。九州大学病院で収録された 50 サンプルの持続母音/e/は，テス

トデータセットの持続母音/a/を発声した話者の音声である。PVQD は，5 つの施設で収録された健常音声と病的音声のデータから構成されるデータベースであり，各音声には GRBAS 尺度の評点が付与されており，音声の中にはノイズが含まれるものも存在する。PVQD には，296 名の話者による発声が含まれるが，今回は，持続母音/a/と/i/の両方が収録されていた 134 名の話者による音声データを用いた。また，オリジナルの PVQD の音声データは，アノテーションが付与されていない生の録音データであった。そこで，今回は，音声データに Praat を用いてアノテーションを付与し，持続母音/a/と/i/の発話区間のみを抽出して使用した。

表 7.6 に，DNN と線形回帰，LightGBM のモデルを使って，これらの音声データに対して GRBAS 尺度の評点の自動推定を行って得られた κ_{reg} の結果を示す。この結果は，九州大学病院で収録された持続母音/a/に対して自動推定を行った時に，最も κ_{reg} が高かったモデルを用いた時に得られたものである。なお，各機械学習モデルは，九州大学病院で収録された持続母音/a/を使って学習した後に，再学習を行わずに，他の母音やデータベースの音声に対して GRBAS 尺度の評点を予測した。結果の詳細について説明する前に，以下の理由によって，κ_{reg} の値の大きさが，必ずしも頑健性の高さに一致するとは限らないことを注意しておく。

- 九州大学病院で収録された持続母音/e/に対する GRBAS 尺度の評点として，代替的に九州大学病院で収録された持続母音/a/に対する GRBAS 尺度の評点を流用した。このように代替的な処理を行ったのは，持続母音/e/は専門家による評価が行われていなかったためである。しかし，持続母音/e/に対する GRBAS 尺度の実際の評点は，持続母音/a/に対する GRBAS 尺度の評点と異なる可能性がある。発声は毎回同じように行われているとは限らない他，発話内容によって GRBAS 尺度の評点は異なる場合があることも報告されている [255–257]。
- PVQD では，アノテーションが付与されていない，様々な母音や短文の発声を含む生の録音データに対して，GRBAS 尺度の評点が付与されている。しかし，前項目で述べたように，各発声に対する GRBAS 尺度の評点と，様々な発声を含む生の録音データに対する GRBAS 尺度の評点は異なる可能性がある。
- 九州大学病院で収録された音声データに対する評者と，PVQD の音声に対する評者は別人であった。加えて，PVQD では，音声データによって評者が異なっている（PVQD 全体では 19 名の評者がいるが，各音声データに対する評者の数は 3 名または 4 名）。実際，評者の臨床経験や所属集団 [33–35]，職業 [36, 37]，熟練度 [33, 36] の違いは，GRBAS 尺度の評点の与え方の傾向に違いをもたらす原因になることが示唆されている。このようにデータベース間で評者が異なるため，得られた実験結果には，機械学習モデルの要因と評者間変動の要因が混在しており，こ

表 7.6 異なるデータベース・母音に対する機械学習手法間の κ_{reg} による性能比較。κ_{reg} の値は「平均 ± 標準偏差」で表している。GRBAS 尺度の各項目，各データベース，各母音につき，最高値を**太字**で示している。この結果は，各機械学習手法，GRBAS 尺度の各項目につき，九州大学病院で収録された母音/a/に対する κ_{reg} が最高値を示したモデルを用いたときに得られたものである。

九州大学病院で収録された持続母音/e/

手法	G	R	B	A	S
DNN	0.869 ± 0.009	0.667 ± 0.021	$\mathbf{0.872 \pm 0.013}$	0.524 ± 0.067	$\mathbf{0.783 \pm 0.039}$
線形回帰	$\mathbf{0.900 \pm 0.005}$	$\mathbf{0.729 \pm 0.035}$	0.864 ± 0.007	$\mathbf{0.551 \pm 0.023}$	0.710 ± 0.016
LightGBM	0.894 ± 0.008	0.621 ± 0.034	0.848 ± 0.017	0.327 ± 0.052	0.710 ± 0.033

PVQD に含まれる持続母音/a/

手法	G	R	B	A	S
DNN	0.530 ± 0.026	$\mathbf{0.591 \pm 0.029}$	0.622 ± 0.049	$\mathbf{0.274 \pm 0.039}$	0.368 ± 0.059
線形回帰	$\mathbf{0.618 \pm 0.030}$	0.572 ± 0.025	$\mathbf{0.695 \pm 0.008}$	0.186 ± 0.028	0.444 ± 0.025
LightGBM	0.603 ± 0.024	0.566 ± 0.018	0.643 ± 0.014	0.163 ± 0.028	$\mathbf{0.474 \pm 0.039}$

PVQD に含まれる持続母音/i/

手法	G	R	B	A	S
DNN	$\mathbf{0.481 \pm 0.050}$	$\mathbf{0.358 \pm 0.040}$	$\mathbf{0.493 \pm 0.042}$	$\mathbf{0.119 \pm 0.034}$	0.225 ± 0.070
線形回帰	0.478 ± 0.045	0.331 ± 0.049	0.443 ± 0.028	0.019 ± 0.032	0.116 ± 0.021
LightGBM	0.387 ± 0.030	0.192 ± 0.105	0.403 ± 0.060	0.086 ± 0.023	$\mathbf{0.352 \pm 0.040}$

れらの要因を切り分けることは実質不可能である。

上述の理由のため，κ_{reg} の値の大きさは，頑健性の高さに完全には一致しないことに注意する必要がある。しかし，それでも，ある程度は頑健性の高さに対応すると考えられるため，本節では，κ_{reg} を考察のために用いた。

DNN を使った場合の κ_{reg} の値は，九州大学病院の持続母音/e/に対しては GRBAS 尺度の 2 つの項目で，PVQD の持続母音/a/に対しては GRBAS 尺度の 2 つの項目で，PVQD の持続母音/i/に対しては GRBAS 尺度の 4 つの項目で，線形回帰や LightGBM のモデルを使った場合の κ_{reg} の値を上回った。典型的には，他の機械学習モデルと比較して DNN は過剰適合を引き起こしやすいと考えられているが，これらの結果は，学習した DNN が十分な汎化性能を有していることを示唆している。なお，すべての機械学習手法で，学習に用いなかった母音やデータベースの音声データに対しては κ_{reg} の値が

低下した。各データベースに関して，持続母音/a/に対する κ_{reg} の値は，他の母音に対する κ_{reg} の値よりも高かった。様々な種類の母音に対する頑健性を高めるためには，訓練データセットに様々な種類の母音を含めるようにすることが必要であると考えられる。PVQD の音声データに対する κ_{reg} の値が九州大学病院の音声データに対する κ_{reg} の値よりも低くなった原因としては，データベース間の発話者が罹患していた音声障害の分布の違いや，録音条件の違い，評者間変動等があると考えられる。録音条件の違いに対する頑健性を向上させるためには，元の音声波形の振幅に，雑音波形の振幅を加算する雑音重畳が効果的である可能性がある。雑音重畳は，話者認証等の課題で有効性が確かめられている [258, 259]。

7.6　まとめ

本章では，300 サンプルの小規模な病的音声のデータセット（200 サンプルの訓練データセット，50 サンプルの検証データセット，50 サンプルのテストデータセット）を元に，対数振幅および位相微分の時間周波数表現を入力特徴量とした，DNN による GRBAS 尺度の評点の自動推定実験を行った。実験の結果，DNN による推定精度は，人間の専門家の信頼性に匹敵することが示された。加えて，サンプル数が少ない場合であっても，DNN による推定精度は，慣習的な音響特徴量に基づく手法の推定精度を上回ることが可能であることが示された。また，提案した DNN は，ノートパソコンによる計算でも高速な推定を行える程度に計算効率が高かった。データ拡張は推定精度の向上に大きく寄与し，中でも最も効果的なデータ拡張手法はランダム速度摂動であった。データ拡張を適用した場合，項目 G，R，A，S ではパワー（対数振幅メル周波数スペクトログラム）が最も有効な時間周波数表現であった。その一方で，項目 B では群遅延が最も有効であり，群遅延と他の時間周波数表現を組み合わせることによって，推定精度がさらに改善することが観察された。この結果より，病的音声の声質の分析においては，振幅情報と位相情報が相補的な役割を果たすことが示唆された。以上の成果は，提案した DNN の臨床応用アプリケーションとしての実現可能性の高さを示すとともに，高い推定精度を実現する上で，データセット構築のための聴覚心理的評価の労力は軽減可能であることを示すものであった。

第 8 章

総括

8.1 本論文のまとめ

本研究では，深層学習を用いた病的音声の声質に対する聴覚心理的評価の評点の自動推定に関する検討を行った。具体的には，時間周波数表現の位相情報の有効性に関する検討と，病的音声データセットが小規模な場合における，推定精度向上のための手段に関する検討を行った。時間周波数表現の位相情報としては，瞬時周波数や群遅延のような位相微分を中心に検討を行い，データセットのサンプル数の少なさを補うための手段として，データ拡張手法についての調査を行った。

第 4 章では，深層学習を用いた声質自動評価実験のために必要となる，専門家によって GRBAS 尺度の評点が付与された病的音声データベースを構築した。ここでは，まず，収集したすべての音声データに対して，1 名の耳鼻咽喉科医による GRBAS 尺度の評点付与を行った。その後，それらの音声データの中から，小規模ながらも多様な声質の音声データを含む 300 サンプルの音声データを選別した。この 300 サンプルの音声データに対しては，経験年数が 2 年以上の専門家（耳鼻咽喉科医または言語聴覚士）8 名，経験年数が 2 年未満の専門家 4 名，非専門家 5 名を実験参加者とする声質評価の聴覚実験を実施した。聴覚実験の結果，経験年数が 2 年以上の専門家群は，GRBAS 尺度のすべての項目で，他の群より評者内信頼性が高い傾向が見られ，他の群より声質評価のための内的基準が確立されていることが推察された。最終的に，機械学習用データセットとして使うための音声データとして，耳鼻咽喉科医 1 名による GRBAS 尺度の評点が付与された 2,545 サンプルを第 5 章の評点自動推定実験で，経験年数が 2 年以上の専門家 8 名による評点が付与された 300 サンプルを第 7 章の評点自動推定実験で採用した。

第 5 章では，聴覚心理的評価の評点の自動推定における位相情報の有効性，特に瞬時周波数に関連する情報の有効性について検討した。ここでは，有効性について検討するために，評点が付与された 2,545 サンプルの病的音声データを元に，位相の時間変化および振

幅に基づく時間周波数表現と RNN を用いて，GRBAS 尺度の評点の自動推定実験を行った。実験の結果，項目 G，R，B，S に対しては，振幅情報よりも位相情報を用いたときに高い推定精度が得られた。この結果は，深層学習を用いた病的音声の声質評価において，位相情報が有効であることを示唆した。また，周波数軸のメル尺度への変換は，振幅情報だけでなく，位相情報に関しても有効であることが示された。

第 6 章では，GRBAS 尺度の評点の自動推定という課題から一度離れ，8 種類の音分類課題に対して，2 次元 CNN を用いて，位相のフェーザ表示，瞬時周波数，群遅延の有用性を調査した。実験の結果，5 種類の音分類課題に対して，振幅情報のみを用いた場合と比較して，位相のフェーザ表示や瞬時周波数，群遅延を用いることで性能向上がもたらされた。また，位相のフェーザ表示が効果的であった課題に対しては，瞬時周波数や群遅延も効果的であった。この結果は，出力目標が音声波形ではない回帰や分類のような課題では，位相の値それ自体よりも，微分で表現されるような位相の隣接関係の方が重要であることを示唆した。その一方で，メル尺度のような周波数軸を持つ位相情報の時間周波数表現を求めるために，拡張 LEAF と呼ぶ手法を提案したが，拡張 LEAF には，計算コストの大きさや，群遅延の計算の正確さに関して課題があった。

第 7 章では，第 6 章の実験で得られた知見を活かしつつ，聴覚心理的評価の評点の自動推定という課題に立ち返って実験を行った。具体的には，8 名の専門家によって評価された 300 サンプルの病的音声データを元に，対数振幅，瞬時周波数，および群遅延から得られる時間周波数表現と，2 次元 CNN と RNN，MLP を組み合わせた DNN を用いて，GRBAS 尺度の評点の自動推定実験を実施した。ここでは，瞬時周波数と群遅延の有効性についてのさらなる検討に加えて，病的音声データセットが小規模な場合における，推定精度向上のための手段についての検討を行った。サンプル数の少なさを補うデータ拡張手法としては，ランダム速度摂動とランダムクロップ，周波数マスキングの有効性をそれぞれ調査した。また，ここでは，周波数軸がメル尺度である位相情報の時間周波数表現を求めるために，計算効率と計算の正確性の両方に関して拡張 LEAF よりも優れた手法を提案した。実験の結果，DNN による推定精度は，GRBAS 尺度のすべての項目で専門家の評者間信頼性に匹敵し，項目 G，B，A，S で専門家の評者内信頼性に匹敵し，項目 G，B，A，S で慣習的な音響特徴量に基づく手法の推定精度を上回った。また，提案した DNN は計算効率が高く，ノートパソコンで高速な推定を行うことができた。最も効果的なデータ拡張手法はランダム速度摂動であり，推定精度改善に大きく貢献した。データ拡張を適用した場合，項目 G，R，A，S では対数振幅が最も有効な時間周波数表現であった。その一方で，項目 B では群遅延が最も有効であり，群遅延と他の時間周波数表現を組み合わせることによって，推定精度がさらに改善することが観察され，振幅情報と位相情報は特徴量として相補的な役割を担うことが示唆された。以上の成果は，提案した DNN の臨床応用アプリケーションとしての実現可能性の高さを示すとともに，高い推定精度を実現す

る上で，データセット構築のための聴覚心理的評価の労力は軽減可能であることを示すものであった。

8.2 今後の展望と課題

8.2.1 CAPE-V や連続音声への対応

本研究では，聴覚心理的評価法として GRBAS 尺度を取り上げた。GRBAS 尺度は現在も広く用いられている一方で，GRBAS 尺度が日本音声言語医学会によって提案されたのは 1979 年と歴史は古い [260]。項目 A を評価項目に含めることに対する疑義が示される [27] 等，GRBAS 尺度には改善の余地があると考えられており，2002 年には，ASHA によって，GRBAS 尺度に影響を受けた聴覚心理的評価法である CAPE-V が提案された [41]。CAPE-V はアメリカ以外の国へも広まり [261]，本論文執筆時点で，日本においても普及に向けた取り組みが行われている最中である [42]。CAPE-V では，GRBAS 尺度の項目 A に対応する声質が定量評価対象から除外されている他，離散尺度（0，1，2，3）ではなく連続尺度（0〜100）を使って各項目を評価する。また，発話内容として，GRBAS 尺度では基本的に持続母音のみを対象としているのに対し，CAPE-V では持続母音に加え，連続発話である短文音読と会話も対象に含まれている。連続発話は日常的な発声である一方で持続母音は非日常的な発声であること [262] や，連続発話と持続母音では知覚される嗄声の程度が異なる場合があること [263] 等が指摘されており，連続発話に対する声質評価を行う意義はあると考えられる。実際，深層学習ではなく慣習的な音響特徴量に基づくアプローチではあるが，連続発話を用いた聴覚心理的評価の評点の自動推定について検討した研究も存在する [44, 53, 54, 264]。本研究では，最も標準的な発話内容である持続母音/a/に焦点を当て，深層学習に基づく GRBAS 尺度の評点の自動推定に関する研究を行ったが，今後は，持続母音だけでなく連続発話も検討対象に含めた，連続尺度で表される評点の自動推定が重要になる可能性がある。

8.2.2 聴覚心理的評価の評点が付与された病的音声の公開データベース

本研究を含め，聴覚心理的評価の評点の自動推定に関するほぼすべての研究は，音声データと評点の両方，あるいはどちらか一方が非公開であるため，結果の比較を行うことが困難である。Moro-Velázquez らによる研究 [45] は，筆者が知る限り音声データと評点の両方が公開されている唯一の研究である。しかし，この研究で使用されている，Massachusetts Eye and Ear Infirmary Voice Laboratory で収録された病的音声データベース [265] は，かつて販売されていたが，現在では購入できなくなっている [254]。その一方で，論文執筆時点では，無償で利用可能な病的音声データベースとして，296

サンプルの音声データから構成される PVQD [254] や，サンプル数が 2,000 を超える Saarbruecken Voice Database [266] が公開されている。今後，聴覚心理的評価の評点の自動推定に関する研究結果の公正な比較を可能にするためには，病的音声の公開データベースに対して，聴覚心理的評価の評点を公開情報として付与することが必要になると考えられる。

なお，PVQD には，既に聴覚心理的評価（GRBAS 尺度と CAPE-V）の評点が付与されているものの，音声データによって評者が異なっている上に，各音声データに対する評者の数は 3 名または 4 名と多くはないため，自動推定実験における正解値としてそのまま使っていいかどうかは疑問が残る。自動推定実験に使える水準の評点付きデータベースを構築するためには，各音声データに対する評者の数を多くしたり，各評者の評者内信頼性を検証したりする等，気を付けなければならないことがあると考えられる。また，SVD はサンプル数が多く，すべての音声データに対して聴覚心理的評価の評点を付与することは多大な人的資源を要するため，多様な声質の音声データから構成される小規模なサブセットを抽出することが重要になるかもしれない。本研究では，多様な声質の音声データのサブセットを抽出するために，1 名の耳鼻咽喉科医によって実際に付与された評点を用いたが，このやり方自体も大きな労力を伴うものであった。多様な声質の音声データのサブセットを効率良く選び出すためには，能動学習（active learning） [267, 268] と呼ばれる，機械学習モデルの性能改善に大きく寄与するサンプルを見つけ出す手法が有効である可能性がある。

