2019 年度 修士論文

日本人の自然発話を対象とした音声感情認識

―感情音声の訓練サンプル数最適化を用いて―

提出日： 2020 年 1 月 29 日
指導： 山名 早人 教授

早稲田大学 基幹理工学研究科 情報理工・情報通信専攻
学籍番号：5118F008

市川 朋輝

概　要

昨今，人間と機械のコミュニケーションが増加している．人間の場合，発話者の思考や感情，顔の表情，体調などを無意識下に感じ取り，こうした情報も考慮しながらコミュニケーションをとっている．したがって，自然発話音声の感情認識は，人間と機械のコミュニケーションにおいて重要な研究課題である．また，日本人の発話や表情には感情が出にくいことが特徴として挙げられる．従来研究では，データ数が少ないため一部の感情の認識率が高く，他方の感情の認識率が低くなるモデルとなっている．そこで本研究は，人間も徐々に他人の感情を理解できるようになる点，人間にとっても比較的分かりづらい感情が存在する点に注目した．人間も徐々に他人の感情を理解できるようになる点より，訓練サンプル数の増加（アップサンプリング）し，たくさんの感情音声により訓練することが重要だと考えた．また，人間にとっても比較的分かりづらい感情が存在する点より，認識しやすい感情の訓練を少なく，認識しにくい感情の訓練を多くするため感情音声サンプル数の最適化を考えた．学習データのアップサンプリング，感情音声サンプル数の最適化により音声感情認識精度の向上・感情認識率のばらつき（標準偏差）の低下を目指した．提案手法により，アップサンプリング・サンプル数最適化前と比較して，感情認識の Accuracy が 25.8%から 36.0%，認識率の標準偏差が 24.0 から 17.5 となった．

目次

第 1 章　はじめに 1
第 2 章　関連研究 2
2.1　概要 2
2.2　音声感情認識 2
2.2.1　入力データ 3
2.2.2　音声の特徴量抽出 3
2.2.3　識別器 7
2.2.4　感情分類 7
2.2.5　音声感情認識のまとめ 8
2.3　関連研究 9
2.3.1　阿部ら[9]の研究 14
2.3.2　Song ら[11]の研究 16
2.3.3　Zhao ら[15]の研究 18
2.3.4　Zhang ら[22]の研究 18
2.4　問題点 22
第 3 章　提案手法 23
3.1　概要 23
3.2　3 次元メルスペクトラム 26
3.3　CNN と LSTM の組み合わせ学習 27
3.3.1　CNN（Convolution Neural Network） 27
3.3.2　BiLSTM（Bidirectional Long Short-term Memory） 28
3.4　音声感情サンプルのアップサンプリング 30
3.4.1　SMOTE（Synthetic Minority Over-sampling Technique） 32
第 4 章　実験・評価 35
4.1　概要 35
4.2　データセット 35
4.2.1　自然発話音声 35
4.2.2　感情ラベル 37
4.3　実験条件 39
4.4　評価実験 41
4.4.1　感情音声のアップサンプリング評価 41
4.4.2　LSTM のハイパーパラメータ最適化実験 49
4.4.3　感情音声サンプル数の最適化実験 51
第 5 章　おわりに 61

第1章　　はじめに

人間はコミュニケーション行う際，発話内容を理解するだけではなく，発話者の思考や感情，顔の表情，体調などを無意識下に感じ取り，こうした情報も考慮しながらコミュニケーションを行っている．人間とコミュニケーションを行う機械においても，同じような能力が備わることでより人間的なコミュニケーションを行うことができると考えらえる．たとえば，スマートフォンに「リラックスできる場所はどこ？」と発話する際，発話の背景として，仕事で疲れたときなのか，休みの朝の元気が良いときなのか，などが考えられる．機械がこうした背景を認識することができれば，最適な提案を行うことができるようになる．実際にコールセンターや，メンタルヘルスケアに実用がされている．このように，人間と機械が音声対話コミュニケーションを行う際に，ユーザの発話に含まれる感情を認識する技術は非常に重要である．

また，日本人の発話や表情には感情が出にくいことが特徴として挙げられる．日本人の国民性として，「礼儀正しく」「親切」という価値観を持っている人が多いためである[1]．実際に表情において，日本人は感情の表出が他国と異なっていることが知られている[2]．本研究では，日本人の自然な発話には感情が表出しにくい点に注目し，日本人の自然発話を対象とした音声感情認識における新しい手法を提案する．具体的には，訓練に用いる音声感情のサンプル数の最適化をする．

本稿は以下の構成をとる．まず，第 2 章で関連研究について述べ，第 3 章で提案手法について説明する．そして，第 4 章で実験・評価を行い，第 5 章で本稿のまとめとする．

第2章 関連研究

2.1 概要

近年音声感情認識の研究は盛んに行われている．「ある人間の発話が，他の人間からどのような感情に聞こえるのか？」を分類することを目標としている．入力は人間の発話音声，もしくは発話音声から特徴量を抽出したものであり，出力は４～７種類（怒り，恐怖，楽しい等）の感情である．収録された音声を入力データとし，その音声に第三者が感情のアノテーションを正解ラベルとして用いる．本節では，音声感情認識の概略と最新研究を示す．

2.2 音声感情認識

音声感情認識の研究は大きく分けて 2 種類存在し，演技発話（セリフに感情を意図的に込めて発話した音声）を対象とした音声感情認識と，自然発話（ラジオやゲーム中の会話など）を対象とした音声感情認識である．演技発話・自然発話を対象とした音声感情認識の入力から出力の流れを図 2.1 に示す．2.2.1 で入力について，2.2.2 で特徴量抽出について，2.2.3 で識別器について，2.2.4 で感情分類についての概略を述べる．

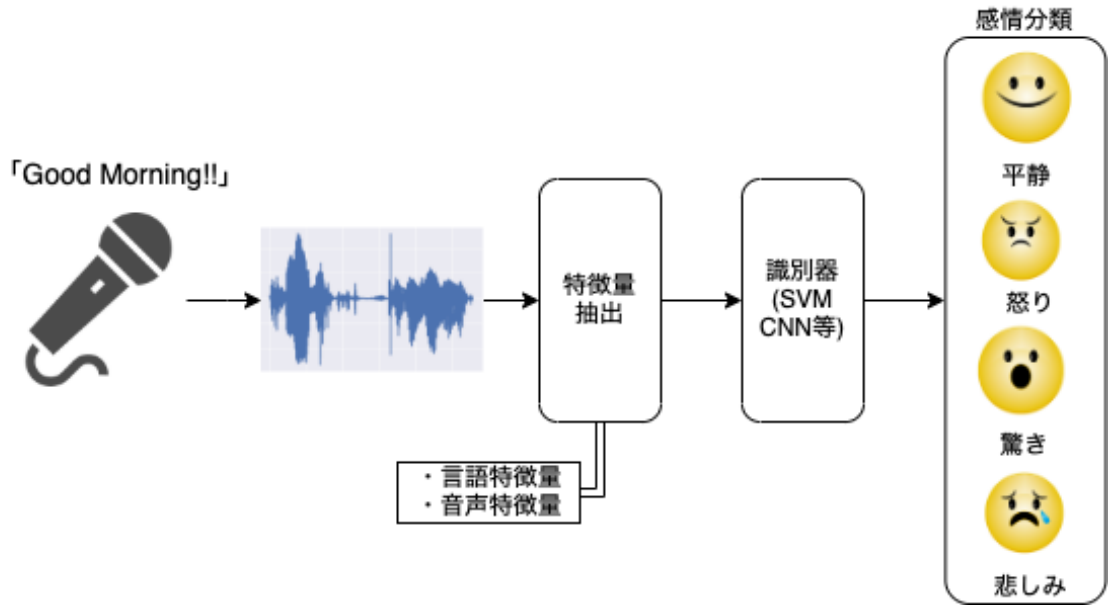

図 2.1:音声感情認識の流れ

2.2.1　入力データ

音声感情認識の分野には入力データとして，収録した音声，もしくは音声と収録した際の顔映像の 2 種類が存在する．本研究では，音声のみを用いた音声感情認識のモデルを構築するため，音声を入力データとする．収録環境として，先述した演技発話の収録か，自然発話の収録か 2 種類の環境設定が存在する．入力データのまとめを表 2.1 に示す．

表 2.1:音声感情認識の入力データまとめ

入力データ種別	入力データ収録環境
1.　音声	1.　演技発話収録
2.　音声＋収録時の顔の映像	2.　自然発話収録

2.2.2　音声の特徴量抽出

特徴量抽出には大別して，2 種類の特徴量が存在する．一つは言語特徴量，もう一つは音声特徴量である．言語特徴量は，「Good」「Morning」といった単語・文脈などが挙げられる．音声特徴量は，音圧・周波数・メル周波数ケプストラム係数（MFCC）などが挙げられる．音声特徴量の利点は，いかなる言語においても共通点が多いことである．それに対して言語特徴量は言語ごとに特徴量が全く異なる点，単語間の関係を読み取ることが難しい点が挙げられる．したがって多くの音声感情認識には，音声特徴量が用いられる．

入力データである音声をデジタル化し，音声特徴量抽出を行う過程を詳述する．以下に出てくるメル尺度とは，音高の知覚的尺度である．メル尺度の差が同じであれば，人間が感じる音高の差が同じになることを意図している．図 2.2 で示したように，人間の知覚する周波数と，音声の物理的な周波数が異なることから導入されている単位である．F_{mels} がメル尺度の周波数を，F_{Hz}は音声の物理的な周波数を示している．物理的な周波数が高くなるほど，人間は周波数の知覚がしにくくなることを示している．

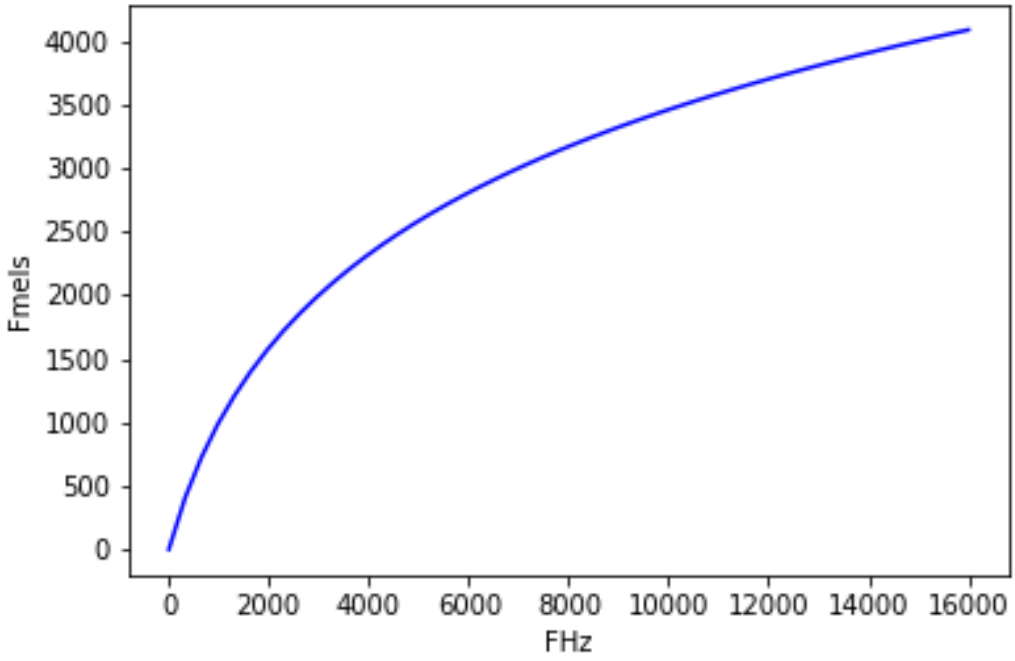

図 2.2:メル尺度と物理的な周波数の関係

まず入力データである音声のデジタル化について詳述する．音声波形はそもそも空気の振動であり，それがマイクロホンにより電圧の変化として観測る．つまりこれはアナログ（連続値）であり，音声を計算機で扱うにはデジタル化（離散化）する必要がある．これをアナログーデジタル（AD）変換と呼び，まずある特定の時間間隔によってデータを切り出す．これを標本化（サンプリング）と呼ぶ．人間の可聴域は20Hz～20,000Hzであり，一般にこの周波数の 2 倍の周波数で標本化することにより，元の音声波形を復元できると言われている.そのためサンプリング周波数は 44.1kHz のものが多く,本実験においても 48kHz で収録された音声を使用する．さらに切り出したデータに対し，量子化を行いデジタルデータへと変換される．この際音声認識の場合は 16bit で十分と言われていることから，本実験においても 16bit で量子化している．

次に，音声特徴量について述べる．音声特徴量には，音圧の最大・最小や標準偏差といった統計量と統計量をもとに算出される特徴量がある．統計量とは量子化した音声波形において，振幅の最大値・最小値，最大値・最小値の位置，振幅の平均，標準偏差などを算出する．一方の特徴量は，この量子化した音声波形をもとにいくつかの過程を経て算出される人間が作り出した尺度である．具体的には，本実験で使用する特徴量である MFCC（メル周波数ケプストラム係数）や RMSenergy などが上げられる．具体例として，本節では音声感情認識において多く利用されている MFCC について詳述する．MFCC を導出するために，まずこの量子化した音声波形のある特定の時間内の周波数成分を算出するためにフーリエ

変換を行う．特定の時間内の周波数成分の算出を，音声データを網羅するように実施することを短時間フーリエ変換（short-time Fourier transform: STFT）と呼ぶ．短時間フーリエ変換結果は，特定の時間内の周波数成分とその振幅を示す．短時間フーリエ変換結果の周波数成分と，振幅を対数スケールに直したものを対数パワースペクトルと呼ぶ．対数パワースペクトルにメルフィルタバンクと呼ばれる，メル周波数領域において特定の周波数を取り出すフィルタをかけることにより，対数パワースペクトルの周波数成分をメル尺度へ変換する．メルフィルタバンクにはフィルタ数を設定することができ，12～128 が一般的である．変換後の出力をさらに逆フーリエ変換（周波数成分軸から時間軸に戻す）を施したものをメルスペクトラムと呼ぶ．このメルスペクトラムに離散コサイン変換を施したものをメル周波数ケプストラム係数（MFCC）と呼ぶ．導出の過程の概略を図 2.3 に示す．

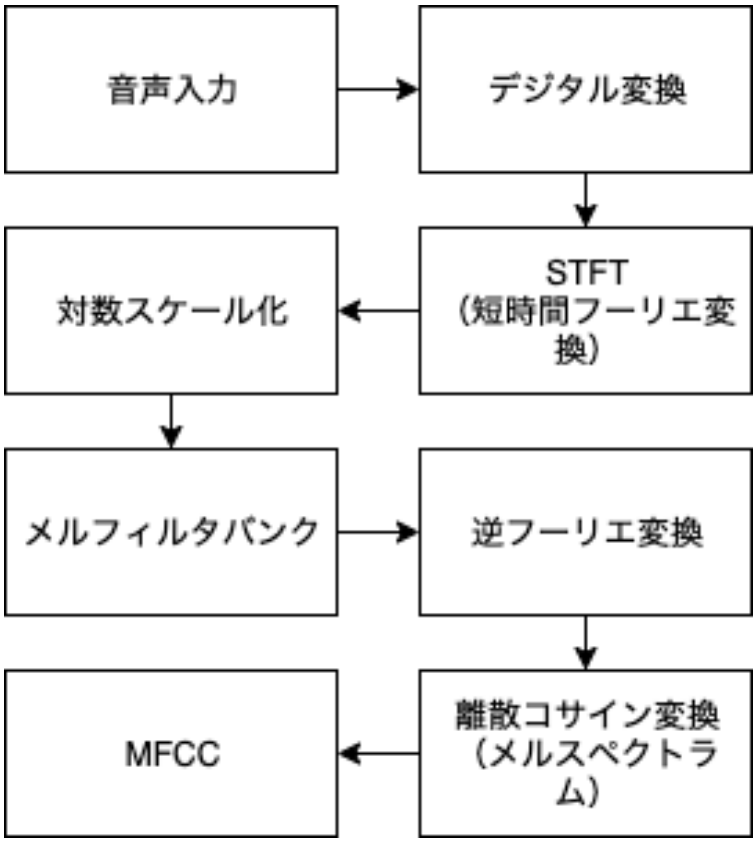

図 2.3: MFCC 導出の過程

最後に，近年の音声からの特徴量抽出の動向を述べる．音声からの特徴量抽出はヒューリスティックな手法から深層学習を用いた手法となった．実際に，Mohamed ら[3]は音声の対数パワースペクトルにメルフィルタバンクをかけた出力を入力特徴量とした DNN（Deep Neural Network）を提案し，MFCC を用いた場合より音声認識性能が高まることを示した．これは MFCC が音声の対数パワースペクトルのメルフィルタバンク出力を，加えて離散コ

サイン変換したものであり，音声の時間的特徴が失われてしまうためである．さらに，Sainath ら[4]は DNN の下層を CNN（Convolution Neural Network）に置き換えることにより音声認識性能が上がることを示した．これは CNN が波形に対する畳み込み演算がフーリエ変換に変わる周波数解析をしているためである．最新の音声感情認識では，順方向と逆方向の時間的特性を考慮できる BiLSTM（Bidirectional Long Short-Term Memory）を用いた学習や，CNN と LSTM を組み合わせたモデルにより音声特徴量の抽出をする．以上の近年の音声特徴量抽出のまとめを表 2.2 に示す．

表 2.2:音声感情認識における音声の特徴量抽出まとめ

特徴量抽出
1. 言語特徴量
(ア) word2vec
2. 音声特徴量
(ア) 統計量
① 振幅の最大値・最小値
② 最大値・最小値の位置
③ 振幅の平均
④ 振幅の標準偏差
(イ) 特徴量
① 対数パワースペクトル
② MFCC

2.2.3 識別器

2.2.2 節で述べたように音声感情認識における識別器として，様々な機械学習のモデルが用いられている. 従来は HMM（Hidden Markov Model）, GMM（Gaussian Mixture Model）, SVM（Support Vector Machine），RF（Random Forest）等の機械学習が，近年では CNN や LSTM により，特徴量抽出と識別器の 2 つの役割を同時に担うモデルも多く存在している．表 2.3 に識別器の例を示す.