謝辞

本研究の実験の大部分は，筆者が九州大学大学院芸術工学府芸術工学専攻博士後期課程在学中に，音声情報研究室において行ったものです。病的音声の声質評価というテーマ自体は，筆者が学士 4 年時の九州大学芸術工学部音響設計学科在学中に着手したものであり，病的音声データベースの作成開始から含めると，2017 年より足掛け 6 年間に渡って取り組んだものです。

本研究を遂行するにあたり，多くの御指導，御助言を賜りました九州大学大学院芸術工学研究院の鏑木時彦教授と若宮幸平助教に心より御礼申し上げます。ご多忙の中にも関わらず，非常に長い時間を割いて私の研究活動全般を支えていただき，事前の連絡もなく急にお伺いしたような時でさえも，親身になって相談に乗っていただきました。先生方には，未熟な私をいつも見守っていただいているような感覚がありました。

九州大学大学院芸術工学研究院の吉永幸靖准教授と高田正幸准教授には，本論文の副査を快く引き受けていただき，専門的な観点からの考察や細部にわたるご指摘等，非常に有益なご助言をいただきました。深く感謝申し上げます。

病的音声データベース構築にあたり，九州大学大学院医学研究院の中川尚志教授，聴覚実験にご協力いただいた実験参加者の皆様，音声データの発話者をしていただいた皆様，九州大学病院耳鼻咽喉科の音声データ収集の関係者の皆様に深く感謝申し上げます。病的音声データベースは，本研究を支える基盤になりました。

本研究は，李庸學博士の存在なしには始まることすらありませんでした。李庸學博士には，研究活動全般において支えていただいたことに加え，その類稀なるバイタリティで様々な場面に連れ出していただき，生き様を含めて多くのことを学ばせていただきました。心より御礼申し上げます。

私が学部 4 年の時から 6 年間所属した音声情報研究室において，時間を共有していただいた学生の皆様には，研究活動以外も含めた私の日常生活の一部を形成していただき，大変お世話になりました。特に，中西萌さんと武田葵さんと共同で病的音声に関する研究に取り組めたことは，研究の発展に繋がっただけでなく，心強いものであり，お二方が居なくては心が折れそうになっていた場面もあったと思います。心より御礼申し上げます。

数学勉強会の皆様には，なけなしの数学知識しか持っていなかった私を温かく迎えていただき，大変お世話になりました。数学勉強会に参加するきっかけは岡田昌大博士にいただき，岩見貴弘さんには，時間周波数表現に関する勉強を一緒にしていただきました。数学勉強会で身に付けた数学的素養は，本研究にも活かされたと考えています。深く感謝申し上げます。

最後に，本研究を遂行するにあたってお世話になったすべての方々，生きる動機づけを与えてくれたすべてのものに，この場を借りて厚く御礼申し上げます。

付録 A

実験で使用した 2 次元 CNN の構造

ここでは，第 6 章と第 7 章の実験で使用した 2 次元 CNN を順に説明する。

A.1　第 6 章で使用した 2 次元 CNN の構造

図 A.1 (A) は，第 6 章の実験で使用した 2 次元 CNN を示す。この CNN は，EfficientNet-B0 [232] と呼ばれる CNN である。但し，LEAF [168] の研究で使われた実装 [225] に基づいて本来の EfficientNet-B0 の一部が修正されており，大域平均プーリングの代わりに大域最大プーリングが採用されている。

EfficientNet-B0 は，精度と計算効率の両方が高くなるように工夫が施された EfficientNet [232] と呼ばれる 2 次元 CNN のファミリーの中で，最も小規模な CNN である。EfficientNet-B0 の構成要素は，ConvBlock（図 A.2 (A)），MBConv1（図 A.2 (B)），MBConv6（図 A.2 (C)）である。ConvBlock は畳み込み層と Batch Normalization，それに活性化関数の直列接続である。「MBConv」で名前が始まる構成要素は，EfficientNet に影響を与えた MnasNet の研究 [269] で導入された構造である。

MBConv1 は，通常の畳み込み層を深さ別畳み込み層と点別畳み込み層のセットに置換することによって，演算量の削減が図られている [204] 他，MBConv1 には Squeeze-and-Excitation（SE） [270, 271] という機構が組み込まれている [272]。SE は，図 A.3 に示す構造をしており，大域平均プーリングによって各チャンネルの大域情報を集約した後（squeeze），0 から 1 までの値を取るチャンネル単位でのゲートを計算し，チャンネル毎にゲートを乗算する（excitation）。SE は，後続の層に伝える情報の重み付けを選択的に行う自己注意（self-attention） [273] としての役割を果たし，比較的軽量な計算オーバーヘッドで CNN の性能向上に寄与する [270, 271]。なお，SE には削減率（reduction rate）$r_{\text{reduction}}$ と呼ばれるハイパーパラメータが存在し，ゲートを計算する際，一時的にチャンネル数を $1/r_{\text{reduction}}$ 倍に削減する。削減率を大きくするほど計算量を削減できる一方，

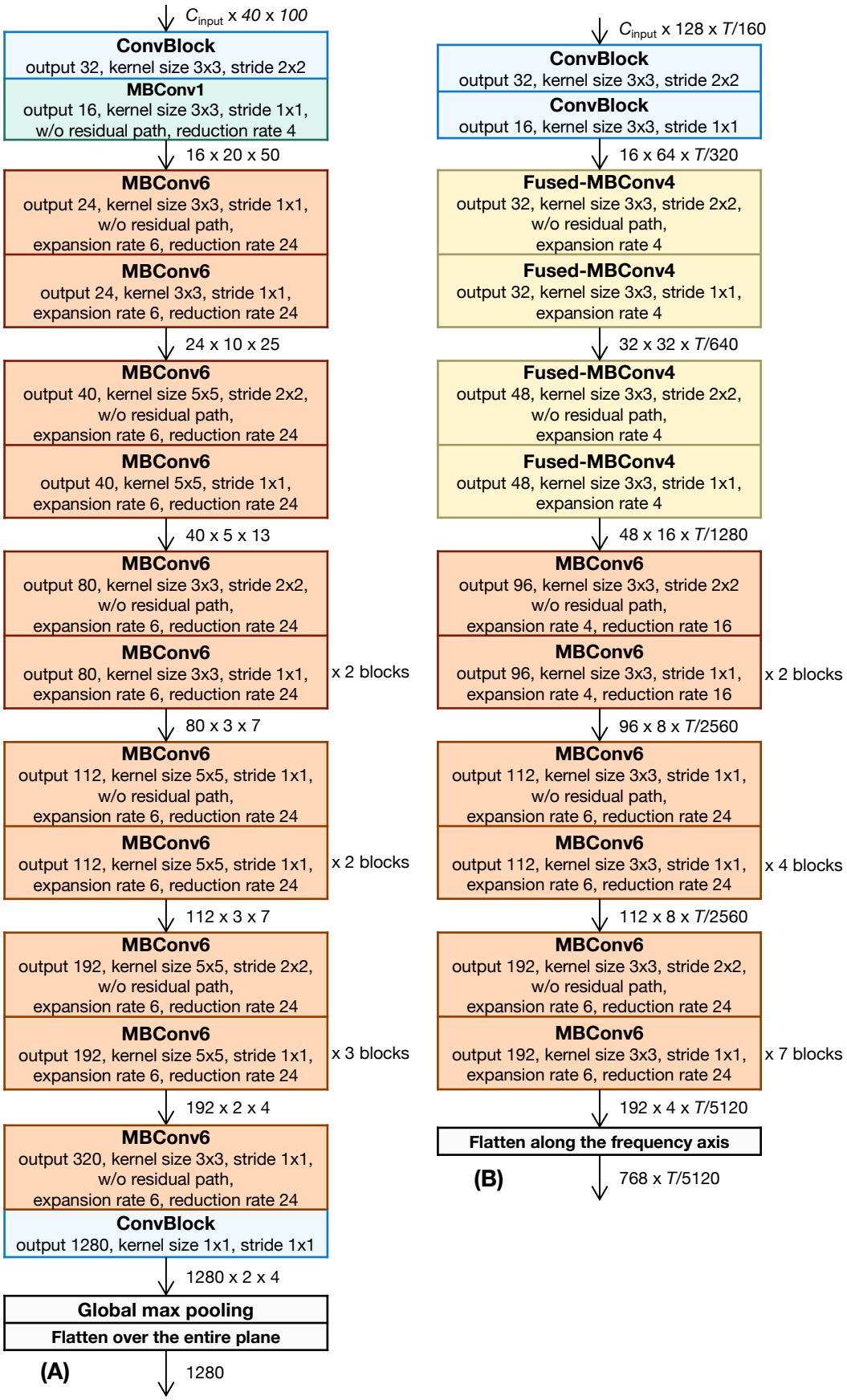


図 A.1 (A) 第 6 章で使用した 2 次元 CNN，(B) 第 7 章で使用した 2 次元 CNN

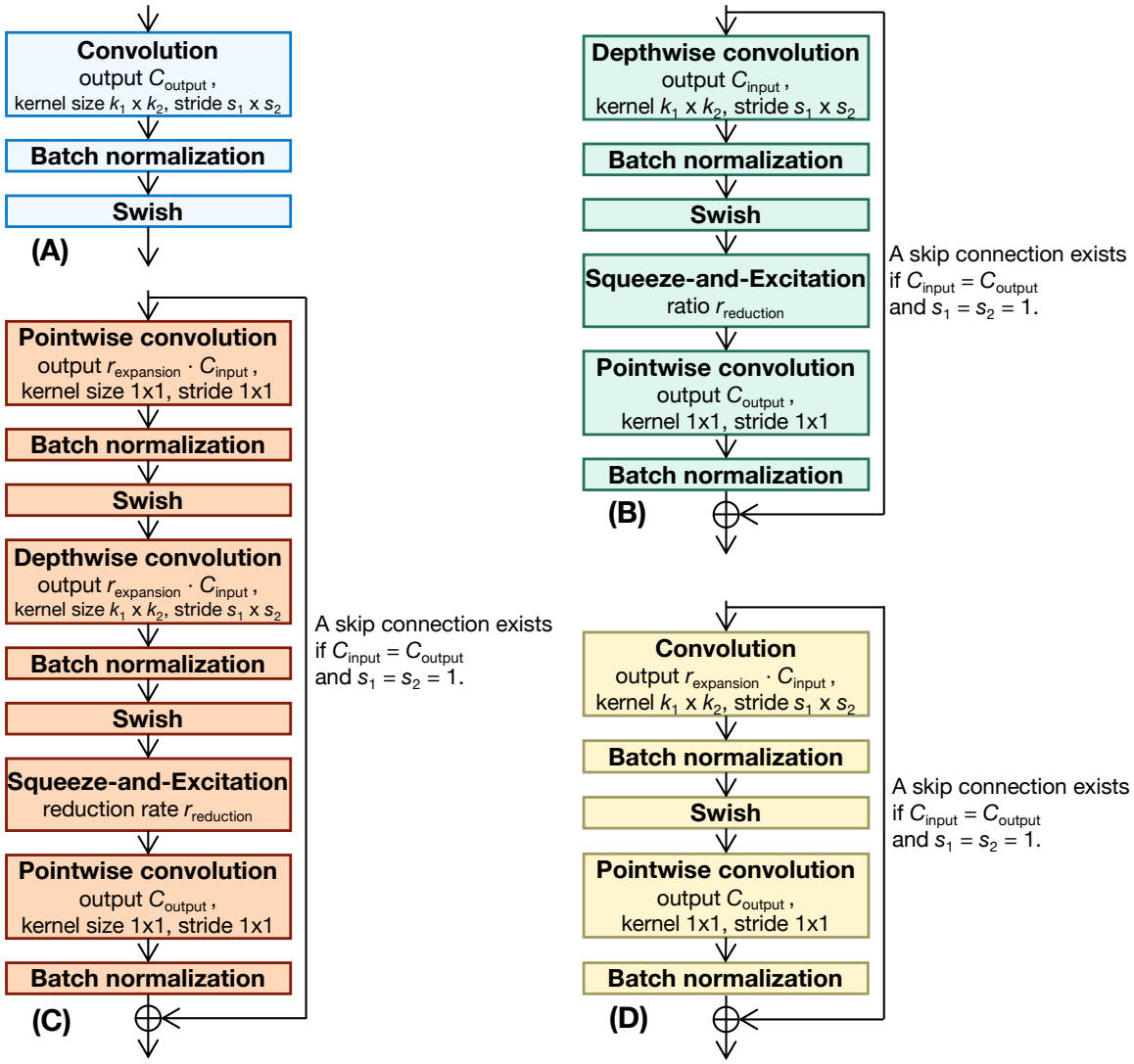


図 A.2 図 A.1 中の CNN の構成要素。(A) ConvBlock，(B) MBConv1，(C) MBConv6，(D) Fused-MBConv4。

削減率を過剰に大きくすると精度が低下するため，削減率は計算量と精度のトレードオフで設定される。

MBConv6 は，逆残差（inverted residual）ブロック [274] と呼ばれる構造に SE を組み込むことによって構成される [272]。図 A.4 は，逆残差ブロックの構造を示す。逆残差ブロックは，点別畳み込み層でチャンネル数をハイパーパラメータである拡大率（expansion rate）$r_{\text{expansion}}$ の割合で増加させ，深さ別畳み込み層を作用させた後に，点別畳み込み層でチャンネル数を元に戻す[*1]。また，最初の点別畳み込み層と深さ別畳み込み層の後では活性化関数を適用するが，最後の点別畳み込み層の後では活性化関数を適用しない。逆残

[*1] チャンネル数が元に戻るのは，図 A.4 において $C_{\text{input}} = C_{\text{output}}$ の場合のみ。$C_{\text{input}} \neq C_{\text{output}}$ の場合はチャンネル数は元に戻らないが，通常，$r_{\text{expansion}} \cdot C_{\text{input}} > C_{\text{output}}$ となる。

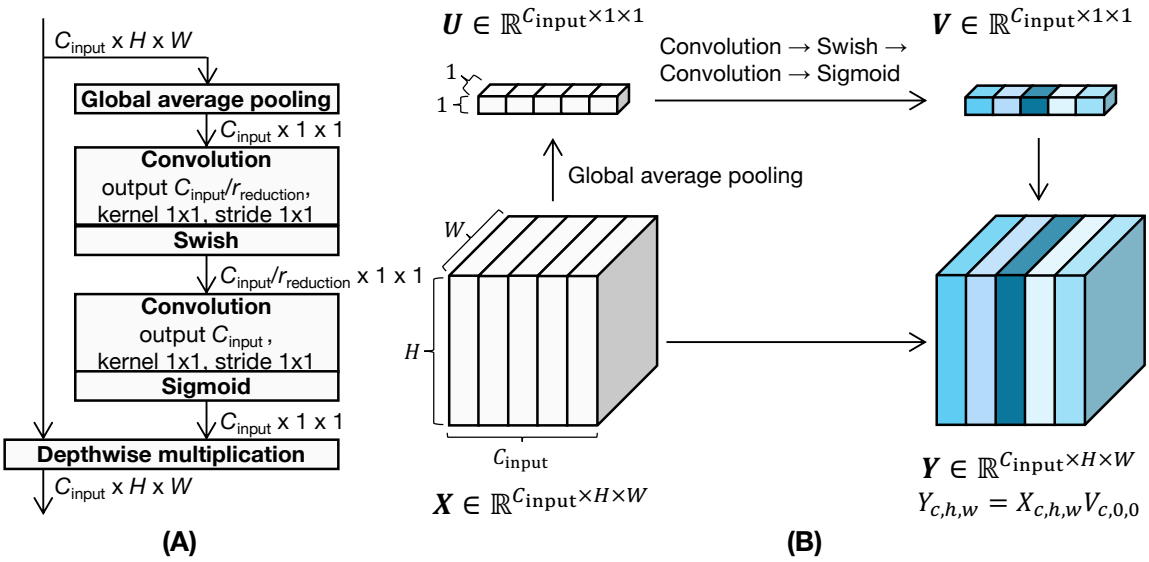


図 A.3 SE の (A) ブロックダイアグラム，(B) テンソルのフロー図

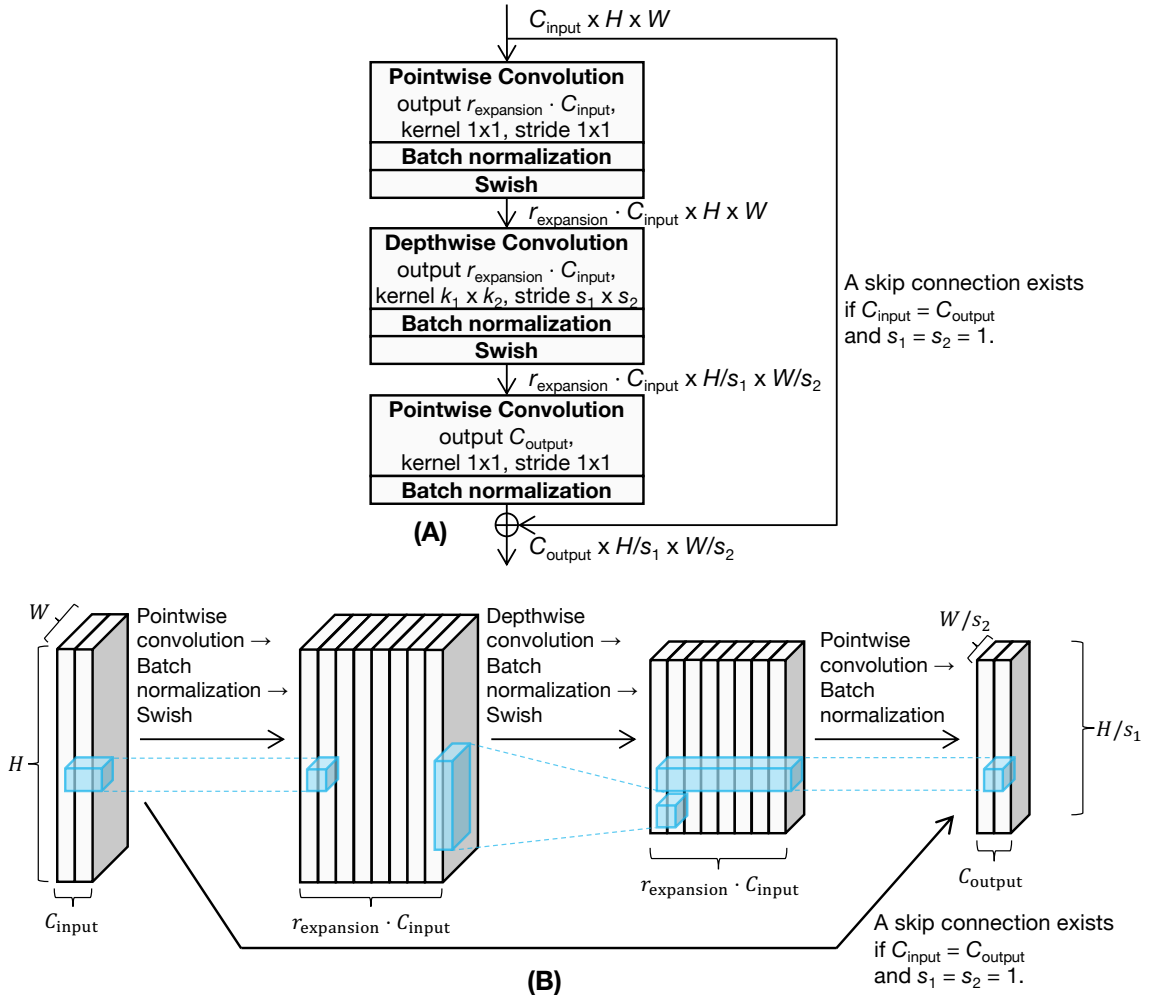


図 A.4 逆残差ブロックの (A) ブロックダイアグラム，(B) テンソルのフロー図

差ブロックは，「現実世界のデータは高次元空間で表現されるが，その分布は低次元多様体に埋め込まれている」という多様体仮説（manifold hypothesis）[64] に着想を得ている。具体的には，「データ自体は低次元多様体に埋め込まれているため，入出力のチャンネル数 C_{input} と C_{output} はある程度小さくしても問題ない一方で，ReLU のような活性化関数を適用すると低次元多様体に関する情報が失われてしまう恐れがあるため，活性化関数を適用するのはチャンネル数が拡大されている段階に限定した方が良い」という洞察に基づき，逆残差ブロックは設計された [274]。実際，逆残差ブロックのチャンネル数を減少させる最後の点別畳み込み層の後では，活性化関数を適用しない方が性能が向上することが実験結果から示唆されている [274]。逆残差ブロックは，性能を維持しながら入出力のチャンネル数を抑制することを可能にし，計算負荷の低減に貢献する。

A.2　第 7 章で使用した 2 次元 CNN の構造

図 A.1 (B) は，第 7 章の実験で使用した 2 次元 CNN を示す。この CNN は，EfficientNetV2-B0 [239, 275] と呼ばれる CNN である。但し，この 2 次元 CNN の直後で LSTM による処理を行うことを考慮して元々の EfficientNetV2-B0 の一部が修正されており，大域平均プーリングが除去されている他，過度に複雑性を増加させないために，チャンネル数を 1,280 に増加させる最後の畳み込み層が省略されている。また，Batch Normalization の定義が可変幅入力を扱えるものに修正されている（第 7.2.2 節参照）。

EfficientNetV2-B0 [239, 275] は，EfficientNet を計算効率と性能の両面で改良した EfficientNetV2 [239] と呼ばれる 2 次元 CNN のファミリーの中で，最も小規模な CNN である[*2]。EfficientNet と比較して，EfficientNetV2 では，畳み込み層のカーネルサイズがすべて 3×3 になる代わりにブロック数（深さ）が増加しており，メモリアクセスのオーバーヘッドを減らすためにより小さな拡大率が設定されているブロックが多い。また，EfficientNetV2 では，入力に近い箇所で Fused-MBConv4（図 A.2 (D)）と呼ばれる構造が使われている。Fused-MBConv4 [239] は，深さ別畳み込み層ではなく，通常の畳み込み層が使われていることが特徴である。Fused-MBConv4 が導入された背景には，深さ別畳み込み層はアクセラレータを十分に活用できないために，理論値ほど計算効率が高くはないという論文執筆時点でのハードウェア事情がある。すべての MBConv6 を Fused-MBConv4 に置換すると，計算量の増加と学習速度の低下を招いてしまうが，入力に近い箇所のみ MBConv6 を Fused-MBConv4 に置換すると計算効率が向上し学習速度が向上することが実験的に観察されており [239]，EfficientNetV2 では入力に近い箇所のみ Fused-MBConv4 が使用されている。

[*2] EfficientNetV2-B0 は，EfficientNetV2 の提案論文 [239] では言及されていないが，公式実装 [275] に含まれている。

付録 B

第 7 章で使用した慣習的な音響特徴量に基づく機械学習手法

この付録では，第 7 章で使用した慣習的な音響特徴量に基づく機械学習手法について説明する。音響特徴量としては，GRBAS 尺度の項目 G [53, 58]，項目 R [58]，項目 B [54, 58]，項目 S [276] の評点の自動推定に関する研究で，有効性が確認された以下のものを採用した。

CPPS [44, 53, 54]

平滑化ケプストラル・ピーク卓越度（smoothed cepstral peak prominence; CPPS）。平滑化したパワーケプストラムの第 1 ピークのケフレンシ（quefurency）における，平滑化したパワーケプストラムの第 1 ピークと平滑化したパワーケプストラムの回帰直線の間のガムニチュード（gamnitude）の差。単位は dB で表される。なお，パワーケプストラムは対数パワースペクトルの逆フーリエ変換の二乗として定義され，スペクトルの周波数（frequency）と振幅（magnitude）に対応するものを，パワーケプストラムではケフレンシとガムニチュードと呼ぶ。平滑化は，時間方向とケフレンシ方向のそれぞれにローパスフィルタを適用することで行う。

GNEmax-4500 Hz [54, 277]

上限周波数を 4,500 Hz として計算された声門対雑音励起比（glottal-to-noise excitation ratio; GNE）。GNE は，声帯振動がどのくらいパルス的であるか（病的振動ではないか）を推定するために考案された音響特徴量であり，以下の手順で計算される。

1. 音声波形を 10 kHz へダウンサンプリングする。
2. 線形予測に基づく逆フィルタリングによって，音声波形から声道特性を除去する。

3. 中心周波数が異なるバンドパスフィルタバンクを通した後に，ヒルベルト変換によって振幅包絡を計算する。
4. 帯域幅の半分以上，中心周波数が離れた振幅包絡の全組み合わせに対して，±0.3 ms の範囲で相関関数を計算する。
5. 相関関数の最大値を求め，これを GNE とする。

GNE の値の範囲は $[0, 1]$ であり，値が大きい場合は声帯振動が正常であり，値が小さい場合は声帯振動が病的であることが期待される。

Hfno-6000 Hz [54, 278]

0–6 kHz のエネルギーと 6–10 kHz のエネルギーの比として定義される高周波雑音の相対的な強さ。単位は dB ではなく，無次元で表される。

HNR [53, 58]

倍音対雑音比（harmonics-to-noise ratio）。単位は dB で表される。

HNR-D [54, 279]

Dejonckere と Lebacq [279] による倍音対雑音比。ケプストラムで基本周波数を推定した後，500–1500 Hz の範囲で倍音対雑音比を計算する。単位は dB で表される。

H1-H2 [54]

第 1 倍音と第 2 倍音の振幅の差。単位は dB で表される。

Jit [54]

周期揺らぎを表す局所ジッタ（local jitter）。分析区間内での周期列が $\{T_n\}_{n=1}^{N}$ であるとき，Jit は $\left(\sum_{j=2}^{N} |T_j - T_{j-1}|/(N-1)\right)/\left(\sum_{k=1}^{N} T_k/N\right)$ と定義される [280]。

Loudness-1st [276, 281]

部分ラウドネス（specific loudness）の 1 次モーメント。ラウドネス（loudness）は音の大きさの心理量であり，1/3 オクターブ分析のデータから，臨界帯域（critical band）ごとの部分ラウドネスを求め，部分ラウドネスの総和を求めることによって推定される [282]。部分ラウドネスは臨界帯域ごとに推定されるが，全帯域の部分ラウドネスを一度に表示する際は，横軸が臨界帯域比 [Bark]，縦軸が部分ラウドネス [sone/Bark] である振幅スペクトルのように表現する。ここで，1 Bark は 1 つの臨界帯域幅に対応し [108]，1 sone は 40 dB の 1 kHz の純音から知覚される音の大きさとして定義される [283]。部分ラウドネスが $\{s_n\}_{n=1}^{N}$，それに対応する臨界帯域比が $\{b_n\}_{n=1}^{N}$ であるとき，Loudness-1st は $\left(\sum_{j=1}^{N} b_j s_j\right)/\left(\sum_{k=1}^{N} s_k\right)$ と定義される。

Loudness-2nd [276, 281]

部分ラウドネスの 2 次の中心モーメントの平方根。部分ラウドネスが $\{s_n\}_{n=1}^{N}$，それに対応する臨界帯域比が $\{b_n\}_{n=1}^{N}$, Loudness-1st が μ であるとき, Loudness-2nd

は $\sqrt{\left(\sum_{k=1}^{N}|b_k-\mu|^2 s_k\right)/\left(\sum_{l=1}^{N} s_l\right)}$ と定義される。Python の音響信号処理ライブラリである librosa [145] の関数 `librosa.feature.spectral_bandwidth` の実装を参考に, 偏差そのものではなく, 偏差の絶対値を用いた (以下の Loudness-3rd, Loudness-4th に関しても同様)。

Loudness-3rd [276, 281]

部分ラウドネスの 3 次の中心モーメントの三乗根。部分ラウドネスが $\{s_n\}_{n=1}^{N}$, それに対応する臨界帯域比が $\{b_n\}_{n=1}^{N}$, Loudness-1st が μ であるとき, Loudness-3rd は $\sqrt[3]{\left(\sum_{k=1}^{N}|b_k-\mu|^3 s_k\right)/\left(\sum_{l=1}^{N} s_l\right)}$ と定義される。

Loudness-4th [276, 281]

部分ラウドネスの 4 次の中心モーメントの四乗根。部分ラウドネスが $\{s_n\}_{n=1}^{N}$, それに対応する臨界帯域比が $\{b_n\}_{n=1}^{N}$, Loudness-1st が μ であるとき, Loudness-4th は $\sqrt[4]{\left(\sum_{k=1}^{N}|b_k-\mu|^4 s_k\right)/\left(\sum_{l=1}^{N} s_l\right)}$ と定義される。

Loudness-skew [276, 281]

部分ラウドネスの歪度 (skewness)。部分ラウドネスが $\{s_n\}_{n=1}^{N}$, それに対応する臨界帯域比が $\{b_n\}_{n=1}^{N}$, Loudness-1st が μ, Loudness-2nd が σ であるとき, Loudness-skew は $\left(\left(\sum_{k=1}^{N}(b_k-\mu)^3 s_k\right)/\left(\sum_{l=1}^{N} s_l\right)\right)/\sigma^3$ と定義される。

Loudness-kurt [276, 281]

部分ラウドネスの尖度 (kurtosis)。部分ラウドネスが $\{s_n\}_{n=1}^{N}$, それに対応する臨界帯域比が $\{b_n\}_{n=1}^{N}$, Loudness-1st が μ, Loudness-2nd が σ であるとき, Loudness-kurt は $\left(\left(\sum_{k=1}^{N}(b_k-\mu)^4 s_k\right)/\left(\sum_{l=1}^{N} s_l\right)\right)/\sigma^4$ と定義される。

PSD [54]

周期の標準偏差 (period standard deviation; PSD)。分析区間内での周期列が $\{T_n\}_{n=1}^{N}$ であるとき, PSD は $\{T_n\}_{n=1}^{N}$ の標準偏差として定義される。

RALA [45, 58]

2 次元画像として表現される振幅変調スペクトル (amplitude modulation spectrum) についての, 線形平均の大きさを超えない要素数に対する, 線形平均の大きさを超える要素数の比 (the ratio of points above linear average; RALA)。振幅変調スペクトルは, 振幅スペクトログラムを時間方向にフーリエ変換した表現の絶対値として計算される。振幅スペクトログラムの横軸が時間を表すのに対し, 振幅変調スペクトルの横軸は変調周波数 (modulation frequency) を表す。振幅変調スペクトルに関して, 健常音声では低域の変調周波数帯にエネルギーが集中するのに対し, 病的音声では広範囲の変調周波数帯にエネルギーが分散する様子が観察されている [45]。RALA は, エネルギーが広範囲の変調周波数帯に分散しているほど値が大きくなるため, 値が大きいほど音声が病的であることが期待される。

Sharpness [276, 284]

音の鋭さの心理量であるシャープネス（sharpness）。シャープネスの単位は acum で表され，1 acum は，音量が 60 dB で，1 kHz を中心周波数とし，帯域幅が臨界帯域幅である雑音から知覚されるシャープネスとして定義される [284]。部分ラウドネスが $s : \mathbb{R} \to \mathbb{R}$ のような臨界帯域比を引数とする関数として表現されるとき，シャープネスを $c\Big(\int_0^{24} s(z)g(z)z\mathrm{d}z\Big)/\Big(\int_0^{24} s(z)\mathrm{d}z\Big)$ と推定することが Fastel と Zwicker [284] によって提案されている。ここで，z は臨界帯域比 [Bark]，$c > 0$ は比例定数，$g : \mathbb{R} \to \mathbb{R}$ は高域成分を強調する重み付け関数である。数値計算時は，上式を離散化して用いる。

Shim [53, 54]

振幅揺らぎを表す局所シマ（local shimmer）。分析区間内での振幅列が $\{A_n\}_{n=1}^N$ であるとき，Shim は $\big(\sum_{j=2}^N |A_j - A_{j-1}|/(N-1)\big)/\big(\sum_{k=1}^N A_k/N\big)$ と定義される [280]。

Shim-dB [53, 54]

単位を dB で表示した振幅を用いて計算される局所シマ。分析区間内での振幅列が $\{A_n\}_{n=1}^N$ であるとき，Shim-dB は $\sum_{j=2}^N |20\log_{10}(A_j/A_{j-1})|/(N-1)$ と定義される [280]。単位は dB で表される。

Slope [53]

対数振幅スペクトルの傾き（slope）。対数振幅スペクトルの 0–1000 Hz のエネルギーと 1000-10000 Hz のエネルギーの差と定義される。単位は dB で表される。

Tilt [53]

対数振幅スペクトルの回帰直線の傾き（tilt）。対数振幅スペクトルの回帰直線の 0–1000 Hz のエネルギーと 1000-10000 Hz のエネルギーの差と定義される。単位は dB で表される。

採用した音響特徴量の内，ラウドネス統計量とシャープネス，RALA の計算には，自作の Python スクリプトを用いた。部分ラウドネスとシャープネスの計算には，音質評価指標の計算ライブラリである MOSQITO [285] を用いた。部分ラウドネスの計算法としては，International Organization for Standardization（ISO）によって標準化された時間変動信号に対する Zwicker の手法 [281] を，シャープネスの重み付けとしては，Deutsches Institut für Normung（DIN） [286] によって標準化された定義を採用した。他の音響特徴量に関しては，Maryn らの論文 [53] の Appendix 1 と Appendix 2，および Barsties らの論文 [54] の Supplementary Data に載っていたスクリプトを元に Praat で計算を行った。なお，すべての音響特徴量の値としては時間軸上の平均値を用い，標本化周波数

は，RALA 以外の音響特徴量計算時は 48 kHz，RALA の計算時は 20 kHz とした[*1]。

機械学習手法としては，線形回帰と LightGBM [248] を採用した。線形回帰モデル内で使用する音響特徴量（説明変数）は，κ_{reg} を最大化するように，すべての組み合わせパターンを探索した（力まかせ探索; brute-force search; exhaustive search）。線形回帰の性能は，5 分割交差検証によって評価した。但し，5 分割交差検証の最中，線形回帰の係数は訓練データセットに対して一意に決定されるため，検証データセットは使用されなかった。LightGBM には，様々なハイパーパラメータが存在する。今回の実験では，最適化ライブラリである Optuna [251] に実装されている tree-structured Parzen estimator（TPE）アルゴリズム [252] を用いて，κ_{reg} を最大化するように，ハイパーパラメータの最適化を行った。LightGBM のハイパーパラメータの最適化のための反復回数は 2,000 回とした。ハイパーパラメータの探索範囲に関して，`lambda_l1` $\in \mathbb{R}_{>0}$ については $[1.0 \times 10^{-8}, 10.0]$，`lambda_l2` $\in \mathbb{R}_{>0}$ については $[1.0 \times 10^{-8}, 10.0]$，`num_leaves` $\in \mathbb{N}$ については $[2, 256]$，`feature_fraction` $\in \mathbb{R}_{>0}$ については $[0.1, 1.0]$，`bagging_fraction` $\in \mathbb{R}_{>0}$ については $[0.1, 1.0]$，`bagging_freq` $\in \mathbb{Z}_{\geq 0}$ については $[0, 7]$，`min_data_in_leaf` $\in \mathbb{Z}_{\geq 0}$ については $[0, 100]$ とした。決定木の最大の深さ（`max_depth`）は 8 とした。LightGBM の性能評価は，DNN の性能評価と同様に，乱数シードを変えて 4 回，5 分割交差検証を行った。