表 2.3: 音声感情認識に用いられる識別器

識別器
1. HMM（Hidden Markov Model）
2. GMM（Gaussian Mixture Model）
3. SVM（Support Vector Machine）
4. RF（Random Forest）
5. CNN
6. LSTM

2.2.4 感情分類

感情分類は，5〜7 感情に分類，もしくは特定の次元によって算出され数値化される．5〜7 感情は，心理学において著名である Ekman ら[5]の基本 6 感情（怒り，嫌悪，恐れ，喜び，悲しみ，驚き），Pluchik ら[6]の 8 感情（怒り，嫌悪，悲しみ，驚き，恐れ，容認，喜び，期待）をもとに抽出される．数値化は Russell ら[7]の回帰分類を用いてなされている．これは感情を Valence（感情価）と Arousal（覚醒度）の 2 軸により数値化する手法である．音声データには感情のアノテーションが第三者より施されている．表 2.4 に感情分類を示す.

表 2.4: 音声感情認識における感情分類まとめ

感情分類
1. カテゴリー分類 (ア) 基本 6 感情（怒り，嫌悪，恐れ，喜び，悲しみ，驚き）＋「平静」のうち 5 ～ 7 感情 (イ) Pluchik ら[6]の 8 感情（怒り，嫌悪，悲しみ，驚き，恐れ，容認，喜び，期待）＋「平静」のうち 5 ～ 7 感情 2. 感情の数値化 (ア) Valence（感情価）と Arousal（覚醒度）による数値化

2.2.5 音声感情認識のまとめ

2.2.1 節から 2.2.4 節で述べた音声感情認識の研究を入力データ，特徴量抽出，識別器，感情分類をそれぞれ表 2.5 に示す.

表 2.5: 音声感情認識のまとめ

入力データ	特徴量抽出	識別器	感情分類
1. 音声 2. 音声＋収録時の顔の映像	1. 言語特徴量 2. 音声特徴量 (ア) 統計量 (イ) 特徴量	1. HMM 2. GMM 3. SVM 4. RF 5. CNN 6. LSTM	1. カテゴリー分類 (ア) 怒り，嫌悪，恐れ，喜び，悲しみ，驚き＋「平静」のうち 5 ～ 7 感情 (イ) 怒り，嫌悪，悲しみ，驚き，恐れ，容認，喜び，期待＋「平静」のうち 5 ～ 7 感情 2. 感情の数値化 (ア) Valence，Arousal

2.3 関連研究

本節では，日本人の自然発話の音声感情認識に関する研究を述べる．当該分野の最新の研究を表 2.6 に示す．本研究で用いるデータセット OGVC[8]を用いた自然発話に関する最新研究に阿部ら[9]があるため，表 2.6 中に示した．また，表 2.7 に表 2.6 に示した最新研究の対象音声・入力データ・特徴量抽出・識別器・感情分類の 5 つを示す．表 2.7 に示した研究のうち，自然発話を対象とした，教師あり学習を用いた音声感情認識の提案手法と実験結果を 2.3.1 節～2.3.6 節にて詳述する．本研究は，自然発話を対象とした教師あり学習であるためである．

表 2.6:音声感情認識の最新研究

提案者	提案年	対象音声	新規性
阿部ら[9]	2016	自然発話	自然発話音声の感情推定に，演技音声を訓練として使用するモデルを提案した.
Deng ら[10]	2018	自然発話	半教師あり自己符号化器を提案した．教師データとして自然発話の感情ラベル付きデータを，教師なしデータとして演技発話の音声を入力とした.
Song ら[11]	2019	演技発話&自然発話	TLSL （ Transfer Linear Subspace Learning）を用いて複数の音声感情コーパス間の相関関係を示した.
Xie ら[12]	2019	演技発話	Attention 機構付き LSTM による音声感情認識モデルを提案した. 入力として 128 次元の音声特徴量，2 層の LSTM 層をもったネットワークを実装した.
Kim ら[13]	2019	演技発話	入力として, 感情音声と映像を用いる. 顔上部，顔下部，音声の 3 つをそれぞれ識別器入力し, 3 つの出力を合わせて感情を推測する．それぞれ 3 つのデータを細かく分割した際に，その分割データに独自

			ラベルをつける手法を提案した.
Guo ら[14]	2019	演技発話	入力特徴量として，人間の経験則により作成した特徴量と，スペクトラム特徴量を用いる手法を提案した．スペクトラム特徴量により，人間が無意識に感じ取っている特徴量を学習することを狙いとした.
Zhao ら[15]	2019	演技発話&自然発話	スペクトラム特徴量を, Attention 機構付き Bidirectional LSTM と CNN の 2 つの識別器に入力し学習するモデルを提案した. 2 つの学習器によりより深い特徴量の発見を狙いとした.
Lotfian ら[16]	2019	自然発話	カリキュラム学習を用いた音声感情認識モデルを提案した．複数のアノテーターが同じ感情ラベルを施した音声から学習することで, 感情認識精度向上を狙った.
Meng ら[17]	2019	演技音声	より重要な特徴量を，dilated CNN と Bidirectional LSTM により発見した．さらに LSTM には Attention 機構を実装, 損失関数に softmax と center loss を用いた.
Hossain ら[18]	2019	演技発話	音声と映像を入力とするディープラーニングを用いた音声感情認識手法を提案した．音声と映像を CNN で学習させている.
Badshah ら[19]	2019	演技発話	CNN の学習において長方形の窓を用いるモデルを提案した．入力特徴量をメルスペクトラムとした．時間的特性をより CNN で捉えることを狙った.

Shahin ら [20]	2019	演技音声	学習器を混合ガウスモデルとニューラルネットワークの組み合わせを提案した．これによりアラビア語の音声感情認識モデル構築を行った．
Ocquaye ら [21]	2019	自然発話	DEAT（dual exclusive attentive transfer）という独自の自己符号化器を提案した．半教師あり学習にて音声感情認識モデルを構築している．
Zhang ら[22]	2019	自然発話	音声感情認識において最も重要な特徴量を抽出できる，時間軸の分割幅を発見した．入力特徴量としてメルスペクトラムを用いて，メルスペクトラムの時間軸方向の分割幅を複数実験した．
Jiang ら[23]	2019	演技音声	音声のスペクトラム特徴量から，CNN と LSTM によって有用な特徴量を抽出する手法を提案した．

それぞれの最新研究を 2.2.5 節にしたがって，対象音声・入力データ・特徴量抽出・識別器・感情分類を表 2.7 に示す．

表 2.7: 最新研究の実装まとめ

提案者	対象音声	入力データ	特徴量抽出	識別器	感情分類
阿部ら [9]	自然発話	音声	44 次元の統計量・特徴量（MFCC 含む）	SVM	5 感情
Deng ら [10]	自然発話	音声	384 次元の統計量・特徴量と自己符号化器による特徴量抽出	NN（Neural Network）	4～7 感情

Song ら [11]	演技発話&自然発話	音声	1582 次元の統計量・特徴量	Transfer Supervised Linear Subspace Learning（次元削減）	5 感情
Xie ら [12]	演技発話	音声	6373 次元の統計量・特徴量	Attention 機構付き LSTM	6 感情
Kim ら [13]	演技発話	音声＋収録時の顔の映像	顔上部＋顔下部＋音声特徴量	SVM，DTW（Dynamic Time Warping）	4 感情
Guo ら [14]	演技発話	音声	384 次元の統計量・特徴量＋メルスペクトラムから CNN により抽出	BiLSTM	4，7 感情
Zhao ら [15]	演技発話&自然発話	音声	メルスペクトラムから LSTM，CNN により抽出	CNN（AlexNet）＋Attention 機構付き LSTM によるアンサンブル学習	5 感情
Lotfian ら[16]	自然発話	音声	6373 次元の統計量・特徴量	NN	Valence, Arousal により数値化
Meng ら [17]	演技音声	音声	3 次元メルスペクトラムから CNN により抽出	CNN＋BiLSTM	4 感情
Hossain ら[18]	演技発話	音声＋収録時の顔映像	メルスペクトラムから CNN により抽出	SVM	6 感情
Badshah	演技発話	音声	メルスペクトラムか	複数の CNN	7 感情