表 B.1 から表 B.5 に，GRBAS 尺度の各項目についての，κ_{reg} を最大化するように最適化したときに選択された，線形回帰モデル内の音響特徴量とその回帰係数を示す。また，表 B.6 に LightGBM の最適化されたハイパーパラメータを，図 B.1 に最適化されたハイパーパラメータを用いた時の LightGBM の特徴量重要度を示す。LightGBM の特徴量重要度が高かった音響特徴量が，必ずしも線形回帰モデルの説明変数で採用されておらず，機械学習手法によって音響特徴量の有効性が異なる可能性が示唆された。なお，TPE を用いて LightGBM のハイパーパラメータを最適化した際の，テストデータセットに対する κ_{reg} の値の遷移は図 B.2 のようになり，最適化によって LightGBM の性能がほぼ上限に収束している様子が観察された。

[*1] 音響特徴量の計算で実際に用いたスクリプトは，GitHub のリポジトリ（https://github.com/onkyo14taro/auto-grbas）および，第 7 章の元になっている原論文 [287] の Supplementary Materials （https://doi.org/10.1016/j.jvoice.2022.10.020）で公開している。

表 B.1 項目 G についての，κ_{reg} を最大化するように最適化したときに選択された，線形回帰モデル内の音響特徴量とその回帰係数。各列は，5 分割交差検証の 1 つの分割に対して最適化した時の係数を表す。但し，最後の列は，5 つの分割に対する結果の平均と標準偏差を表す。

音響特徴量	分割 1	分割 2	分割 3	分割 4	分割 5	平均 ± 標準偏差
切片	−2.351	−3.993	−2.299	−1.311	−3.315	−2.654 ± 1.031
CPPS	−0.029	−0.022	−0.014	−0.031	−0.03	−0.025 ± 0.007
GNEmax-4500 Hz	−0.481	−0.519	−0.605	−0.456	−0.305	−0.474 ± 0.110
HNR	−0.074	−0.074	−0.084	−0.072	−0.073	−0.076 ± 0.005
PSD	64.074	74.845	59.223	88.483	66.364	70.598 ± 11.487
Slope	−0.003	−0.012	−0.009	−0.007	0.004	−0.005 ± 0.006
Tilt	−0.053	−0.07	−0.057	−0.045	−0.068	−0.059 ± 0.010
Loudness-1st	0.402	0.482	0.362	0.376	0.456	0.416 ± 0.052
Loudness-2nd	3.882	4.178	2.994	3.542	4.111	3.741 ± 0.486
Loudness-3rd	−8.122	−8.753	−6.875	−6.959	−8.873	−7.916 ± 0.957
Loudness-4th	4.371	4.776	4.023	3.47	4.901	4.308 ± 0.583
Loudness-skew	1.394	1.531	0.894	1.375	1.508	1.340 ± 0.259

表 B.2 項目 R についての，κ_{reg} を最大化するように最適化したときに選択された，線形回帰モデル内の音響特徴量とその回帰係数。各列は，5 分割交差検証の 1 つの分割に対して最適化した時の係数を表す。但し，最後の列は，5 つの分割に対する結果の平均と標準偏差を表す。

音響特徴量	分割 1	分割 2	分割 3	分割 4	分割 5	平均 ± 標準偏差
切片	1.68	−0.663	7.336	6.671	2.24	3.453 ± 3.428
CPPS	0.053	0.057	0.056	0.049	0.039	0.051 ± 0.007
HNR	−0.082	−0.074	−0.092	−0.072	−0.08	−0.080 ± 0.008
H1-H2	−0.021	−0.017	−0.024	−0.024	−0.021	−0.021 ± 0.003
PSD	34.63	17.589	20.737	47.279	48.769	33.801 ± 14.490
Shim	−0.009	−0.087	−0.046	−0.104	−0.094	−0.068 ± 0.040
ShimdB	0.227	1.304	0.631	1.331	1.05	0.909 ± 0.473
Slope	0.004	0.005	−0.005	−0.006	0.011	0.002 ± 0.007
Tilt	−0.071	−0.058	−0.055	−0.058	−0.08	−0.064 ± 0.011
Loudness-1st	−0.6	−0.386	−0.984	−0.939	−0.585	−0.699 ± 0.255
Loudness-2nd	2.098	2.228	0.378	0.284	0.527	1.103 ± 0.972
Loudness-4th	−1.859	−1.811	−0.594	−0.649	−0.475	−1.078 ± 0.694
Loudness-skew	0.994	1.357	−0.333	0.357	0.541	0.583 ± 0.644
Loudness-kurt	0.672	0.682	0.029	−0.142	0.19	0.286 ± 0.376
Sharpness	3.058	2.171	3.954	4.691	3.117	3.398 ± 0.959

表 B.3 項目 B についての，κ_{reg} を最大化するように最適化したときに選択された，線形回帰モデル内の音響特徴量とその回帰係数。各列は，5 分割交差検証の 1 つの分割に対して最適化した時の係数を表す。但し，最後の列は，5 つの分割に対する結果の平均と標準偏差を表す。

音響特徴量	分割 1	分割 2	分割 3	分割 4	分割 5	平均 ± 標準偏差
切片	4.39	4.004	2.917	3.868	3.904	3.817 ± 0.544
CPPS	−0.11	−0.115	−0.104	−0.117	−0.114	-0.112 ± 0.005
GNEmax-4500 Hz	−1.967	−1.753	−1.863	−1.759	−1.678	-1.804 ± 0.112
Jit	−0.052	−0.035	−0.048	−0.06	−0.014	-0.042 ± 0.018
PSD	56.626	54.498	32.543	76.912	25.113	49.138 ± 20.669
Slope	−0.021	−0.024	−0.027	−0.023	−0.019	-0.023 ± 0.003
Tilt	0.047	0.022	0.008	0.023	0.028	0.026 ± 0.014
Loudness-skew	−1.385	−1.182	−1.239	−1.386	−1.198	-1.278 ± 0.101
Loudness-kurt	0.295	0.194	0.28	0.278	0.264	0.262 ± 0.040
Sharpness	−0.216	−0.142	0.442	−0.056	−0.158	-0.026 ± 0.268

表 B.4 項目 A についての，κ_{reg} を最大化するように最適化したときに選択された，線形回帰モデル内の音響特徴量とその回帰係数。各列は，5 分割交差検証の 1 つの分割に対して最適化した時の係数を表す。但し，最後の列は，5 つの分割に対する結果の平均と標準偏差を表す。

音響特徴量	分割 1	分割 2	分割 3	分割 4	分割 5	平均 ± 標準偏差
切片	−2.234	−1.076	−1.333	0.531	0.983	-0.626 ± 1.343
CPPS	−0.051	−0.051	−0.059	−0.059	−0.045	-0.053 ± 0.006
GNEmax-4500 Hz	−0.017	0.1	0.044	0.01	0.041	0.036 ± 0.044
Hfno-6000 Hz	0.048	−0.169	−0.088	−0.063	−0.175	-0.089 ± 0.091
HNR	0.002	0.005	0.012	0.003	0.002	0.005 ± 0.004
HNR-D	−0.006	−0.009	−0.013	−0.008	−0.005	-0.008 ± 0.003
H1-H2	0.008	0.008	0.011	0.01	0.012	0.010 ± 0.002
Jit	−0.033	−0.052	−0.045	−0.034	−0.013	-0.035 ± 0.015
Shim	0.01	0.016	0.006	0.006	0.009	0.009 ± 0.004
Loudness-1st	0.473	0.476	0.48	0.309	0.301	0.408 ± 0.094
Loudness-2nd	3.066	3.548	3.105	3.016	3.288	3.205 ± 0.218
Loudness-3rd	−7.136	−7.806	−7.025	−6.424	−6.981	-7.075 ± 0.493
Loudness-4th	4.215	4.395	4.014	3.405	3.632	3.932 ± 0.409
Sharpness	−2.02	−2.428	−2.081	−1.592	−1.643	-1.953 ± 0.344

表 B.5 項目 S についての，κ_{reg} を最大化するように最適化したときに選択された，線形回帰モデル内の音響特徴量とその回帰係数。各列は，5 分割交差検証の 1 つの分割に対して最適化した時の係数を表す。但し，最後の列は，5 つの分割に対する結果の平均と標準偏差を表す。

音響特徴量	分割 1	分割 2	分割 3	分割 4	分割 5	平均 ± 標準偏差
切片	−5.527	−8.121	−7.093	−3.992	−7.391	-6.425 ± 1.657
CPPS	0.019	0.036	0.033	0.028	0.027	0.028 ± 0.006
HNR	−0.081	−0.07	−0.081	−0.084	−0.081	-0.079 ± 0.006
Jit	−0.026	0.012	0.017	0.017	0.001	0.004 ± 0.018
Shim	−0.016	−0.022	−0.023	−0.031	−0.025	-0.024 ± 0.005
Slope	0.004	0.003	0.002	0.006	0.01	0.005 ± 0.003
Tilt	−0.074	−0.079	−0.084	−0.057	−0.086	-0.076 ± 0.012
RALA	0.165	1.584	1.761	1.855	0.55	1.183 ± 0.772
Loudness-1st	0.695	0.779	0.724	0.624	0.786	0.722 ± 0.066
Loudness-2nd	4.577	4.468	3.233	2.789	3.883	3.790 ± 0.775
Loudness-3rd	−8.067	−8.496	−6.283	−4.1	−7.674	-6.924 ± 1.784
Loudness-4th	3.455	4.112	3.036	1.144	3.779	3.105 ± 1.166
Loudness-skew	3.019	2.992	2.865	3.109	3.047	3.006 ± 0.090

表 B.6 LightGBM の最適化されたハイパーパラメータ

ハイパーパラメータ	G	R	B	A	S
`lambda_l1`	3.04×10^{-5}	5.93×10^{-8}	1.54×10^{-4}	5.96×10^{-1}	6.56×10^{-5}
`lambda_l2`	1.62×10^{-6}	1.27×10^{-2}	4.16×10^{-7}	3.53×10^{-3}	2.29×10^{-8}
`num_leaves`	53	76	40	46	89
`feature_fraction`	0.39	0.687	0.324	0.78	0.861
`bagging_fraction`	0.891	0.844	0.158	0.705	0.287
`bagging_freq`	4	0	0	1	0
`min_data_in_leaf`	18	0	23	0	8

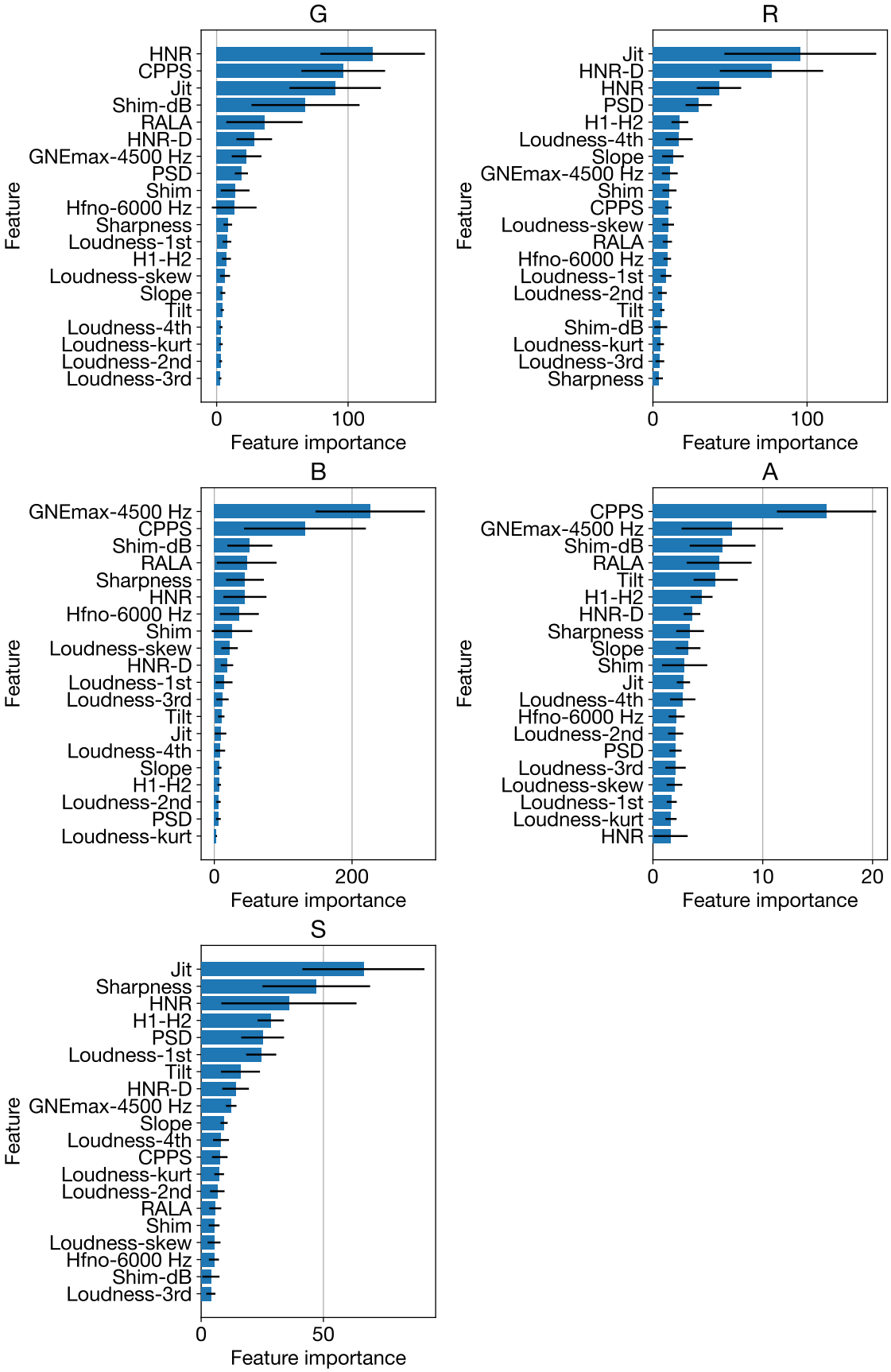


図 B.1 LightGBM の最適化されたモデルの特徴量重要度。エラーバーは標準偏差を表す。

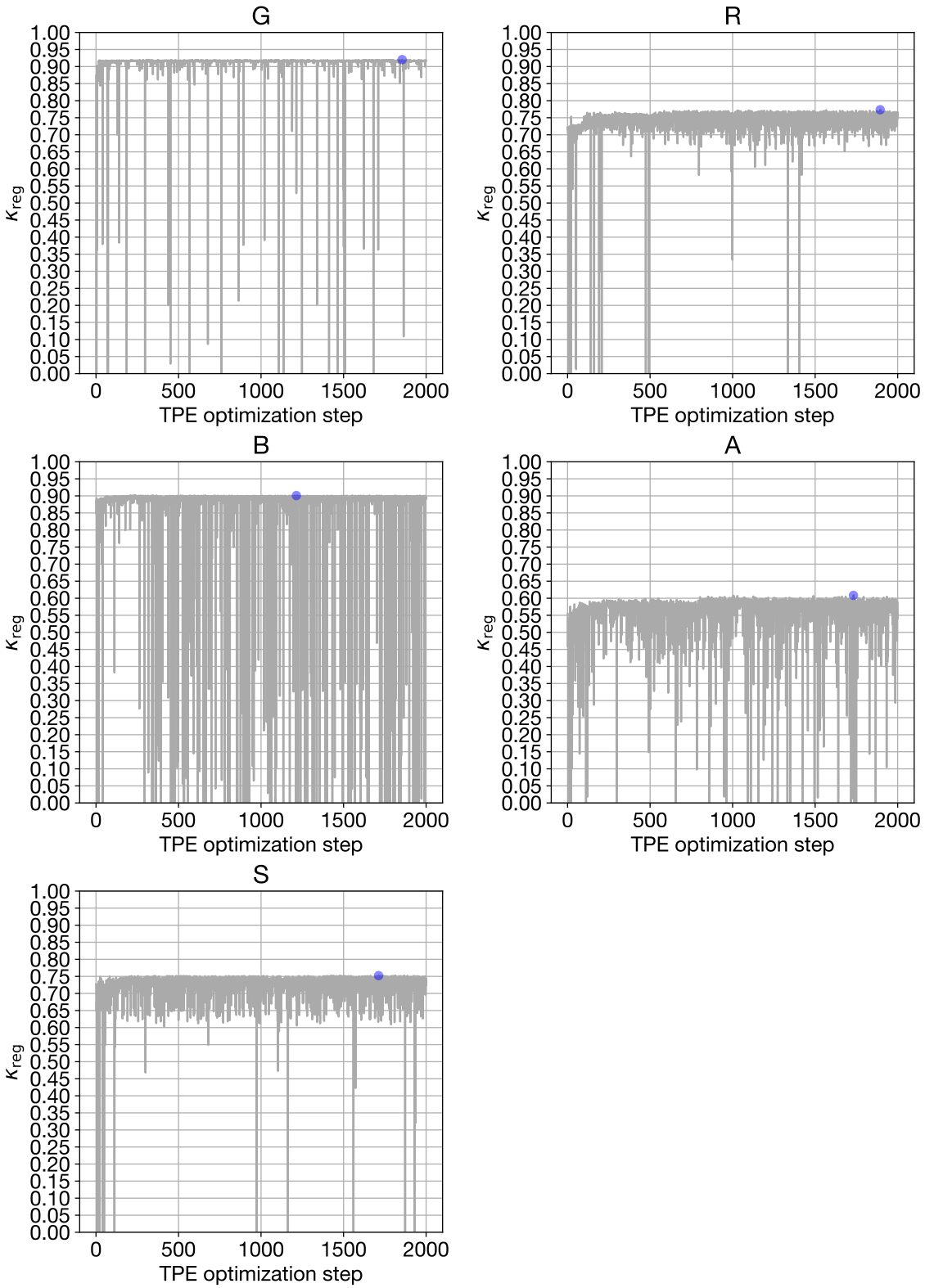


図 B.2 TPE を用いて LightGBM のハイパーパラメータを最適化した際の，テストデータセットに対する κ_{reg} の値（5 分割交差検証の平均値）の遷移。青い点は最大値を表す。

参考文献

[1] 森 大毅，前川 喜久雄，粕谷 英樹，“音声による情報伝達”，音声は何を伝えているか - 感情・パラ言語情報・個人性の音声科学 -，森 大毅，前川 喜久雄，粕谷 英樹, Eds. 東京: コロナ社，2014，pp. 1–21.

[2] D. Felce, and J. Perry, “Quality of life: its definition and measurement,” *Res. Dev. Disabil.*, vol. 16, no. 1, pp. 51–74, 1995. doi: 10.1016/0891-4222(94)00028-8.

[3] B. H. Jacobson, A. Johnson, C. Grywalski, A. Silbergleit, G. Jacobson, M. S. Benninger, and C. W. Newman, “The voice handicap index (VHI): development and validation,” *Am. J. Speech. Lang. Pathol.*, vol. 6, no. 3, pp. 66–70, 1997. doi: 10.1044/1058-0360.0603.66.

[4] N. D. Hogikyan, and G. Sethuraman, “Validation of an instrument to measure voice-related quality of life (V-RQOL),” *J. Voice*, vol. 13, no. 4, pp. 557–569, 1999. doi: 10.1016/s0892-1997(99)80010-1.

[5] C. D. Mulrow, C. Aguilar, J. E. Endicott, R. Velez, M. R. Tuley, W. S. Charlip, and J. A. Hill, “Association between hearing impairment and the quality of life of elderly individuals,” *J. Am. Geriatr. Soc.*, vol. 38, no. 1, pp. 45–50, 1990. doi: 10.1111/j.1532-5415.1990.tb01595.x.

[6] D. Monzani, G. M. Galeazzi, E. Genovese, A. Marrara, and A. Martini, “Psychological profile and social behaviour of working adults with mild or moderate hearing loss,” *Acta Otorhinolaryngol. Ital.*, vol. 28, no. 2, pp. 61–66, 2008.

[7] American Speech-Language-Hearing Association, “Definitions of communication disorders and variations,” Rockville, MD, Tech. Rep., 1993.

[8] 日本音声言語医学会，新編 声の検査法．東京: 医歯薬出版株式会社，2009.

[9] 日本音声言語医学会，日本喉頭科学会，音声障害診療ガイドライン 2018 年版．東京: 金原出版，2018.

[10] 日本音声言語医学会, 第 2 版 声の検査法 基礎編. 東京: 医師薬出版株式会社, 1994.

[11] ——, 第 2 版 声の検査法 臨床編. 東京: 医師薬出版株式会社, 1994.

[12] 西澤 典子, 苅安 誠, 三枝 英人, 椎名 英貴, 田中 康博, 中森 正博, 中谷 謙, 南都 智紀, 福永 真哉, 細見 直永, 益田 慎, 中川 尚志, "Dysarthria の翻訳名称について", 音声言語医学, vol. 64, no. 1, pp. 24–32, 2023.

[13] 西尾 正輝, "かつてわれわれは dysarthria を「構音障害」と呼んでいた, といえる新たなる時代を迎えよう: 小島氏への異論", 音声言語医学, vol. 41, no. 1, pp. 73–73, 2000. doi: 10.5112/jjlp.41.73.

[14] R. J. Stachler, D. O. Francis, S. R. Schwartz, C. C. Damask, G. P. Digoy, H. J. Krouse, S. J. McCoy, D. R. Ouellette, R. R. Patel, C. C. W. Reavis, L. J. Smith, M. Smith, S. W. Strode, P. Woo, and L. C. Nnacheta, "Clinical practice guideline: Hoarseness (dysphonia) (update)," *Otolaryngol. Head Neck Surg.*, vol. 158, no. 1_suppl, pp. S1–S42, 2018. doi: 10.1177/0194599817751030.

[15] K. Verdolini, C. A. Rosen, and R. C. Branski, *Classification manual for voice disorders-I.* New York: Psychology Press, 2013.

[16] N. Roy, R. M. Merrill, S. D. Gray, and E. M. Smith, "Voice disorders in the general population: prevalence, risk factors, and occupational impact," *Laryngoscope*, vol. 115, no. 11, pp. 1988–1995, 2005. doi: 10.1097/01.mlg.0000179174.32345.41.

[17] J. H. Hah, S. Sim, S.-Y. An, M.-W. Sung, and H. G. Choi, "Evaluation of the prevalence of and factors associated with laryngeal diseases among the general population," *Laryngoscope*, vol. 125, no. 11, pp. 2536–2542, 2015. doi: 10.1002/lary.25424.

[18] M. S. Benninger, A. S. Ahuja, G. Gardner, and C. Grywalski, "Assessing outcomes for dysphonic patients," *J. Voice*, vol. 12, no. 4, pp. 540–550, 1998. doi: 10.1016/s0892-1997(98)80063-5.

[19] N. Mirza, C. Ruiz, E. D. Baum, and J. P. Staab, "The prevalence of major psychiatric pathologies in patients with voice disorders," *Ear Nose Throat J.*, vol. 82, no. 10, pp. 808–10, 812, 814, 2003.

[20] S. M. Cohen, W. D. Dupont, and M. S. Courey, "Quality-of-life impact of non-neoplastic voice disorders: a meta-analysis," *Ann. Otol. Rhinol. Laryngol.*, vol. 115, no. 2, pp. 128–134, 2006. doi: 10.1177/000348940611500209.

[21] D. O. Francis, M. E. McKiever, C. G. Garrett, B. Jacobson, and D. F. Penson, "Assessment of patient experience with unilateral vocal fold immobility: a preliminary study," *J. Voice*, vol. 28, no. 5, pp. 636–643, 2014. doi:

10.1016/j.jvoice.2014.01.006.

[22] A. Y. Chen, N. M. Schrag, M. Halpern, A. Stewart, and E. M. Ward, "Health insurance and stage at diagnosis of laryngeal cancer: does insurance type predict stage at diagnosis?" *Arch. Otolaryngol. Head. Neck Surg.*, vol. 133, no. 8, pp. 784–790, 2007. doi: 10.1001/archotol.133.8.784.

[23] A. Ma, K. K. Lau, and D. Thyagarajan, "Voice changes in parkinson's disease: What are they telling us?" *J. Clin. Neurosci.*, vol. 72, pp. 1–7, 2020. doi: 10.1016/j.jocn.2019.12.029.

[24] 小川 真, "音声障害の評価と治療", 日本耳鼻咽喉科学会会報, vol. 123, no. 7, pp. 587–591, 2020. doi: 10.3950/jibiinkoka.123.587.

[25] J. Kreiman, B. R. Gerratt, G. B. Kempster, A. Erman, and G. S. Berke, "Perceptual evaluation of voice quality: review, tutorial, and a framework for future research," *J. Speech Hear. Res.*, vol. 36, no. 1, pp. 21–40, 1993. doi: 10.1044/jshr.3601.21.

[26] 今泉 敏, "声の聴覚心理的評価", 第 2 版 声の検査法 基礎編, 日本音声言語医学会, Ed. 東京: 医師薬出版株式会社, 1994, pp. 151–172.

[27] 高橋 宏明, 城本 修, 平野 実, "声の聴覚的評価", 第 2 版 声の検査法 臨床編, 日本音声言語医学会, Ed. 東京: 医師薬出版株式会社, 1994, pp. 187–214.

[28] 今泉 敏, "評価尺度法による検査", 新編 声の検査法, 日本音声言語医学会, Ed. 東京: 医歯薬出版株式会社, 2009, pp. 235–249.

[29] M. Hirano, "Psycho-acoustic evaluation of voice," in *Clinical examination of voice : disorders of human communication.* New York: Springer-Verlag, 1981, pp. 81–84.

[30] 阿部 千佳, 土師 知行, 水田 匡信, 城本 修, "音声課題が与える聴覚心理的評価への影響について —持続母音と短文—", 音声言語医学, vol. 64, no. 1, pp. 1–9, 2023.

[31] 一色 信彦, "嗄声の分類記載法", 音声言語医学, vol. 7, no. 1, pp. 15–21, 1966.

[32] 楯谷 一郎, 二藤 隆春, 城本 修, 佐藤 剛史, 岸本 曜, 末廣 篤, "音声障害", 音声言語認定医・認定士テキスト, 日本音声言語医学会, Ed. 東京: インテルナ出版, 2022, pp. 15–54.

[33] J. Kreiman, B. R. Gerratt, and K. Precoda, "Listener experience and perception of voice quality," *J. Speech Hear. Res.*, vol. 33, no. 1, pp. 103–115, 1990. doi: 10.1044/jshr.3301.103.

[34] J. Kreiman, B. R. Gerratt, K. Precoda, and G. S. Berke, "Individual differences in voice quality perception," *J. Speech Hear. Res.*, vol. 35, no. 3, pp. 512–520,

1992. doi: 10.1044/jshr.3503.512.

[35] H. Yamaguchi, R. Shrivastav, M. L. Andrews, and S. Niimi, "A comparison of voice quality ratings made by japanese and american listeners using the GRBAS scale," *Folia Phoniatr. Logop.*, vol. 55, no. 3, pp. 147–157, 2003. doi: 10.1159/000070726.

[36] M. S. De Bodt, F. L. Wuyts, P. H. Van de Heyning, and C. Croux, "Test-retest study of the GRBAS scale: influence of experience and professional background on perceptual rating of voice quality," *J. Voice*, vol. 11, no. 1, pp. 74–80, 1997. doi: 10.1016/s0892-1997(97)80026-4.

[37] J. L. Sofranko, and R. A. Prosek, "The effect of levels and types of experience on judgment of synthesized voice quality," *J. Voice*, vol. 28, no. 1, pp. 24–35, 2014. doi: 10.1016/j.jvoice.2013.06.001.

[38] B. R. Gerratt, J. Kreiman, N. Antonanzas-Barroso, and G. S. Berke, "Comparing internal and external standards in voice quality judgments," *J. Speech Hear. Res.*, vol. 36, no. 1, pp. 14–20, 1993. doi: 10.1044/jshr.3601.14.

[39] P. H. Dejonckere, M. Remacle, E. Fresnel-Elbaz, V. Woisard, L. Crevier-Buchman, and B. Millet, "Differentiated perceptual evaluation of pathological voice quality: reliability and correlations with acoustic measurements," *Rev. Laryngol. Otol. Rhinol.*, vol. 117, no. 3, pp. 219–224, 1996.

[40] P. H. Dejonckere, "Assessment of voice and respiratory function," in *Surgery of Larynx and Trachea*, M. Remacle, and H. E. Eckel, Eds. Berlin, Heidelberg: Springer Berlin Heidelberg, 2010, pp. 11–26.

[41] G. B. Kempster, B. R. Gerratt, K. Verdolini Abbott, J. Barkmeier-Kraemer, and R. E. Hillman, "Consensus auditory-perceptual evaluation of voice: development of a standardized clinical protocol," *Am. J. Speech. Lang. Pathol.*, vol. 18, no. 2, pp. 124–132, 2009. doi: 10.1044/1058-0360(2008/08-0017).

[42] K. Kondo, M. Mizuta, Y. Kawai, T. Sogami, S. Fujimura, T. Kojima, C. Abe, R. Tanaka, O. Shiromoto, R. Uozumi, Y. Kishimoto, I. Tateya, K. Omori, and T. Haji, "Development and validation of the japanese version of the consensus Auditory-Perceptual evaluation of voice," *J. Speech Lang. Hear. Res.*, vol. 64, no. 12, pp. 4754–4761, 2021. doi: 10.1044/2021_JSLHR-21-00269.

[43] S. N. Awan, and N. Roy, "Toward the development of an objective index of dysphonia severity: a four-factor acoustic model," *Clin. Linguist. Phon.*, vol. 20, no. 1, pp. 35–49, 2006. doi: 10.1080/02699200400008353.

[44] J. Hillenbrand, and R. A. Houde, "Acoustic correlates of breathy vocal quality:

dysphonic voices and continuous speech," *J. Speech Hear. Res.*, vol. 39, no. 2, pp. 311–321, 1996. doi: 10.1044/jshr.3902.311.

[45] L. Moro-Velázquez, J. A. Gómez-García, J. I. Godino-Llorente, and G. Andrade-Miranda, "Modulation spectra morphological parameters: A new method to assess voice pathologies according to the GRBAS scale," *Biomed Res. Int.*, vol. 2015, 2015. doi: 10.1155/2015/259239.

[46] Y. Zhang, J. J. Jiang, S. M. Wallace, and L. Zhou, "Comparison of nonlinear dynamic methods and perturbation methods for voice analysis," *J. Acoust. Soc. Am.*, vol. 118, no. 4, pp. 2551–2560, 2005. doi: 10.1121/1.2005907.

[47] J. J. Jiang, Y. Zhang, and C. McGilligan, "Chaos in voice, from modeling to measurement," *J. Voice*, vol. 20, no. 1, pp. 2–17, 2006. doi: 10.1016/j.jvoice.2005.01.001.

[48] 粕谷 英樹, "音響分析による検査", 新編 声の検査法, 日本音声言語医学会, Ed. 東京: 医歯薬出版株式会社, 2009, pp. 209–234.

[49] S. Fujimura, T. Kojima, Y. Okanoue, K. Shoji, M. Inoue, K. Omori, and R. Hori, "Classification of voice disorders using a One-Dimensional convolutional neural network," *J. Voice*, vol. 36, no. 1, pp. 15–20, 2022. doi: 10.1016/j.jvoice.2020.02.009.

[50] E. Yumoto, W. J. Gould, and T. Baer, "Harmonics ‐ to ‐ noise ratio as an index of the degree of hoarseness," *J. Acoust. Soc. Am.*, vol. 71, no. 6, pp. 1544–1550, 1982. doi: 10.1121/1.387808.

[51] J. Kreiman, and B. R. Gerratt, "Perception of aperiodicity in pathological voice," *J. Acoust. Soc. Am.*, vol. 117, no. 4 Pt 1, pp. 2201–2211, 2005. doi: 10.1121/1.1858351.

[52] P. N. Carding, I. N. Steen, A. Webb, K. MacKenzie, I. J. Deary, and J. A. Wilson, "The reliability and sensitivity to change of acoustic measures of voice quality," *Clin. Otolaryngol. Allied Sci.*, vol. 29, no. 5, pp. 538–544, 2004. doi: 10.1111/j.1365-2273.2004.00846.x.

[53] Y. Maryn, P. Corthals, P. Van Cauwenberge, N. Roy, and M. D. Bodt, "Toward improved ecological validity in the acoustic measurement of overall voice quality: Combining continuous speech and sustained vowels," *J. Voice*, vol. 24, no. 5, pp. 540–555, 2010. doi: 10.1016/j.jvoice.2008.12.014.

[54] B. Barsties V Latoszek, Y. Maryn, E. Gerrits, and M. De Bodt, "The acoustic breathiness index (ABI): A multivariate acoustic model for breathiness," *J. Voice*, vol. 31, no. 4, pp. 511.e11–511.e27, 2017. doi:

10.1016/j.jvoice.2016.11.017.

[55] N. Sáenz-Lechón, J. I. Godino-Llorente, V. Osma-Ruiz, M. Blanco-Velasco, and F. Cruz-Roldán, "Automatic assessment of voice quality according to the GRBAS scale," *Conf. Proc. IEEE Eng. Med. Biol. Soc.*, vol. 2006, pp. 2478–2481, 2006. doi: 10.1109/IEMBS.2006.260603.

[56] Z. Wang, P. Yu, N. Yan, L. Wang, and M. L. Ng, "Automatic assessment of pathological voice quality using multidimensional acoustic analysis based on the GRBAS scale," *J. Signal Process. Syst.*, vol. 82, no. 2, pp. 241–251, 2016. doi: 10.1007/s11265-015-1016-2.

[57] J. M. Miramont, M. A. Colominas, and G. Schlotthauer, "Emulating perceptual evaluation of voice using scattering transform based features," *IEEE/ACM Trans. Audio, Speech Lang. Process.*, vol. 30, pp. 1892–1901, 2022. doi: 10.1109/TASLP.2022.3178239.

[58] J. A. Gómez-García, L. Moro-Velázquez, J. Mendes-Laureano, G. Castellanos-Dominguez, and J. I. Godino-Llorente, "Emulating the perceptual capabilities of a human evaluator to map the GRB scale for the assessment of voice disorders," *Eng. Appl. Artif. Intell.*, vol. 82, pp. 236–251, 2019. doi: 10.1016/j.engappai.2019.03.027.

[59] J. D. Arias-Londoño, J. I. Godino-Llorente, N. Sáenz-Lechón, V. Osma-Ruiz, and J. M. Gutiérrez-Arriola, "Automatic GRBAS assessment using complexity measures and a multiclass GMM-based detector," in *MAVEBA*, 2011, pp. 111–114.

[60] S. Xie, N. Yan, P. Yu, M. L. Ng, L. Wang, and Z. Ji, "Deep neural networks for voice quality assessment based on the GRBAS scale," in *Interspeech*, 2016, pp. 2656–2660. doi: 10.21437/interspeech.2016-986.

[61] J. D. Arias-Londoño, J. A. Gómez-García, and J. I. Godino-Llorente, "Multimodal and Multi-Output deep learning architectures for the automatic assessment of voice quality using the GRB scale," *IEEE J. Sel. Top. Signal Process.*, vol. 14, no. 2, pp. 413–422, 2020. doi: 10.1109/JSTSP.2019.2956410.

[62] T. Kojima, S. Fujimura, K. Hasebe, Y. Okanoue, O. Shuya, R. Yuki, K. Shoji, R. Hori, Y. Kishimoto, and K. Omori, "Objective assessment of pathological voice using artificial intelligence based on the GRBAS scale," *J. Voice*, 2021. doi: 10.1016/j.jvoice.2021.11.021.