ら[19]			ら CNN により抽出	（AlexNet）でのアンサンブル学習	
Shahin ら[20]	演技音声	音声	MFCC	GMM ＋DNN	6 感情
Ocquaye ら[21]	自然発話	音声	13 次元メルスペクトラムを入力とし自己符号化器により抽出	SVM CNN PCA	Valence, Arousal により数値化
Zhang ら[22]	自然発話	音声	3 次元メルスペクトラムから CNN（AlexNet）により抽出	LSTM	6〜7 感情
Jiang ら[23]	演技音声	音声	3 次元メルスペクトラムより CNN と LSTM により抽出	CNN（AlexNet） ＋LSTM	6〜７感情

2.3.1 阿部ら[9]の研究

阿部ら[9]は，OGVC（Online Gaming Voice Chat corpus with emotional label）[8]を用いて，日本人の自然発話を対象としたSVM（Support Vector Machine）による感情認識の手法を提案した．OGVC[8]とは，オンラインゲームをプレイしている際の男性 9 名，女性 4 名の音声チャットを自然発話として収録し，3 名の第三者により感情のアノテーションをされているデータセットである．加えて，男女 2 名ずつのプロの俳優に，各発話をアノテーションがつけられた感情で演じた，演技発話も収録している．演技発話は感情強度を 1〜4 として，感情を演じる度合いを設定し収録している．

阿部ら[9]は特徴量抽出により MFCC（1〜12 次）や RMSenergy など，加えてそれら特徴量の変化量を含む 44 特徴量を用いた．学習には怒り・喜び・悲しみ・驚き・平静の 5 感情の音声感情を選び，計 2438 の発話を SVM で学習させた．提案手法概要を表 2.8 に，実験結果を表 2.9 に阿部ら[9]より引用し記載する．各感情の平均 Accuracy は 38.6%であった．

表 2.8: 阿部ら[9]の提案手法概要

対象音声	入力データ	特徴量抽出	識別器	感情分類
自然発話	音声	統計量・特徴量 （MFCC 含む）	SVM	5 感情

表 2.9: SVM による 5 感情分類[%]（阿部ら[9]より抜粋）（認識率の標準偏差＝21.39）

		推測された感情					発話数
		怒り	喜び	悲しみ	驚き	平静	
正解ラベル	怒り	**7.2**	27.4	2.5	19.8	43.0	240
	喜び	5.9	**44.9**	2.2	11.9	35.1	595
	悲しみ	1.2	18.5	**21.8**	11.5	46.9	243
	驚き	5.7	16.8	3.2	**52.0**	22.3	565
	平静	3.0	17.5	4.1	8.4	**66.9**	798
	平均	38.6					

また，OGVC の演技発話の感情強度 1，2 のデータも学習に用いることで Accuracy が 40.9%に上昇したことを報告している．原因を感情強度 1，２の演技発話は自然発話と大き

く異ならないため，学習データの不足であると報告した．特に「怒り」「悲しみ」の発話数が少ないラベルの Accuracy が低いことから見て取れる．阿部ら[9]のモデルの認識率の標準偏差と，各感情の Precision（適合率），Recall（再現率），F 値を計算したところ表 2.10 に示す結果となった．

表 2.10: 阿部ら[9]のモデルの統計量（認識率の標準偏差＝21.4）

	Precision	Recall	F 値
怒り	15.5	7.20	9.83
喜び	43.6	44.9	44.3
悲しみ	43.1	21.8	29.0
驚き	57.9	52.0	54.8
平静	49.2	66.9	56.7
平均	41.9	38.6	38.9

2.3.2 Song ら[11]の研究

Song ら[11]の提案手法概要を表 2.11 に示す．訓練・評価用に，ドイツ語の演技発話コーパスである Berlin コーパス（Emo-DB）[1]，英語の演技音声である eNTERFACE コーパス[2]，ドイツ語の自然発話である FAU Aibo データセット[3]の 3 つを用いている．自然発話データセットである FAU Aibo データセットの結果を表 2.12，表 2.13 に示す．

表 2.11: Song ら[11]の提案手法概要

対象音声	入力データ	特徴量抽出	識別器	感情分類
演技発話&自然発話	音声	1582 次元の統計量・特徴量	Transfer Supervised Linear Subspace Learning（次元削減）	5 感情

表 2.12: Song ら[11]の提案手法による実験結果（単位[%]）

		推測された感情				
		怒り	嫌悪	恐れ	喜び	悲しみ
正解ラベル	怒り	**36**	30	25	8.0	1.0
	嫌悪	16	**75**	0.0	7.0	2.0
	恐れ	6.0	51	**19**	14	10
	喜び	4.0	48	17	**26**	5.0
	悲しみ	5.0	1.0	19	3.0	**72**
	平均	45.6				

[1] http://emodb.bilderbar.info/docu/
[2] http://enterface.net/enterface05/main.php?frame=emotion
[3] https://www5.cs.fau.de/nc/our-team/

表 2.13: Song ら[11]のモデルの統計量

	Precision	Recall	F 値
怒り	53.7	36	43.1
嫌悪	36.6	75.	49.2
恐れ	23.8	19	21.1
喜び	44.8	26	32.9
悲しみ	80	72	75.8
平均	47.7	45.6	44.4

2.3.3 Zhao ら[15]の研究

Zhao ら[15]の提案手法の概要を表 2.14 に示す．訓練・評価用に，FAU Aibo を用いている．5 感情分類（怒り，驚き，平静，喜び，悲しみ）において，平均 Accuracy が 45.4%であったと報告している．各感情の詳細の精度については記載がなかった．

表 2.14: Zhao ら[15]の提案手法概要

対象音声	入力データ	特徴量抽出	識別器	感情分類
演技発話&自然発話	音声	メルスペクトラムから LSTM，CNN により抽出	CNN（AlexNet）＋Attention 機構付き LSTM によるアンサンブル学習	5 感情

2.3.4 Zhang ら[22]の研究

Zhang ら[22]は，使用しているデータセットが異なるため厳密な比較はできないが，最新手法として詳述する．Zhang ら[22]は，AFEW5.0[24]と BAUM-1s[25]を用いて自然発話を対象とした音声感情認識の新たな手法を提案した．AFEW5.0[24]は感情のアノテーション付きの映像データ，BAUM-1s[25]は３１人のトルコ人による自然対話を収録した感情アノテーション付きの映像データである．AFEW5.0[24]は，約 1600 発話に 3 人の第三者が 7 種類の感情アノテーションを施し，BAUM-1s[25]は，約 500 発話に第三者が 8 種類の感情アノテーションを施したものである．提案手法は，入力を発話音声の対数パワースペクトルにメルフィルタバンクをかけて出力したメルスペクトラム，メルスペクトラムの変化量，メルスペクトラムの変化量の変化量の 3 次元メルスペクトラムする．その入力を CNN と LSTM によって学習させる．提案手法の概要を表 2.15 に示す．

表 2.15: Zhang ら[22]の提案手法概要

対象音声	入力データ	特徴量抽出	識別器	感情分類
自然発話	音声	3 次元メルスペクトラムから CNN（AlexNet）により抽出	LSTM	6～7 感情

AFEW5.0[24]の実験結果を図 2.5 に，BAUM の実験結果を図 2.6 に Zhang ら[22]より引用し記載する．各感情の最大の平均 Accuracy は，AFEW5.0[24]において 40.73%，BAUM-1s[25]において 50.22%であった．各感情の Precision，Recall，F 値を算出し，表 2.16，表 2.17 に記載する．

図 2.4: AFEW5.0[24]における結果（認識率の標準偏差=18.93）（Zhang ら[22]より引用．数値の単位[%]）

	怒り	嫌悪	恐れ	喜び	悲しみ	驚き	平静
怒り	65.6	0.0	1.6	15.6	7.8	1.6	7.8
嫌悪	12.5	20.0	5.0	32.5	5.0	0.0	25.0
恐れ	10.9	4.4	52.2	13.0	6.5	4.4	8.7
喜び	11.1	0.0	12.7	49.2	9.5	0.0	17.5
悲しみ	4.9	3.3	13.1	16.4	34.4	6.6	21.3
驚き	17.4	0.0	2.2	43.5	4.4	10.9	21.7
平静	6.4	3.2	1.6	25.4	3.2	1.6	58.7
平均	41.6						

表 2.16: AFEW5.0[24]における最大 Accuracy（40.73%）時の統計量（Zhang ら[22]より抜粋．数値の単位[%]）

感情	Precision	Recall	F-score
怒り	50.97	65.63	57.38
嫌悪	64.94	20.00	30.58
恐れ	59.08	52.17	55.41
喜び	25.15	49.21	33.29
悲しみ	48.63	34.43	40.32
驚き	43.60	10.87	17.40
平静	36.53	58.73	45.05
平均	46.98	41.58	39.92

図 2.5: BAUM-1s[25]における結果（認識率の標準偏差＝25.81）（Zhang ら[22]より抜粋．数値の単位[%]）

	怒り	嫌悪	恐れ	喜び	悲しみ	驚き
怒り	34.88	11.63	2.33	0.00	41.86	9.30
嫌悪	13.54	23.96	5.21	37.50	17.71	2.08
恐れ	7.14	32.14	3.57	17.86	25.00	14.29
喜び	1.95	11.69	0.65	79.87	4.55	1.30
悲しみ	13.04	13.04	13.06	4.34	49.07	6.83
驚き	5.13	10.26	17.95	5.13	15.38	46.15
平均	39.58					