[63] M. A. García, and A. L. Rosset, *"Deep neural network for automatic assessment of dysphonia*," *arXiv preprint arXiv:2202.12957*, 2022. doi:

10.48550/arxiv.2202.12957.

[64] Y. Bengio, A. Courville, and P. Vincent, "Representation learning: a review and new perspectives," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 35, no. 8, pp. 1798–1828, 2013. doi: 10.1109/TPAMI.2013.50.

[65] A. Baevski, Y. Zhou, A. Mohamed, and M. Auli, "wav2vec 2.0: A framework for Self-Supervised learning of speech representations," in *NeurIPS*, 2020, pp. 12 449–12 460.

[66] A. Gulati, J. Qin, C.-C. Chiu, N. Parmar, Y. Zhang, J. Yu, W. Han, S. Wang, Z. Zhang, Y. Wu, and R. Pang, "Conformer: Convolution-augmented Transformer for Speech Recognition," in *Interspeech*, 2020, pp. 5036–5040. doi: 10.21437/interspeech.2020-3015.

[67] A. van den Oord, S. Dieleman, H. Zen, K. Simonyan, O. Vinyals, A. Graves, N. Kalchbrenner, A. Senior, and K. Kavukcuoglu, *"WaveNet: A generative model for raw audio,"* *arXiv preprint arXiv:1609.03499*, 2016. doi: 10.48550/arxiv.1609.03499.

[68] Y. Jia, Y. Zhang, R. Weiss, Q. Wang, J. Shen, F. Ren, Z. Chen, P. Nguyen, R. Pang, I. Lopez Moreno, and Y. Wu, "Transfer learning from speaker verification to multispeaker Text-To-Speech synthesis," in *NeurIPS*, 2018.

[69] J. Kong, J. Kim, and J. Bae, "HiFi-GAN: Generative adversarial networks for efficient and high fidelity speech synthesis," in *NeurIPS*, 2020, pp. 17 022–17 033.

[70] Y. Xu, J. Du, L.-R. Dai, and C.-H. Lee, "An experimental study on speech enhancement based on deep neural networks," *IEEE Signal Process. Lett.*, vol. 21, no. 1, pp. 65–68, 2014. doi: 10.1109/LSP.2013.2291240.

[71] D. S. Williamson, Y. Wang, and D. Wang, "Complex ratio masking for monaural speech separation," *IEEE/ACM Trans Audio Speech Lang Process*, vol. 24, no. 3, pp. 483–492, 2016. doi: 10.1109/TASLP.2015.2512042.

[72] Y. Hu, Y. Liu, S. Lv, M. Xing, S. Zhang, Y. Fu, J. Wu, B. Zhang, and L. Xie, "DCCRN: Deep complex convolution recurrent network for phase-aware speech enhancement," in *Interspeech*, 2020, pp. 2472–2476. doi: 10.21437/interspeech.2020-2537.

[73] Y. Luo, and N. Mesgarani, "TasNet: Time-Domain audio separation network for Real-Time, Single-Channel speech separation," in *ICASSP*, 2018, pp. 696–700. doi: 10.1109/ICASSP.2018.8462116.

[74] D. Stoller, S. Ewert, and S. Dixon, "Wave-U-Net: A Multi-Scale neural network

for End-to-End audio source separation," in *ISMIR*, 2018, pp. 334–340.

[75] A. Nagrani, J. S. Chung, and A. Zisserman, "VoxCeleb: A Large-Scale Speaker Identification Dataset," in *Interspeech*, 2017, pp. 2616–2620. doi: 10.21437/Interspeech.2017-950.

[76] J. S. Chung, A. Nagrani, and A. Zisserman, "VoxCeleb2: Deep Speaker Recognition," in *Interspeech*, 2018, pp. 1086–1090. doi: 10.21437/interspeech.2018-1929.

[77] M. India, P. Safari, and J. Hernando, "Self Multi-Head Attention for Speaker Recognition," in *Interspeech*, 2019, pp. 4305–4309. doi: 10.21437/interspeech.2019-2616.

[78] D. Snyder, D. Garcia-Romero, D. Povey, and S. Khudanpur, "Deep neural network embeddings for text-independent speaker verification," in *Interspeech*, 2017, pp. 999–1003. doi: 10.21437/interspeech.2017-620.

[79] L. Wan, Q. Wang, A. Papir, and I. L. Moreno, "Generalized End-to-End loss for speaker verification," in *ICASSP*, 2018, pp. 4879–4883. doi: 10.1109/ICASSP.2018.8462665.

[80] Z. Fan, M. Li, S. Zhou, and B. Xu, "Exploring wav2vec 2.0 on speaker verification and language identification," in *Interspeech*, 2021, pp. 1509–1513. doi: 10.21437/interspeech.2021-1280.

[81] J. Lee, and H.-J. Choi, "Deep learning approaches for pathological voice detection using heterogeneous parameters," *IEICE Trans. Inf. Syst.*, vol. E103.D, no. 8, pp. 1920–1923, 2020. doi: 10.1587/transinf.2020EDL8031.

[82] J.-Y. Lee, "Experimental evaluation of deep learning methods for an intelligent pathological voice detection system using the saarbruecken voice database," *NATO Adv. Sci. Inst. Ser. E Appl. Sci.*, vol. 11, no. 15, p. 7149, 2021. doi: 10.3390/app11157149.

[83] H.-C. Hu, S.-Y. Chang, C.-H. Wang, K.-J. Li, H.-Y. Cho, Y.-T. Chen, C.-J. Lu, T.-P. Tsai, and O. K.-S. Lee, "Deep learning application for vocal fold disease prediction through voice recognition: Preliminary development study," *J. Med. Internet Res.*, vol. 23, no. 6, p. e25247, 2021. doi: 10.2196/25247.

[84] C. Zhou, Y. Wu, Z. Fan, X. Zhang, D. Wu, and Z. Tao, "Gammatone spectral latitude features extraction for pathological voice detection and classification," *Appl. Acoust.*, vol. 185, p. 108417, 2022. doi: 10.1016/j.apacoust.2021.108417.

[85] R. Maskeliūnas, A. Kulikajevas, R. Damaševičius, K. Pribuišis, N. Ulozaitė-Stanienė, and V. Uloza, "Lightweight deep learning model for assessment of substitution voicing and speech after laryngeal carcinoma surgery," *Cancers*,

vol. 14, no. 10, 2022. doi: 10.3390/cancers14102366.

[86] T. Drugman, T. Dubuisson, and T. Dutoit, "Phase-based information for voice pathology detection," in *ICASSP*, 2011, pp. 4612–4615. doi: 10.1109/ICASSP.2011.5947382.

[87] G. Degottex, and N. Obin, "Phase distortion statistics as a representation of the glottal source: application to the classification of voice qualities," in *Interspeech*, 2014, pp. 1633–1637. doi: 10.21437/interspeech.2014-387.

[88] M. Koutsogiannaki, O. Simantiraki, G. Degottex, and Y. Stylianou, "The importance of phase on voice quality assessment," in *Interspeech*, 2014, pp. 1653–1657. doi: 10.21437/interspeech.2014-391.

[89] P. Janbakhshi, and I. Kodrasi, "Experimental investigation on STFT phase representations for deep Learning-Based dysarthric speech detection," in *ICASSP*, 2022, pp. 6477–6481. doi: 10.1109/ICASSP43922.2022.9747205.

[90] A. J. Rozsypal, and B. F. Millar, "Perception of jitter and shimmer in synthetic vowels," *J. Phon.*, vol. 7, no. 4, pp. 343–355, 1979. doi: 10.1016/S0095-4470(19)31069-1.

[91] R. Yamasaki, A. Montagnoli, E. Z. Murano, E. Gebrim, A. Hachiya, J. V. Lopes da Silva, M. Behlau, and D. Tsuji, "Perturbation measurements on the degree of naturalness of synthesized vowels," *J. Voice*, vol. 31, no. 3, pp. 389.e1–389.e8, 2017. doi: 10.1016/j.jvoice.2016.09.020.

[92] Y. Zhang, and J. J. Jiang, "Acoustic analyses of sustained and running voices from patients with laryngeal pathologies," *J. Voice*, vol. 22, no. 1, pp. 1–9, 2008. doi: 10.1016/j.jvoice.2006.08.003.

[93] K. Omori, "Diagnosis of voice disorders," *Japan Med. Assoc. J.*, vol. 54, no. 4, pp. 248–253, 2011.

[94] H. Herzel, D. Berry, I. R. Titze, and M. Saleh, "Analysis of vocal disorders with methods from nonlinear dynamics," *J. Speech Hear. Res.*, vol. 37, no. 5, pp. 1008–1019, 1994. doi: 10.1044/jshr.3705.1008.

[95] H. Herzel, J. Holzfuss, Z. J. Kowalik, B. Pompe, and R. Reuter, "Detecting bifurcations in voice signals," in *Nonlinear Analysis of Physiological Data*. Springer Berlin Heidelberg, 1998, pp. 325–344. doi: 10.1007/978-3-642-71949-3_19.

[96] K. Gröchenig, *Foundations of Time-Frequency Analysis*. Springer Science & Business Media, 2001.

[97] J. C. Brown, "Calculation of a constant Q spectral transform," *J. Acoust. Soc.*

Am., vol. 89, no. 1, pp. 425–434, 1991. doi: 10.1121/1.400476.

[98] O. Rioul, and M. Vetterli, "Wavelets and signal processing," *IEEE Signal Process. Mag.*, vol. 8, no. 4, pp. 14–38, 1991. doi: 10.1109/79.91217.

[99] S. S. Stevens, J. Volkmann, and E. B. Newman, "A scale for the measurement of the psychological magnitude pitch," *J. Acoust. Soc. Am.*, vol. 8, no. 3, pp. 185–190, 1937. doi: 10.1121/1.1915893.

[100] S. S. Stevens, and J. Volkmann, "The relation of pitch to frequency: A revised scale," *Am. J. Psychol.*, vol. 53, no. 3, pp. 329–353, 1940. doi: 10.2307/1417526.

[101] S. Davis, and P. Mermelstein, "Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences," *IEEE Trans. Acoust.*, vol. 28, no. 4, pp. 357–366, 1980. doi: 10.1109/TASSP.1980.1163420.

[102] J. Andén, and S. Mallat, "Deep scattering spectrum," *IEEE Trans. Signal Process.*, vol. 62, no. 16, pp. 4114–4128, 2014. doi: 10.1109/TSP.2014.2326991.

[103] M. Huzaifah, *"Comparison of Time-Frequency representations for environmental sound classification using convolutional neural networks*," *arXiv preprint arXiv:1706.07156*, 2017. doi: 10.48550/arxiv.1706.07156.

[104] J. Shen, R. Pang, R. J. Weiss, M. Schuster, N. Jaitly, Z. Yang, Z. Chen, Y. Zhang, Y. Wang, R. Skerrv-Ryan, R. A. Saurous, Y. Agiomvrgiannakis, and Y. Wu, "Natural TTS synthesis by conditioning wavenet on MEL spectrogram predictions," in *ICASSP*, 2018, pp. 4779–4783. doi: 10.1109/ICASSP.2018.8461368.

[105] T. Arias-Vergara, P. Klumpp, J. C. Vasquez-Correa, E. Nöth, J. R. Orozco-Arroyave, and M. Schuster, "Multi-channel spectrograms for speech processing applications using deep learning methods," *Pattern Anal. Appl.*, vol. 24, no. 2, pp. 423–431, 2021. doi: 10.1007/s10044-020-00921-5.

[106] Z. Tüske, P. Golik, R. Schlüter, and H. Ney, "Acoustic modeling with deep neural networks using raw time signal for LVCSR," in *Interspeech*, 2014, pp. 890–894. doi: 10.21437/interspeech.2014-223.

[107] Y. Hoshen, R. J. Weiss, and K. W. Wilson, "Speech acoustic modeling from raw multichannel waveforms," in *ICASSP*, 2015, pp. 4624–4628. doi: 10.1109/ICASSP.2015.7178847.

[108] E. Zwicker, "Subdivision of the audible frequency range into critical bands (frequenzgruppen)," *J. Acoust. Soc. Am.*, vol. 33, no. 2, pp. 248–248, 1961. doi: 10.1121/1.1908630.

[109] B. C. Moore, and B. R. Glasberg, "Suggested formulae for calculating

auditory-filter bandwidths and excitation patterns," *J. Acoust. Soc. Am.*, vol. 74, no. 3, pp. 750–753, 1983. doi: 10.1121/1.389861.

[110] Y. Masuyama, K. Yatabe, Y. Koizumi, Y. Oikawa, and N. Harada, "Phase reconstruction based on recurrent phase unwrapping with deep neural networks," in *ICASSP*, 2020, pp. 826–830. doi: 10.1109/ICASSP40776.2020.9053234.

[111] K. K. Paliwal, and L. Alsteris, "Usefulness of phase spectrum in human speech perception," in *Eurospeech*, 2003, pp. 2117–2120. doi: 10.21437/eurospeech.2003-611.

[112] K. K. Paliwal, and L. D. Alsteris, "On the usefulness of STFT phase spectrum in human listening tests," *Speech Commun.*, vol. 45, no. 2, pp. 153–170, 2005. doi: 10.1016/j.specom.2004.08.001.

[113] K. Paliwal, K. Wójcicki, and B. Shannon, "The importance of phase in speech enhancement," *Speech Commun.*, vol. 53, no. 4, pp. 465–494, 2011. doi: 10.1016/j.specom.2010.12.003.

[114] K. Tan, and D. Wang, "Complex spectral mapping with a convolutional recurrent network for monaural speech enhancement," in *ICASSP*, 2019, pp. 6865–6869. doi: 10.1109/ICASSP.2019.8682834.

[115] A. Pandey, and D. Wang, "Exploring deep complex networks for complex spectrogram enhancement," in *ICASSP*, 2019, pp. 6885–6889. doi: 10.1109/ICASSP.2019.8682169.

[116] Z. Wang, K. Tan, and D. Wang, "Deep learning based phase reconstruction for speaker separation: A trigonometric perspective," in *ICASSP*, 2019, pp. 71–75. doi: 10.1109/ICASSP.2019.8683231.

[117] H. Kawahara, T. Irino, and M. Morise, "An interference-free representation of instantaneous frequency of periodic signals and its application to F0 extraction," in *ICASSP*, 2011, pp. 5420–5423. doi: 10.1109/ICASSP.2011.5947584.

[118] 森勢 将雅，"基本波検出に基づく f0 推定法の耐雑音性向上"，情報処理学会研究報告，vol. 2016-SLP-110，no. 5，pp. 1–6，2016.

[119] D. Friedman, "Instantaneous-frequency distribution vs. time: An interpretation of the phase structure of speech," in *ICASSP*, vol. 10, 1985, pp. 1121–1124. doi: 10.1109/ICASSP.1985.1168461.

[120] D. J. Nelson, "Cross-spectral methods for processing speech," *J. Acoust. Soc. Am.*, vol. 110, no. 5 Pt 1, pp. 2575–2592, 2001. doi: 10.1121/1.1402616.

[121] H. A. Murthy, and B. Yegnanarayana, "Group delay functions and its applications in speech technology," *Sadhana - Academy Proceedings in*

Engineering Sciences, vol. 36, no. 5, pp. 745–782, 2011. doi: 10.1007/s12046-011-0045-1.

[122] B. Bozkurt, L. Couvreur, and T. Dutoit, "Chirp group delay analysis of speech signals," *Speech Commun.*, vol. 49, no. 3, pp. 159–176, 2007. doi: 10.1016/j.specom.2006.12.004.

[123] B. Bozkurt, and L. Couvreur, "On the use of phase information for speech recognition," in *EUSIPCO*, 2005, pp. 1–4.

[124] E. Loweimi, S. M. Ahadi, and T. Drugman, "A new phase-based feature representation for robust speech recognition," in *ICASSP*, 2013, pp. 7155–7159. doi: 10.1109/icassp.2013.6639051.

[125] Y. Masuyama, K. Yatabe, and Y. Oikawa, "Phase-aware harmonic/percussive source separation via convex optimization," in *ICASSP*, 2019, pp. 985–989. doi: 10.1109/ICASSP.2019.8683821.

[126] N. Akaishi, K. Yatabe, and Y. Oikawa, "Harmonic and percussive sound separation based on mixed partial derivative of phase spectrogram," in *ICASSP*, 2022, pp. 301–305. doi: 10.1109/ICASSP43922.2022.9747057.

[127] V. Kandia, and Y. Stylianou, "Detection of clicks based on group delay," *Can. Acoustics*, vol. 36, no. 1, pp. 48–54, 2008.

[128] A. Holzapfel, and Y. Stylianou, "Beat tracking using group delay based onset detection," in *ISMIR*, 2008, pp. 653–658.

[129] A. Holzapfel, Y. Stylianou, A. C. Gedik, and B. Bozkurt, "Three dimensions of pitched instrument onset detection," *IEEE Trans. Audio Speech Lang. Processing*, vol. 18, no. 6, pp. 1517–1527, 2010. doi: 10.1109/TASL.2009.2036298.

[130] S. Böck, and G. Widmer, "Local group delay based vibrato and tremolo suppression for onset detection," in *ISMIR*, 2013, pp. 361–366.

[131] M. Morise, "D4C, a band-aperiodicity estimator for high-quality speech synthesis," *Speech Commun.*, vol. 84, pp. 57–65, 2016. doi: 10.1016/j.specom.2016.09.001.

[132] H. Kawahara, K.-I. Sakakibara, M. Morise, H. Banno, and T. Toda, "A modulation property of Time-Frequency derivatives of filtered phase and its application to aperiodicity and fo estimation," in *Interspeech*, 2017, pp. 424–428. doi: 10.21437/Interspeech.2017-436.

[133] 河原 英紀，榊原 健一，森勢 将雅，坂野 秀樹，"瞬時周波数および群遅延に基づく非周期成分推定法再考"，情報処理学会研究報告，vol. 2017-MUS-114，no. 6，pp.

1–6, 2017.

[134] S.-W. Fu, T.-Y. Hu, Y. Tsao, and X. Lu, "Complex spectrogram enhancement by convolutional neural network with multi-metrics learning," in *MLSP*, 2017, pp. 1–6. doi: 10.1109/MLSP.2017.8168119.

[135] K. Yatabe, Y. Masuyama, T. Kusano, and Y. Oikawa, "Representation of complex spectrogram via phase conversion," *Acoust. Sci. Technol.*, vol. 40, no. 3, pp. 170–177, 2019. doi: 10.1250/ast.40.170.

[136] 矢田部 浩平, 升山 義紀, 草野 翼, 及川 靖広, "位相変換による複素スペクトログラムの表現", 日本音響学会誌, vol. 75, no. 3, pp. 147–155, 2019. doi: 10.20697/jasj.75.3_147.

[137] F. Léonard, "Phase spectrogram and frequency spectrogram as new diagnostic tools," *Mech. Syst. Signal Process.*, vol. 21, no. 1, pp. 125–137, 2007. doi: 10.1016/j.ymssp.2005.08.011.

[138] 小野 順貴, "短時間フーリエ変換の基礎と応用", 日本音響学会誌, vol. 72, no. 12, pp. 764–769, 2016. doi: 10.20697/jasj.72.12_764.

[139] 矢田部 浩平, "第三回：短時間フーリエ変換", 日本音響学会誌, vol. 77, no. 6, pp. 396–403, 2021. doi: 10.20697/jasj.77.6_396.

[140] ——, "第五回：実装における諸注意", 日本音響学会誌, vol. 77, no. 8, pp. 537–544, 2021. doi: 10.20697/jasj.77.8_537.

[141] F. Auger, and P. Flandrin, "Improving the readability of Time-Frequency and Time-Scale representations by the reassignment method," *IEEE Trans. Signal Process.*, vol. 43, no. 5, pp. 1068–1089, 1995. doi: 10.1109/78.382394.

[142] 矢田部 浩平, "第六回：時間周波数領域のスパース表現", 日本音響学会誌, vol. 77, no. 9, pp. 609–616, 2021. doi: 10.20697/jasj.77.9_609.

[143] P. Balazs, D. Bayer, F. Jaillet, and P. Søndergaard, "The pole behavior of the phase derivative of the short-time fourier transform," *Appl. Comput. Harmon. Anal.*, vol. 40, no. 3, pp. 610–621, 2016. doi: 10.1016/j.acha.2015.10.001.

[144] S. Umesh, L. Cohen, and D. Nelson, "Fitting the mel scale," in *ICASSP*, 1999, pp. 217–220. doi: 10.1109/ICASSP.1999.758101.

[145] B. McFee, C. Raffel, D. Liang, D. P. Ellis, M. McVicar, E. Battenberg, and O. Nieto, "librosa: Audio and music signal analysis in python," in *Proc. of the 14th Python in Science Conf.*, vol. 8, 2015, pp. 18–25.

[146] M. Slaney, "Auditory toolbox: A MATLAB toolbox for auditory modeling work version 2," Interval Research Corporation, Tech. Rep. 1998-010, 1998.

[147] S. Young, G. Evermann, M. Gales, T. Hain, D. Kershaw, X. Liu, G. Moore,

J. Odell, D. Ollason, D. Povey, and Others, *The HTK book*, 3rd ed. Cambridge: Cambridge University, 2006.

[148] D. O'shaughnessy, *Speech communications: Human and machine*. New York: Addison-Wesley, 1987.

[149] N. Zeghidour, N. Usunier, I. Kokkinos, T. Schaiz, G. Synnaeve, and E. Dupoux, "Learning filterbanks from raw speech for phone recognition," in *ICASSP*, 2018, pp. 5509–5513. doi: 10.1109/ICASSP.2018.8462015.

[150] H. Purwins, B. Li, T. Virtanen, J. Schlüter, S.-Y. Chang, and T. Sainath, "Deep learning for audio signal processing," *IEEE J. Sel. Top. Signal Process.*, vol. 13, no. 2, pp. 206–219, 2019. doi: 10.1109/JSTSP.2019.2908700.

[151] K. Choi, G. Fazekas, and M. Sandler, *"Automatic tagging using deep convolutional neural networks*," *arXiv preprint arXiv:1606.00298*, 2016. doi: 10.48550/arxiv.1606.00298.

[152] K. Venkataramanan, and H. R. Rajamohan, *"Emotion recognition from speech*," *arXiv preprint arXiv:1912.10458*, 2019. doi: 10.48550/arxiv.1912.10458.

[153] B. N. Suhas, J. Mallela, A. Illa, B. K. Yamini, N. Atchayaram, R. Yadav, D. Gope, and P. K. Ghosh, "Speech task based automatic classification of ALS and parkinson's disease and their severity using log mel spectrograms," in *SPCOM*, 2020, pp. 1–5. doi: 10.1109/SPCOM50965.2020.9179503.

[154] D. Ngo, H. Hoang, A. Nguyen, T. Ly, and L. Pham, *"Sound context classification basing on join learning model and Multi-Spectrogram features*," *arXiv preprint arXiv:2005.12779*, 2020. doi: 10.48550/arxiv.2005.12779.

[155] F. Rosenblatt, "The perceptron: A probabilistic model for information storage and organization in the brain," *Psychol. Rev.*, vol. 65, no. 6, pp. 386–408, 1958. doi: 10.1037/h0042519.

[156] W. S. McCulloch, and W. Pitts, "A logical calculus of the ideas immanent in nervous activity," *Bull. Math. Biophys.*, vol. 5, no. 4, pp. 115–133, 1943. doi: 10.1007/bf02478259.

[157] 岡谷 貴之，深層学習 改訂第 2 版．東京: 講談社，2022.

[158] N. S. Keskar, D. Mudigere, J. Nocedal, M. Smelyanskiy, and P. T. P. Tang, "On large-batch training for deep learning: Generalization gap and sharp minima," in *ICLR*, 2017.

[159] P. Goyal, P. Dollár, R. Girshick, P. Noordhuis, L. Wesolowski, A. Kyrola, A. Tulloch, Y. Jia, and K. He, *"Accurate, large minibatch SGD: Training ImageNet in 1 hour*," *arXiv preprint arXiv:1706.02677*, 2017. doi:

10.48550/arxiv.1706.02677.

[160] J. Duchi, E. Hazan, and Y. Singer, "Adaptive subgradient methods for online learning and stochastic optimization," *J. Mach. Learn. Res.*, vol. 12, no. 61, pp. 2121–2159, 2011.

[161] I. Sutskever, J. Martens, G. Dahl, and G. Hinton, "On the importance of initialization and momentum in deep learning," in *ICML*, 2013, pp. 1139–1147.

[162] S. Ruder, *"An overview of gradient descent optimization algorithms,"* *arXiv preprint arXiv:1609.04747*, 2016. doi: 10.48550/arxiv.1609.04747.

[163] A. Graves, *"Generating sequences with recurrent neural networks,"* *arXiv preprint arXiv:1308.0850*, 2013. doi: 10.48550/arxiv.1308.0850.

[164] D. P. Kingma, and J. Ba, *"Adam: A method for stochastic optimization,"* *arXiv preprint arXiv:1412.6980*, 2014. doi: 10.48550/arxiv.1412.6980.

[165] A. Ng, "Machine Learning — Andrew Ng, Stanford University [FULL COURSE] - YouTube," https://www.youtube.com/playlist?list=PLLssT5z_DsK-h9vYZkQkYNWcItqhlRJLN. Available at: https://www.youtube.com/playlist?list=PLLssT5z_DsK-h9vYZkQkYNWcItqhlRJLN. Accessed: 2023-1-26.

[166] D. R. Cox, "The regression analysis of binary sequences," *J. R. Stat. Soc.*, vol. 20, no. 2, pp. 215–232, 1958. doi: 10.1111/j.2517-6161.1958.tb00292.x.

[167] M. Ravanelli, and Y. Bengio, "Speaker recognition from raw waveform with SincNet," in *SLT*, 2018, pp. 1021–1028. doi: 10.1109/SLT.2018.8639585.

[168] N. Zeghidour, O. Teboul, F. de Chaumont Quitry, and M. Tagliasacchi, "LEAF: A learnable frontend for audio classification," in *ICLR*, 2021.

[169] T. Karras, M. Aittala, J. Hellsten, S. Laine, J. Lehtinen, and T. Aila, "Training generative adversarial networks with limited data," in *NeurIPS*, 2020, pp. 12 104–12 114.

[170] S. Zhao, Z. Liu, J. Lin, J.-Y. Zhu, and S. Han, "Differentiable augmentation for Data-Efficient GAN training," in *NeurIPS*, 2020, pp. 7559–7570.

[171] Y. Bengio, P. Simard, and P. Frasconi, "Learning Long-Term dependencies with gradient descent is difficult," *IEEE Trans. Neural Netw.*, vol. 5, no. 2, pp. 157–166, 1994. doi: 10.1109/72.279181.

[172] R. Pascanu, T. Mikolov, and Y. Bengio, "On the difficulty of training recurrent neural networks," in *ICML*, 2013, pp. 1310–1318.

[173] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov, "Dropout: A simple way to prevent neural networks from overfitting," *J. Mach.*

Learn. Res., vol. 15, no. 1, pp. 1929–1958, 2014.

[174] V. Nair, and G. E. Hinton, "Rectified linear units improve restricted boltzmann machines," in *ICML*, 2010, pp. 807–814.

[175] X. Glorot, A. Bordes, and Y. Bengio, "Deep sparse rectifier neural networks," in *AISTATS*, 2011, pp. 315–323.

[176] P. Ramachandran, B. Zoph, and Q. V. Le, *"Searching for activation functions*," *arXiv preprint arXiv:1710.05941*, 2017. doi: 10.48550/arxiv.1710.05941.

[177] I. Goodfellow, D. Warde-Farley, M. Mirza, A. Courville, and Y. Bengio, "Maxout networks," in *ICML*, 2013, pp. 1319–1327.

[178] D.-A. Clevert, T. Unterthiner, and S. Hochreiter, *"Fast and accurate deep network learning by exponential linear units (ELUs)*," *arXiv preprint arXiv:1511.07289*, 2015. doi: 10.48550/arxiv.1511.07289.

[179] D. Hendrycks, and K. Gimpel, *"Gaussian error linear units (GELUs)*," *arXiv preprint arXiv:1606.08415*, 2016. doi: 10.48550/arxiv.1606.08415.

[180] S. Ioffe, and C. Szegedy, "Batch normalization: Accelerating deep network training by reducing internal covariate shift," in *ICML*, 2015, pp. 448–456.

[181] S. Santurkar, D. Tsipras, A. Ilyas, and A. Madry, "How does batch normalization help optimization?" in *NeurIPS*, 2018.

[182] P. Luo, X. Wang, W. Shao, and Z. Peng, "Towards understanding regularization in batch normalization," in *ICLR*, 2019.

[183] S. De, and S. Smith, "Batch normalization biases residual blocks towards the identity function in deep networks," in *NeurIPS*, 2020, pp. 19 964–19 975.

[184] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *CVPR*, 2016, pp. 770–778.

[185] ——, "Identity mappings in deep residual networks," in *ECCV*, 2016, pp. 630–645. doi: 10.1007/978-3-319-46493-0_38.

[186] A. Veit, M. J. Wilber, and S. Belongie, "Residual networks behave like ensembles of relatively shallow networks," in *NIPS*, 2016.

[187] G. Huang, Y. Sun, Z. Liu, D. Sedra, and K. Q. Weinberger, "Deep networks with stochastic depth," in *ECCV*, 2016, pp. 646–661. doi: 10.1007/978-3-319-46493-0_39.

[188] J. L. Elman, "Finding structure in time," *Cogn. Sci.*, vol. 14, no. 2, pp. 179–211, 1990. doi: 10.1207/s15516709cog1402_1.

[189] P. J. Werbos, "Backpropagation through time: What it does and how to do it," *Proc. IEEE*, vol. 78, no. 10, pp. 1550–1560, 1990. doi: 10.1109/5.58337.

[190] M. Schuster, and K. K. Paliwal, "Bidirectional recurrent neural networks," *IEEE Trans. Signal Process.*, vol. 45, no. 11, pp. 2673–2681, 1997. doi: 10.1109/78.650093.

[191] A. Graves, and J. Schmidhuber, "Framewise phoneme classification with bidirectional LSTM and other neural network architectures," *Neural Netw.*, vol. 18, no. 5-6, pp. 602–610, 2005. doi: 10.1016/j.neunet.2005.06.042.

[192] S. Hochreiter, and J. Urgen Schmidhuber, "Long Short-Term memory," *Neural Comput.*, vol. 9, no. 8, pp. 1735–1780, 1997. doi: 10.1162/neco.1997.9.8.1735.

[193] F. A. Gers, J. Schmidhuber, and F. Cummins, "Learning to forget: continual prediction with LSTM," *Neural Comput.*, vol. 12, no. 10, pp. 2451–2471, 2000. doi: 10.1162/089976600300015015.

[194] K. Cho, B. van Merrienboer, C. Gulcehre, D. Bahdanau, F. Bougares, H. Schwenk, and Y. Bengio, "Learning phrase representations using RNN encoder–decoder for statistical machine translation," in *EMNLP*, 2014, pp. 1724–1734. doi: 10.3115/v1/d14-1179.

[195] M. Ravanelli, P. Brakel, M. Omologo, and Y. Bengio, "Light gated recurrent units for speech recognition," *IEEE Transactions on Emerging Topics in Computational Intelligence*, vol. 2, no. 2, pp. 92–102, 2018. doi: 10.1109/TETCI.2017.2762739.

[196] Y. Su, and C.-C. J. Kuo, "On extended long short-term memory and dependent bidirectional recurrent neural network," *Neurocomputing*, vol. 356, pp. 151–161, 2019. doi: 10.1016/j.neucom.2019.04.044.

[197] D. Ulyanov, A. Vedaldi, and V. Lempitsky, "Deep image prior," in *CVPR*, 2018, pp. 9446–9454. doi: 10.1109/cvpr.2018.00984.

[198] Z. Huang, M. Dong, Q. Mao, and Y. Zhan, "Speech emotion recognition using CNN," in *ACM Multimedia*, 2014, pp. 801–804. doi: 10.1145/2647868.2654984.

[199] M. S. Hossain, and G. Muhammad, "Emotion recognition using deep learning approach from audio–visual emotional big data," *Inf. Fusion*, vol. 49, pp. 69–78, 2019. doi: 10.1016/j.inffus.2018.09.008.

[200] P. Heracleous, K. Takai, K. Yasuda, Y. Mohammad, and A. Yoneyama, "Comparative study on spoken language identification based on deep learning," in *EUSIPCO*, 2018, pp. 2265–2269. doi: 10.23919/EUSIPCO.2018.8553347.

[201] S. Revay, and M. Teschke, *"Multiclass language identification using deep learning on spectral images of audio signals*," *arXiv preprint arXiv:1905.04348*, 2019. doi: 10.48550/arxiv.1905.04348.

[202] A. Jansson, E. J. Humphrey, N. Montecchio, R. M. Bittner, A. Kumar, and T. Weyde, "Singing voice separation with deep U-Net convolutional networks," in *ISMIR*, 2017, pp. 745–751.

[203] M. A. Islam, S. Jia, and N. D. B. Bruce, "How much position information do convolutional neural networks encode?" in *ICLR*, 2020.

[204] A. G. Howard, M. Zhu, B. Chen, D. Kalenichenko, W. Wang, T. Weyand, M. Andreetto, and H. Adam, *"MobileNets: Efficient convolutional neural networks for mobile vision applications*," *arXiv preprint arXiv:1704.04861*, 2017. doi: 10.48550/arxiv.1704.04861.

[205] K. Simonyan, and A. Zisserman, *"Very deep convolutional networks for Large-Scale image recognition*," *arXiv preprint arXiv:1409.1556*, 2014. doi: 10.48550/arxiv.1409.1556.

[206] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich, "Going deeper with convolutions," in *CVPR*. IEEE, 2015, pp. 1–9. doi: 10.1109/cvpr.2015.7298594.

[207] D. D. Deliyski, H. S. Shaw, and M. K. Evans, "Adverse effects of environmental noise on acoustic voice quality measurements," *J. Voice*, vol. 19, no. 1, pp. 15–28, 2005. doi: 10.1016/j.jvoice.2004.07.003.

[208] K. M. K. Chan, and E. M.-L. Yiu, "The effect of anchors and training on the reliability of perceptual voice evaluation," *J. Speech Lang. Hear. Res.*, vol. 45, no. 1, pp. 111–126, 2002. doi: 10.1044/1092-4388(2002/009).

[209] S. N. Awan, and L. L. Lawson, "The effect of anchor modality on the reliability of vocal severity ratings," *J. Voice*, vol. 23, no. 3, pp. 341–352, 2009. doi: 10.1016/j.jvoice.2007.10.006.

[210] T. L. Eadie, and M. Kapsner-Smith, "The effect of listener experience and anchors on judgments of dysphonia," *J. Speech Lang. Hear. Res.*, vol. 54, no. 2, pp. 430–447, 2011. doi: 10.1044/1092-4388(2010/09-0205).

[211] 河原 一彦，"聴能形成の訓練システムと運用の改善と展開"，Ph.D. dissertation，九州大学，2017.

[212] J. Peirce, J. R. Gray, S. Simpson, M. MacAskill, R. Höchenberger, H. Sogo, E. Kastman, and J. K. Lindeløv, "PsychoPy2: Experiments in behavior made easy," *Behav. Res. Methods*, vol. 51, no. 1, pp. 195–203, 2019. doi: 10.3758/s13428-018-01193-y.

[213] 日本音声言語医学会，動画で見る音声障害 ver. 2.0（DVD-ROM）．東京: インテルナ出版，2018.