表 2.17: BAUM-1s[25]における統計量（Zhang ら[22]より抜粋）

感情	Precision	Recall	F-score
怒り	46.09	34.88	39.71
嫌悪	23.32	23.96	23.64
恐れ	8.23	3.57	4.98
喜び	55.19	79.86	65.27
悲しみ	31.95	49.07	38.71
驚き	57.72	46.15	51.29
平均	46.99	41.58	39.92

2.4 問題点

自然発話を対象とした音声感情認識に関する関連研究として，2.3.1 節にて阿部ら[9]の研究を，2.3.2 節にて Song ら[11]の研究を，2.3.3 節にて Zhao ら[15]の研究を，2.3.4 節にて Zhang ら[22]の研究を述べた．以上 4 つの研究における問題点は，認識率の標準偏差が大きいことである．音声感情認識において，推測できる感情に偏りが発生していることを意味する．ある感情は正しく認識できるが他の感情は正しく認識できないモデル（怒りだけ認識できない，悲しみだけ認識できない）となることを意味するため適切でない．本来であれば，Accuracy が高く，分散が小さいモデルが音声感情認識としては適切である．表 2.18 に各最新研究の Accuracy と，感情認識率標準偏差を示す．Zhao ら[15]については，各感情の精度が論文中に記載されていなかったため算出できなかった．表 2.18 よりすべての標準偏差が 20 程度になっていることがわかる．したがって本研究では，精度を落とさず標準偏差を小さくすることを目標とする．

表 2.18: 最新研究の Accuracy と標準偏差

提案者	分類感情数	訓練と評価に用いた発話数	Accuracy[%]	標準偏差
阿部ら[9]	5 感情	2,438	38.6	21.4
Song ら[11]	5 感情	18,216	45.6	23.4
Zhao ら[15]	5 感情	18,216	45.4	–
Zhang ら[22]	7 感情	1,426	41.6	18.9

第3章　　提案手法

3.1 概要

本研究では，日本人の自然対話における音声感情認識の精度向上の一助となることを目的としている．そこで，訓練に用いる感情音声のサンプル数の最適化を提案する．Accuracy を高め，感情認識標準偏差（ばらつき）を小さくする．仮説として，人間も徐々に他人の感情を理解できるようになること，他人の感情には分かりづらい感情（悲しい，恐れ等）もあることを考えた．徐々に理解できるようになるためアップサンプリングを行う．また，分かりづらい感情の認識率を高めるためにサンプル数の比の最適化を行う．具体的には，怒り，喜び，悲しみ，驚き，平静で表される 5 感情の学習に最適な比率を提案する．また，この際の学習には CNN＋BiLSTM を用いる．入力から出力の流れを図 3.1 に示す．さらに，阿部ら[9]，Song ら[11]，Zhao ら[15]，Zhang ら[22]との比較のため提案手法の概要を表 3.1 に示す．特徴量抽出は Meng ら[17]の手法と同様に，識別器は Zhang ら[22]と CNN のみ異なるモデルを構築した．3.2 節にて 3 次元メルスペクトラムの抽出手法を，3.3 節にて識別器を，3.4 節にて音声感情サンプル数のアップサンプリング手法を詳説する．

表 3.1: 提案手法と最新研究の音声感情認識モデル

提案者	特徴量抽出	識別器	感情分類
阿部ら[9]	44 次元の統計量・特徴量（MFCC 含む）	SVM	5 感情（怒り，喜び，悲しみ，驚き，平静）
Song ら[11]	1582 次元の統計量・特徴量	Transfer Supervised Linear Subspace Learning（次元削減）	5 感情（怒り，喜び，悲しみ，驚き，平静）
Zhao ら[15]	メルスペクトラムを BiLSTM(2 層・128 出力), CNN（AlexNet, VGG16, VGG19）に入力	CNN（AlexNet）＋Attention 機構付き LSTM によるアンサンブル学習	5 感情（怒り，喜び，悲しみ，驚き，平静）
Zhang ら[22]	3 次元メルスペクトラムから CNN（AlexNet）により抽出	LSTM（3 層 256 出力 or3 層 512 出力）	7 感情（怒り，嫌悪，恐れ，喜び，悲しみ，驚き，平静）
提案手法	3 次元メルスペクトラムから CNN（4 入力，3 層）により抽出	BiLSTM（1〜3 層 128 出力〜512 出力）	5 感情（怒り，喜び，悲しみ，驚き，平静）

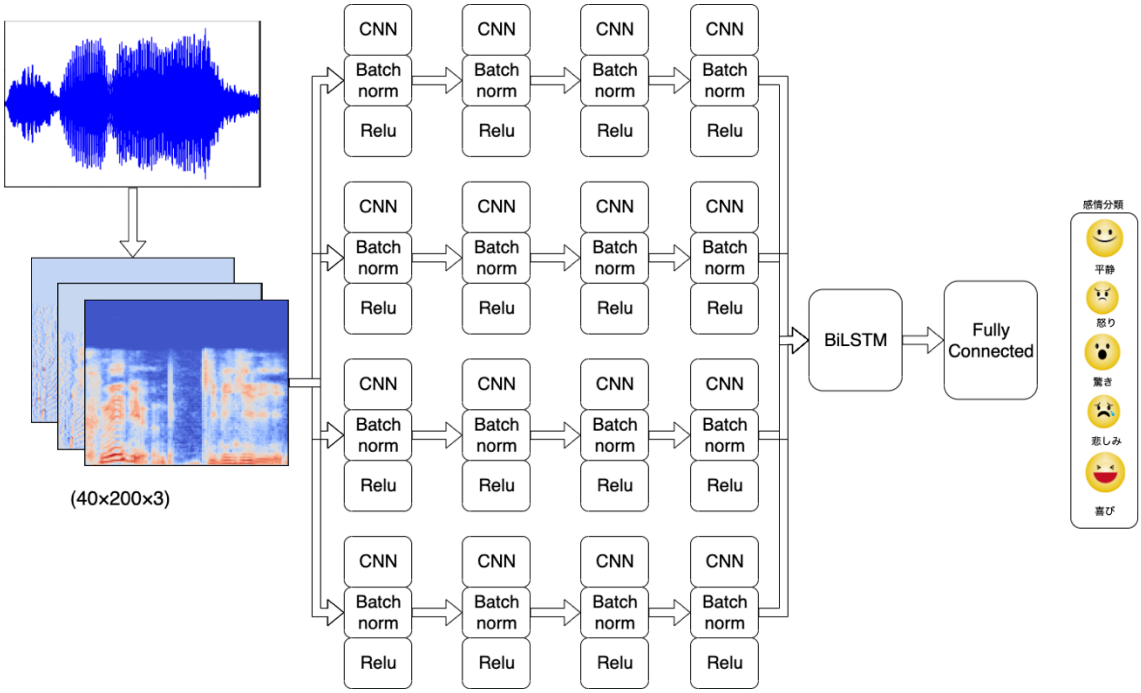

図 3.1: 提案手法の入力から出力の流れ（入力を 3 次元メルスペクトラムとし，出力を怒り，喜び，悲しみ，驚き，平静の 5 感情．学習器には CNN と LSTM を組み合わせる．）

3.2 3 次元メルスペクトラム

本節では, CNN への入力である 3 次元メルスペクトラムについて詳説する. Meng ら[17]の手法と同様の手順で，3 次元メルスペクトラムの特徴量抽出を行う．1 次元メルスペクトラムとは 2.2.2 節，図 2.3 にて示した MFCC を導出する際の，離散コサイン変換を施す前の特徴量を言う．図 3.2 に示すように STFT（short-time Fourier transform）を施す．この出力を対数スケールにし（対数パワースペクトルに変換し），メルフィルタバンクを施し逆フーリエ変換することにより 1 次元メルスペクトラムを算出する. 1 次元メルスペクトラムと 1 次元メルスペクトラムの変化量, 1 次元メルスペクトラムの変化量の変化量を合わせて 3 次元メルスペクトラムと呼ぶ．この 3 次元メルスペクトラムから CNN により特徴量抽出を行う.

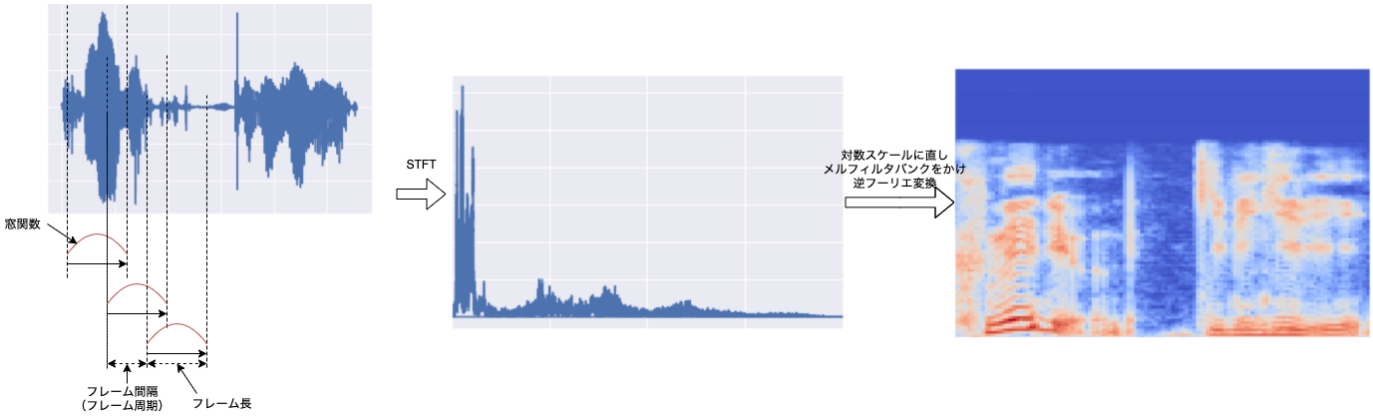

図 3.2: メルスペクトラムの導出過程

窓関数とは，STFT（short-time Fourier transform）を行う際に特定の時間内の周波数を切りだすために用いる関数である．ハイパーパラメータである窓関数，フレーム長，フレーム間隔，メルフィルタバンク数を Meng ら[17]の手法と同様に，窓関数をハミング窓，フレーム長を 25ms，フレーム間隔を 10ms，メルフィルタバンク数は 40 とした.

3.3 CNN と LSTM の組み合わせ学習

本節では，学習器に用いる CNN（Convolution Neural Network）と LSTM（Long Short-term Memory）の組み合わせについて詳説する．本実験では Zhang ら[22]の手法と CNN のみ異なる，同様の識別器を用いた．Zhang ら[22]は CNN に AlexNet[26]を用いている．本実験では，3 次元メルスペクトラムがもともとは画像ではなく音声であることから，3 次元メルスペクトラムの時間軸方向の特徴量をより抽出できるような CNN を独自に構築した．3.3.1 節にて CNN について，3.3.2 節にて LSTM について述べる．

3.3.1 CNN（Convolution Neural Network）

CNN は畳み込みニューラルネットワークと呼ばれる MLP（多層パーセプトロン）の一種であり，LeCun ら[27]によって提案された．層の間のユニットの結合が全結合でなく，結合重みが複数エッジで共有されているニューラルネットワークである．CNN は以下の式で表される単層パーセプトロンの組み合わせで構成される．

$$u_{ij} = \sum_{p=0}^{P-1} \sum_{q=0}^{Q-1} w_{pq} x_{i+p,j+q} + b \tag{1}$$

$$y_{ij} = f(u_{ij}) \tag{2}$$

ここで，図 3.3 のように縦方向にA画素，横方向にB画素からなる 2 次元の画像において，位置(i,j)にある画素の画素値を$x_{i,j}$とする．P,Qを画像中の任意の画素数（$\mathrm{P} \leq \mathrm{A}, \mathrm{Q} \leq \mathrm{B}$）と考え，縦方向にP，横方向に$Q$のサイズを持った矩形領域をフィルタと呼ぶ．$f(x)$を任意の活性化関数とし，$y_{ij}$をパーセプトロンの出力とする．CNN は$A \times B$画素の画像内において，このフィルタを一定の画素数$d$の幅だけ縦方向，あるいは横方向，もしくは縦方向横方向ともにずらしながら出力y_{ij}を求める．dをストライドと呼ぶ．wを各ノードが持つ重みとし，位置任意の位置(i,j)におけるフィルタの重みをw_{ij}とする．bはフィルタのバイアスである．

本実験では，サイズが$(8, 16, 32, 64)$となる 4 つのサイズの窓を用いて，ストライド幅を 1 または 2 とする CNN を使用した．さらに図 3.1 のように，入力である 3 次元メルスペクトラムを 4 つの窓サイズが異なる CNN に入力し，その出力を BiLSTM（Bidirectional Long short-term Memory）の入力としている．

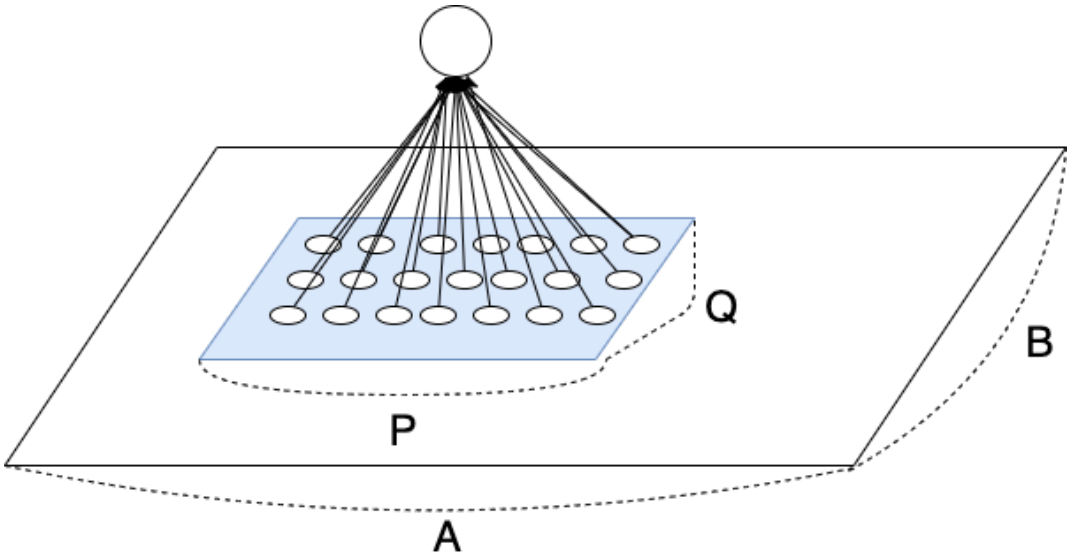

図 3.3: CNN（Convolution Neural Network）のフィルタの例．$P = 7, Q = 3$の場合．

3.3.2 BiLSTM（Bidirectional Long Short-term Memory）

BiLSTM は RNN（Recurrent Neural Network：再帰型ニューラルネットワーク）の一種である. RNN とは, あるユニット間に有向な閉路をもつニューラルネットワークである. RNN は過去の時刻における入力の影響が時間軸方向に指数的に減衰してしまう欠点があるため，長時間にわたる特徴量を学習できない問題点がある．LSTM はこの問題を解決するために，以下の 5 つの式により出力を演算する．LSTM はメモリユニットと呼ばれる要素を基本単位として, RNN の隠れ層のユニットに置き換えて用いている. メモリユニットは, 図 3.4 に示すように一つのメモリM，5 つのユニット(A, B, I, F, O)と 3 つのゲート（入力ゲート，忘却ゲート，出力ゲート）により構成される．入力を$(x_1, x_2, \dots, x_T)$とし出力を$(y_1, y_2, \dots, y_T)$とし，$t = 1$から$t = T$まで演算を行う.

$$i_t = \sigma(W_{xi}x_t + W_{hi}h_{t-1} + W_{ci}c_{t-1} + b_i) \tag{3}$$

$$f_t = \sigma\left(W_{xf}x_t + W_{hf}h_{t-1} + W_f c_{t-1} + b_f\right) \tag{4}$$

$$c_t = f_t c_{t-1} + i_t \tanh(W_{xc}x_t + W_{hc}h_{t-1} + b_c) \tag{5}$$

$$o_t = \sigma(W_{xo}x_t + W_{ho}h_{t-1} + W_{co}c_t + b_o) \tag{6}$$

$$y_t = \sigma_t \tanh(c_t) \tag{7}$$

（ただし，　$\sigma(x) = sigmoid(x) = \frac{1}{1+e^{-x}}$）

i_t, f_t, c_tはそれぞれ入力ゲート，忘却ゲート，メモリの計算を表し，o_tは出力ゲートの計

算を表す．$W_{\alpha\beta}$はα，β間の重みを示す（α，βはインプットゲート，出力ゲート，忘却ゲート，メモリ，入力のいずれかである）．

以上の LSTM は時間軸の順方向のみの影響を考慮したモデルであるため，本実験では時間軸の逆方向の影響も考慮するため，BiLSTM を用いる．BiLSTM は LSTM を時間軸に順方向と逆方向の 2 層に重ねたものである．