[214] 日本音声言語医学会発声機能検査法検討委員会，嗄声のサンプルテープ．東京: メディカルリサーチセンター，1981.

[215] J. Cohen, “A coefficient of agreement for nominal scales,” *Educ. Psychol. Meas.*, vol. 20, no. 1, pp. 37–46, 1960. doi: 10.1177/001316446002000104.

[216] ——, “Weighted kappa: Nominal scale agreement provision for scaled disagreement or partial credit,” *Psychol. Bull.*, vol. 70, no. 4, pp. 213–220, 1968. doi: 10.1037/h0026256.

[217] K. L. Gwet, *Handbook of Inter-Rater Reliability, 4th Edition: The Definitive Guide to Measuring The Extent of Agreement Among Raters*. Advanced Analytics, LLC, 2014.

[218] A. J. Conger, “Integration and generalization of kappas for multiple raters,” *Psychol. Bull.*, vol. 88, no. 2, pp. 322–328, 1980. doi: 10.1037/0033-2909.88.2.322.

[219] G. de Krom, “A cepstrum-based technique for determining a harmonics-to-noise ratio in speech signals,” *J. Speech Hear. Res.*, vol. 36, no. 2, pp. 254–266, 1993. doi: 10.1044/jshr.3602.254.

[220] İ. Bayram, and M. E. Kamasak, “A simple prior for audio signals,” *IEEE Trans. Audio Speech Lang. Processing*, vol. 21, no. 6, pp. 1190–1200, 2013. doi: 10.1109/TASL.2013.2245652.

[221] K. Yatabe, and Y. Oikawa, “Phase corrected total variation for audio signals,” in *ICASSP*, 2018, pp. 656–660. doi: 10.1109/ICASSP.2018.8461541.

[222] A. Paszke, S. Gross, F. Massa, A. Lerer, J. Bradbury, G. Chanan, T. Killeen, Z. Lin, N. Gimelshein, L. Antiga, A. Desmaison, A. Kopf, E. Yang, Z. DeVito, M. Raison, A. Tejani, S. Chilamkurthy, B. Steiner, L. Fang, J. Bai, and S. Chintala, “PyTorch: An imperative style, High-Performance deep learning library,” in *NeurIPS*, 2019, pp. 8024–8035.

[223] J.-W. Jung, H.-J. Shim, H.-S. Heo, and H.-J. Yu, “Replay attack detection with complementary high-resolution information using end-to-end DNN for the ASVspoof 2019 challenge,” in *Interspeech 2019*, Sep. 2019. doi: 10.21437/interspeech.2019-1991.

[224] J. Yang, H. Wang, R. K. Das, and Y. Qian, “Modified Magnitude-Phase spectrum information for spoofing detection,” *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 29, pp. 1065–1078, 2021. doi: 10.1109/TASLP.2021.3060810.

[225] “LEAF: a LEarnable audio frontend,” https://github.com/google-research/leaf-

audio. Available at: https://github.com/google-research/leaf-audio. Accessed: 2021-7-11.

[226] Y. Wang, P. Getreuer, T. Hughes, R. F. Lyon, and R. A. Saurous, "Trainable frontend for robust and far-field keyword spotting," in *ICASSP*, 2017, pp. 5670–5674. doi: 10.1109/ICASSP.2017.7953242.

[227] J. Engel, C. Resnick, A. Roberts, S. Dieleman, M. Norouzi, D. Eck, and K. Simonyan, "Neural audio synthesis of musical notes with wavenet autoencoders," in *ICML*, 2017, pp. 1068–1077.

[228] D. Stowell, M. D. Wood, H. Pamuła, Y. Stylianou, and H. Glotin, "Automatic acoustic detection of birds through deep learning: The first bird audio detection challenge," *Methods Ecol. Evol.*, vol. 10, no. 3, pp. 368–380, 2018. doi: 10.1111/2041-210x.13103.

[229] T. Heittola, A. Mesaros, and T. Virtanen, "TUT urban acoustic scenes 2018, development dataset," 2018. Available at: https://doi.org/10.5281/zenodo.1228142.

[230] P. Warden, *"Speech commands: A dataset for Limited-Vocabulary speech recognition,"* *arXiv preprint arXiv:1804.03209*, 2018. doi: 10.48550/arxiv.1804.03209.

[231] H. Cao, D. G. Cooper, M. K. Keutmann, R. C. Gur, A. Nenkova, and R. Verma, "CREMA-D: Crowd-sourced emotional multimodal actors dataset," *IEEE Trans Affect Comput*, vol. 5, no. 4, pp. 377–390, 2014. doi: 10.1109/TAFFC.2014.2336244.

[232] M. Tan, and Q. Le, "EfficientNet: Rethinking model scaling for convolutional neural networks," in *ICML*, 2019, pp. 6105–6114.

[233] C. Trabelsi, O. Bilaniuk, D. Serdyuk, S. Subramanian, J. F. Santos, S. Mehri, N. Rostamzadeh, Y. Bengio, and C. J. Pal, "Deep complex networks," in *ICLR*, 2018.

[234] D. S. Park, W. Chan, Y. Zhang, C.-C. Chiu, B. Zoph, E. D. Cubuk, and Q. V. Le, "SpecAugment: A simple data augmentation method for automatic speech recognition," in *Interspeech*, 2019, pp. 2613–2617. doi: 10.21437/Interspeech.2019-2680.

[235] R. Müller, S. Kornblith, and G. Hinton, *"When does label smoothing help?"* *arXiv preprint arXiv:1906.02629*, 2019. doi: 10.48550/arxiv.1906.02629.

[236] Y. N. Dauphin, A. Fan, M. Auli, and D. Grangier, "Language modeling with gated convolutional networks," in *ICML*, D. Precup, and Y. W. Teh, Eds., 2017, pp. 933–941.

[237] T. Ko, V. Peddinti, D. Povey, and S. Khudanpur, "Audio augmentation for speech recognition," in *Interspeech*, 2015, pp. 3586–3589. doi: 10.21437/Interspeech.2015-711.

[238] T. Fujimura, Y. Koizumi, K. Yatabe, and R. Miyazaki, *"Noisy-target training: A training strategy for DNN-based speech enhancement without clean speech,"* *arXiv preprint arXiv:2101.08625*, 2021. doi: 10.48550/arxiv.2101.08625.

[239] M. Tan, and Q. Le, "EfficientNetV2: Smaller models and faster training," in *ICML*, 2021, pp. 10 096–10 106.

[240] R. Wightman, "PyTorch image models," 2019. Available at: https://github.com/rwightman/pytorch-image-models.

[241] B.-B. Gao, C. Xing, C.-W. Xie, J. Wu, and X. Geng, "Deep label distribution learning with label ambiguity," *IEEE Trans. Image Process.*, vol. 26, no. 6, pp. 2825–2838, 2017. doi: 10.1109/TIP.2017.2689998.

[242] S. Kullback, and R. A. Leibler, "On information and sufficiency," *Ann. Math. Stat.*, vol. 22, no. 1, pp. 79–86, 1951. doi: 10.1214/aoms/1177729694.

[243] K. Krippendorff, "Estimating the reliability, systematic error and random error of interval data," *Educ. Psychol. Meas.*, vol. 30, no. 1, pp. 61–70, 1970. doi: 10.1177/001316447003000105.

[244] M. Geng, X. Xie, S. Liu, J. Yu, S. Hu, X. Liu, and H. Meng, "Investigation of data augmentation techniques for disordered speech recognition," in *Interspeech*, 2020, pp. 696–700. doi: 10.21437/Interspeech.2020-1161.

[245] G. de Krom, "Consistency and reliability of voice quality ratings for different types of speech fragments," *J. Speech Hear. Res.*, vol. 37, no. 5, pp. 985–1000, 1994. doi: 10.1044/jshr.3705.985.

[246] J. Revis, A. Giovanni, F. Wuyts, and J. M. Triglia, "Comparison of different voice samples for perceptual analysis," *Folia Phoniatr. Logop.*, vol. 51, no. 3, pp. 108–116, 1999. doi: 10.1159/000021485.

[247] M. Morise, F. Yokomori, and K. Ozawa, "WORLD: A Vocoder-Based High-Quality speech synthesis system for Real-Time applications," *IEICE Trans. Inf. Syst.*, vol. E99.D, no. 7, pp. 1877–1884, 2016. doi: 10.1587/transinf.2015edp7457.

[248] G. Ke, Q. Meng, T. Finley, T. Wang, W. Chen, W. Ma, Q. Ye, and T.-Y. Liu, "LightGBM: A highly efficient gradient boosting decision tree," in *NIPS*, 2017, pp. 3149–3157.

[249] P. Harar, Z. Galaz, J. B. Alonso-Hernandez, J. Mekyska, R. Burget, and

Z. Smekal, “Towards robust voice pathology detection,” *Neural Comput. Appl.*, vol. 32, no. 20, pp. 15 747–15 757, 2020. doi: 10.1007/s00521-018-3464-7.

[250] G. V. D. Kumar, V. Deepa, N. Vineela, and G. Emmanuel, “Detection of parkinson’s disease using LightGBM classifier,” in *ICCMC*, 2022, pp. 1292–1297. doi: 10.1109/ICCMC53470.2022.9753909.

[251] T. Akiba, S. Sano, T. Yanase, T. Ohta, and M. Koyama, “Optuna: A next-generation hyperparameter optimization framework,” in *KDD*, 2019, pp. 2623–2631. doi: 10.1145/3292500.3330701.

[252] J. Bergstra, R. Bardenet, Y. Bengio, and B. Kégl, “Algorithms for hyper-parameter optimization,” in *NIPS*, 2011, pp. 1–9.

[253] S. Anand, M. D. Skowronski, R. Shrivastav, and D. A. Eddins, “Perceptual and quantitative assessment of dysphonia across vowel categories,” *J. Voice*, vol. 33, no. 4, pp. 473–481, 2019. doi: 10.1016/j.jvoice.2017.12.018.

[254] P. R. Walden, “Perceptual voice qualities database (PVQD): Database characteristics,” *J. Voice*, 2020. doi: 10.1016/j.jvoice.2020.10.001.

[255] T. Sakata, N. Kubota, H. Yonekawa, S. Imaizumi, and S. Niimi, “GRBAS evaluation of running speech and sustained phonations,” *Ann, Bull. RILP*, vol. 28, pp. 51–56, 1994.

[256] V. Wolfe, R. Cornell, and J. Fitch, “Sentence/vowel correlation in the evaluation of dysphonia,” *J. Voice*, vol. 9, no. 3, pp. 297–303, 1995. doi: 10.1016/s0892-1997(05)80237-1.

[257] F.-L. Lu, and S. Matteson, “Speech tasks and interrater reliability in perceptual voice evaluation,” *J. Voice*, vol. 28, no. 6, pp. 725–732, 2014. doi: 10.1016/j.jvoice.2014.01.018.

[258] H. Yamamoto, K. A. Lee, K. Okabe, and T. Koshinaka, “Speaker augmentation and bandwidth extension for deep speaker embedding,” in *Interspeech*, 2019, pp. 406–410. doi: 10.21437/Interspeech.2019-1508.

[259] C.-P. Chen, S.-Y. Zhang, C.-T. Yeh, J.-C. Wang, T. Wang, and C.-L. Huang, “Speaker characterization using TDNN-LSTM based speaker embedding,” in *ICASSP*, 2019, pp. 6211–6215. doi: 10.1109/ICASSP.2019.8683185.

[260] 日本音声言語医学会，日本音声言語医学会 40 年史．東京: 日本音声言語医学会，1995.

[261] S. Mahalingam, Y. Venkatraman, and P. Boominathan, “Cross-Cultural adaptation and validation of consensus auditory perceptual evaluation of voice (CAPE-V): A systematic review,” *J. Voice*, 2021. doi:

10.1016/j.jvoice.2021.10.022.

[262] V. Parsa, and D. G. Jamieson, "Acoustic discrimination of pathological voice: sustained vowels versus continuous speech," *J. Speech Lang. Hear. Res.*, vol. 44, no. 2, pp. 327–339, 2001. doi: 10.1044/1092-4388(2001/027).

[263] Y. Maryn, and N. Roy, "Sustained vowels and continuous speech in the auditory-perceptual evaluation of dysphonia severity," *J. Soc. Bras. Fonoaudiol.*, vol. 24, no. 2, pp. 107–112, 2012. doi: 10.1590/S2179-64912012000200003.

[264] S. N. Awan, N. Roy, and C. Dromey, "Estimating dysphonia severity in continuous speech: application of a multi-parameter spectral/cepstral model," *Clin. Linguist. Phon.*, vol. 23, no. 11, pp. 825–841, 2009. doi: 10.3109/02699200903242988.

[265] M. E. A. Infirmary, *Voice Disorders Database, Version 1.03.* Lincoln Park, NJ: Kay Elemetrics Corp., 1994.

[266] M. Pützer, and W. J. Barry, "Saarbruecken voice database," http://www.stimmdatenbank.coli.uni-saarland.de/. Available at: http://www.stimmdatenbank.coli.uni-saarland.de/. Accessed: 2022-7-1.

[267] P. Ren, Y. Xiao, X. Chang, P.-Y. Huang, Z. Li, B. B. Gupta, X. Chen, and X. Wang, "A survey of deep active learning," *ACM Comput. Surv.*, vol. 54, no. 9, pp. 1–40, 2021. doi: 10.1145/3472291.

[268] T. Kadota, H. Hayashi, R. Bise, K. Tanaka, and S. Uchida, "Deep bayesian Active-Learning-to-Rank for endoscopic image data," in *MIUA*, 2022, pp. 609–622. doi: 10.1007/978-3-031-12053-4_45.

[269] M. Tan, B. Chen, R. Pang, V. Vasudevan, M. Sandler, A. Howard, and Q. V. Le, "MnasNet: Platform-Aware neural architecture search for mobile," in *CVPR*, 2019, pp. 2815–2823. doi: 10.1109/CVPR.2019.00293.

[270] J. Hu, L. Shen, and G. Sun, "Squeeze-and-Excitation networks," in *CVPR*, 2018, pp. 7132–7141. doi: 10.1109/CVPR.2018.00745.

[271] J. Hu, L. Shen, S. Albanie, G. Sun, and E. Wu, "Squeeze-and-Excitation networks," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 42, no. 8, pp. 2011–2023, 2020. doi: 10.1109/TPAMI.2019.2913372.

[272] A. Howard, M. Sandler, B. Chen, W. Wang, L.-C. Chen, M. Tan, G. Chu, V. Vasudevan, Y. Zhu, R. Pang, H. Adam, and Q. Le, "Searching for MobileNetV3," in *ICCV*, 2019, pp. 1314–1324. doi: 10.1109/ICCV.2019.00140.

[273] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. U. Kaiser, and I. Polosukhin, "Attention is all you need," in *NIPS*, 2017.

[274] M. Sandler, A. Howard, M. Zhu, A. Zhmoginov, and L.-C. Chen, "MobileNetV2: Inverted residuals and linear bottlenecks," in *CVPR*, 2018, pp. 4510–4520. doi: 10.1109/CVPR.2018.00474.

[275] "EfficientNetV2," https://github.com/google/automl/tree/master/efficientnetv2. Available at: https://github.com/google/automl/tree/master/efficientnetv2. Accessed: 2022-12-4.

[276] S. Anand, L. M. Kopf, R. Shrivastav, and D. A. Eddins, "Objective indices of perceived vocal strain," *J. Voice*, vol. 33, no. 6, pp. 838–845, 2019. doi: 10.1016/j.jvoice.2018.06.005.

[277] D. Michaelis, T. Gramss, and H. W. Strube, "Glottal-to-Noise excitation ratio - a new measure for describing pathological voices," *Acustica*, vol. 83, no. 4, pp. 700–706, 1997.

[278] P. H. Dejonckere, "Recognition of hoarseness by means of LTAS," *Int. J. Rehabil. Res.*, vol. 7, pp. 73–74, 1983.

[279] P. H. Dejonckere, and J. Lebacq, "Harmonic emergence in formant zone of a sustained [a] as a parameter for evaluating hoarseness," *Acta Otorhinolaryngol. Belg.*, vol. 41, no. 6, pp. 988–996, 1987.

[280] J. P. Teixeira, and A. Gonçalves, "Algorithm for jitter and shimmer measurement in pathologic voices," *Procedia Comput. Sci.*, vol. 100, pp. 271–279, 2016. doi: 10.1016/j.procs.2016.09.155.

[281] International Organization for Standardization, "Acoustics — methods for calculating loudness — part 1: Zwicker method," Tech. Rep. ISO 532-1:2017, 2017.

[282] 難波 精一郎, "知っているようで知らないラウドネス", 日本音響学会誌, vol. 73, no. 12, pp. 765–773, 2017. doi: 10.20697/jasj.73.12_765.

[283] S. S. Stevens, "The measurement of loudness," *J. Acoust. Soc. Am.*, vol. 27, no. 5, pp. 815–829, 1955. doi: 10.1121/1.1908048.

[284] H. Fastl, and E. Zwicker, "Sharpness and sensory pleasantness," in *Psychoacoustics: Facts and Models*, H. Fastl, and E. Zwicker, Eds. Berlin, Heidelberg: Springer Berlin Heidelberg, 2007, pp. 239–246.

[285] G. F. Coop, "MOSQITO," https://zenodo.org/record/6675733, 2022. Available at: https://zenodo.org/record/6675733.

[286] Deutsches Institut für Normung, "Measurement technique for the simulation of the auditory sensation of sharpness," Tech. Rep. DIN 45692, 2009.

[287] S. Hidaka, Y. Lee, M. Nakanishi, K. Wakamiya, T. Nakagawa, and T. Kaburagi, "Automatic GRBAS scoring of pathological voices using deep

learning and a small set of labeled voice data," *J. Voice*, 2022. doi: 10.1016/j.jvoice.2022.10.020.