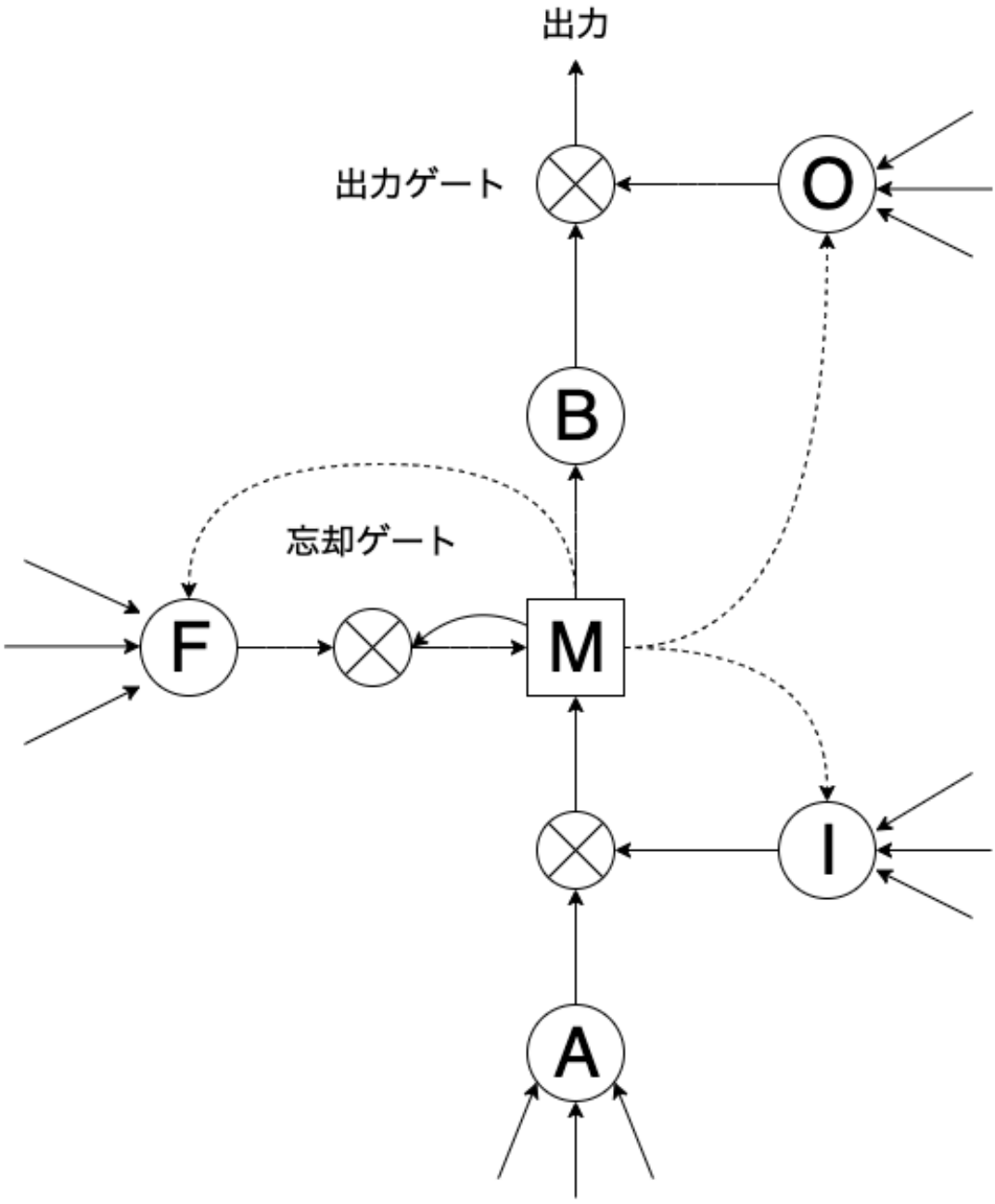

図 3.4: LSTM（Long Short-term Memory）

3.4 音声感情サンプルのアップサンプリング

本実験では，音声感情サンプルのアップサンプリングを特徴量の時間軸をシフトすることにより実施する．4.2 節にて詳述する，本実験で使用するデータセット OGVC[8]では，「1 発話」の基準を 400ms 以上のポーズによって挟まれた音声の範囲としている．たとえば「○もう疲れた○」（○は 400ms 以上のポーズを表す）という発話のとき「もう疲れた」を 1 発話として見なす．そのため「○もう疲れた，今日はもう休みたいな○」のような発話の場合，「もう疲れた，今日はもうもう休みたいな」を 1 発話とする．文章の意味ではなく，発話時間を基準に 1 発話を定義している．したがって，発話データにより発話の時間幅に差が存在している．

本実験では(40,200)を入力サイズとし，40 は 3.2 節で述べたメルフィルタバンク数，200 はスペクトラムの時間軸である．200 は発話時間に直すと，約 2 秒である．2 秒以上の発話時間となるものは，直感的に単一の感情でないと考えられるため，スペクトラムの時間軸が 201 以上となる発話は訓練・評価データからは除外した．Meng ら[17]，Zhang ら[22]はスペクトラムの時間軸の上限を 227 としており，228 以上となる発話は除外している．図 3.5 に示すようにある音声の 1 次元メルスペクトラムの特徴量が(40,80)のとき，図 3.3 のように(40,200)に変換するときに 0 パティング（足りない部分を 0 で埋めてサイズ調整する）を行う．これを 3 パターンで実施することでアップサンプリングしている．たとえば，「今日は疲れた」という音声が，「○○今日は疲れた」，「○今日は疲れた○」「今日は疲れた○○」（○は無声区間を示す.）のように 3 つのサンプルになる．この際，話し始める時間が異なるだけで，感情に差はない．アップサンプリング前の訓練・評価データには，0 パティングを前後に施したデータ（「○今日は疲れた○」となるデータ）を用いている．Meng ら[17]，Zhang ら[22]は，0 パティングを前後に施したデータ（「○今日は疲れた○」となるデータ）を訓練・評価データとして用いている．

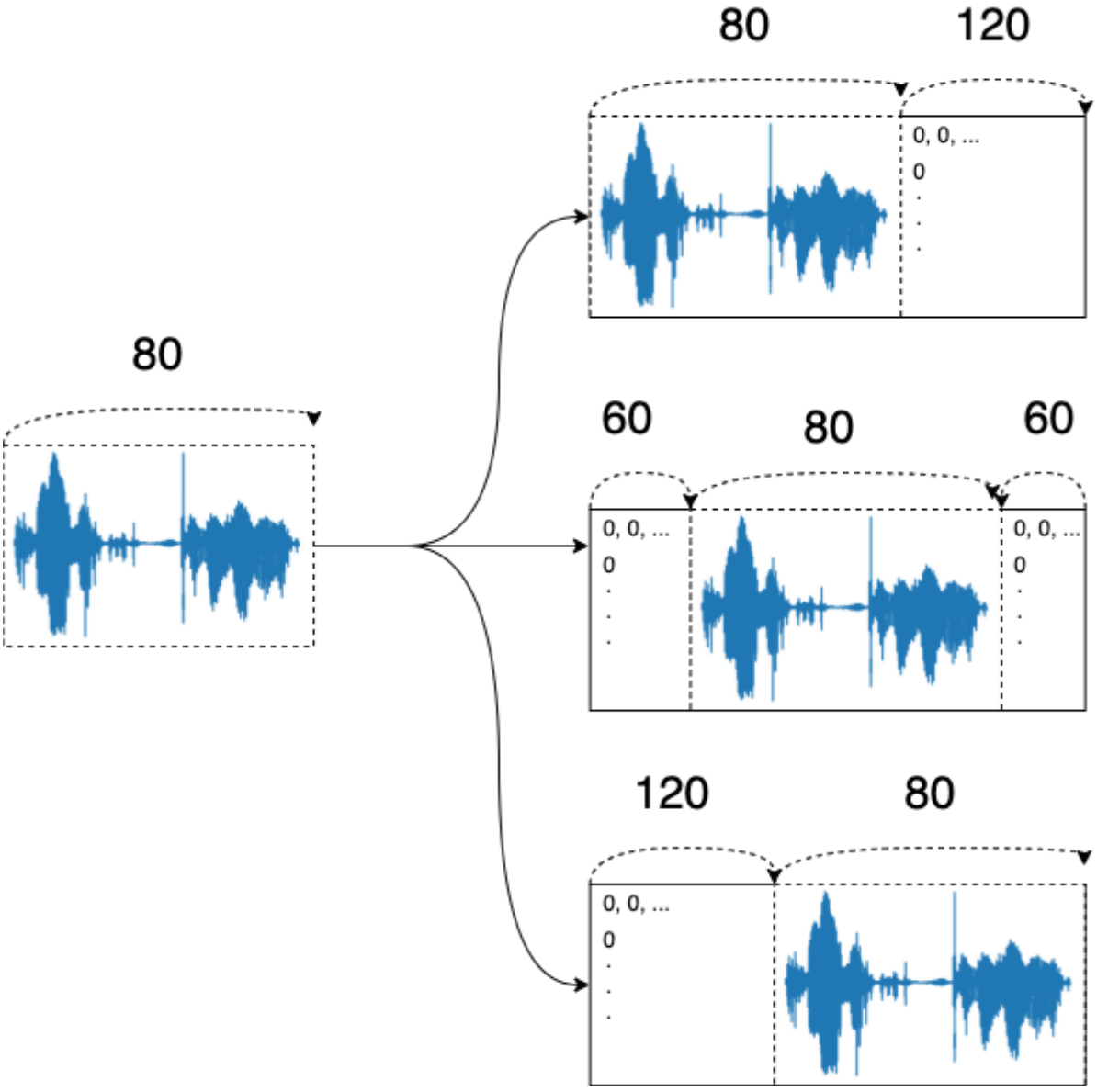

図 3.5: 音声の時間軸をずらすアップサンプリング手法

3.4.1 SMOTE（Synthetic Minority Over-sampling Technique）

4.4.1 節にて一般的なアップサンプリング手法である Chawla ら[28]の提案した SMOTE(Synthetic Minority Over-sampling Technique)を用いた手法との比較を実施する．本節では，SMOTE の概説をする．図 3.6 のようなクラス分類をしているデータセットを例として，SMOTE の手順を以下に示す．

手順1. サンプルの選定：少数派サンプルを一つ(m_aとする)ランダムに選択する．（図 3.7）

手順2. k-最近傍の特定：m_aのk-最近傍を特定する．kは Chawla ら[28]と同様に，$k = 5$とした．（図 3.8）

手順3. k-最近傍からランダムにサンプル選択：k-最近傍のサンプルの中からランダムにサンプル(m_rとする)を選ぶ．（図 3.9）

手順4. サンプルの追加：m_aとm_rを結ぶ線分上のランダムな位置に新しいサンプルを生成する．（図 3.10）

手順5. 手順 3 と手順 4 の繰り返し：手順 3 と手順 4 を（SMOTE によって追加する少数化サンプル数/元の少数派サンプル数）回繰り返す．

手順6. 手順 1～手順 5 の繰り返し：手順 1～手順 5 を少数派の各クラスに対して実行する．

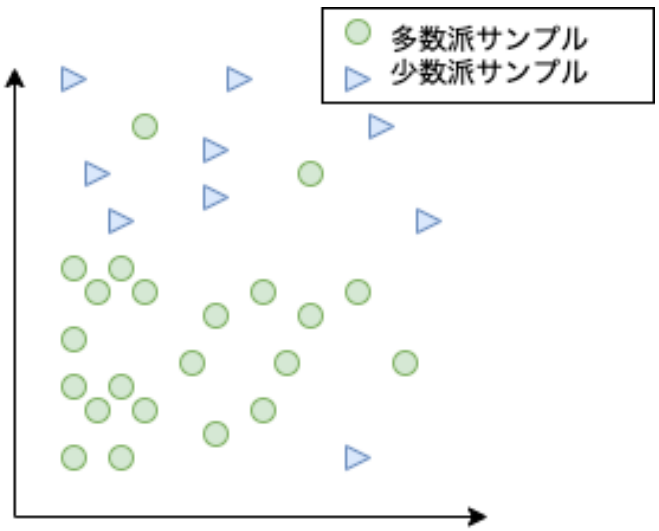

図 3.6: データセット例

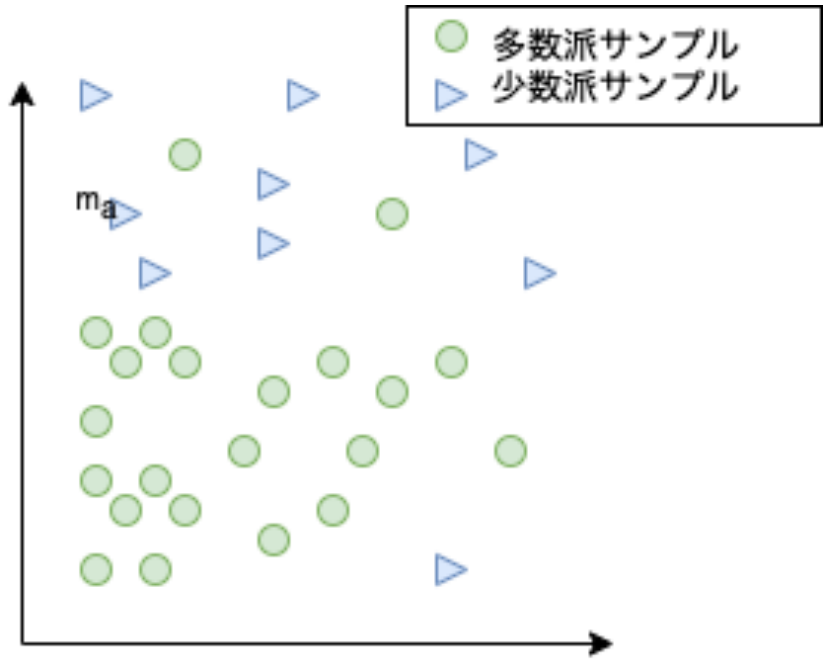

図 3.7: 手順 1 ランダムなm_aの選定

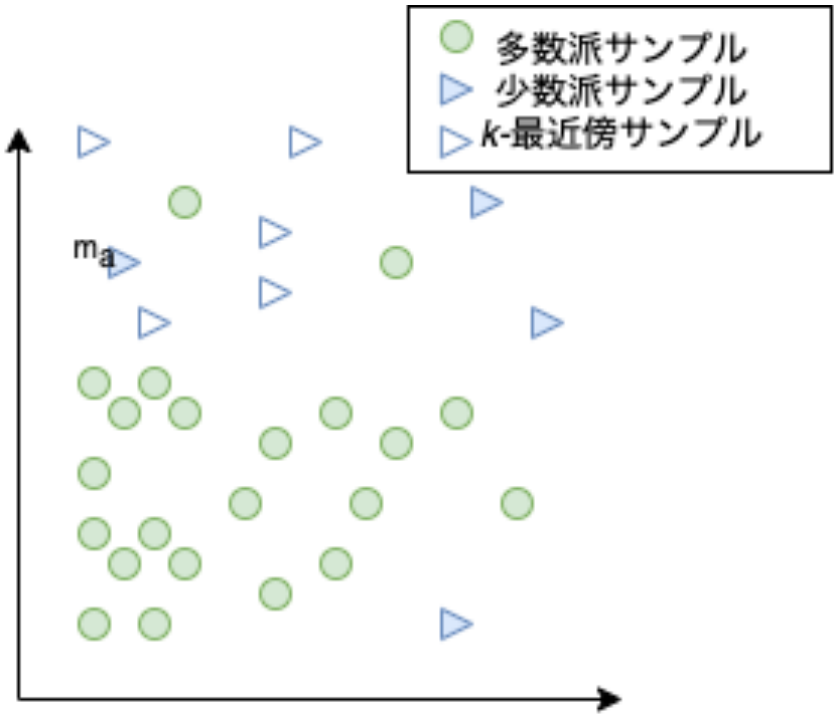

図 3.8: 手順 2 k-最近傍の特定

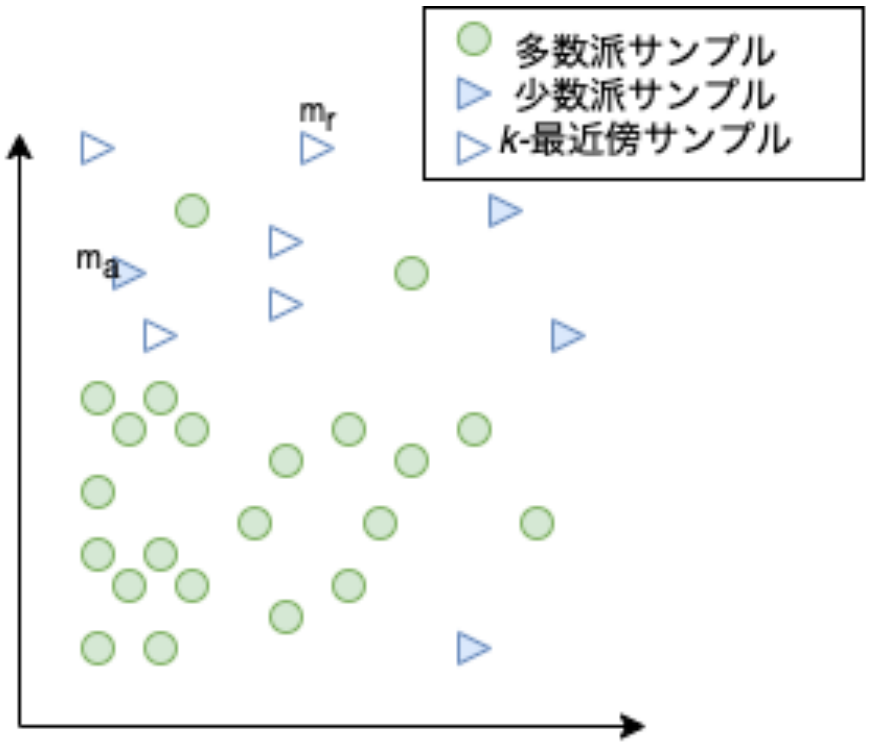

図 3.9: 手順 3 ランダムなm_rの選定

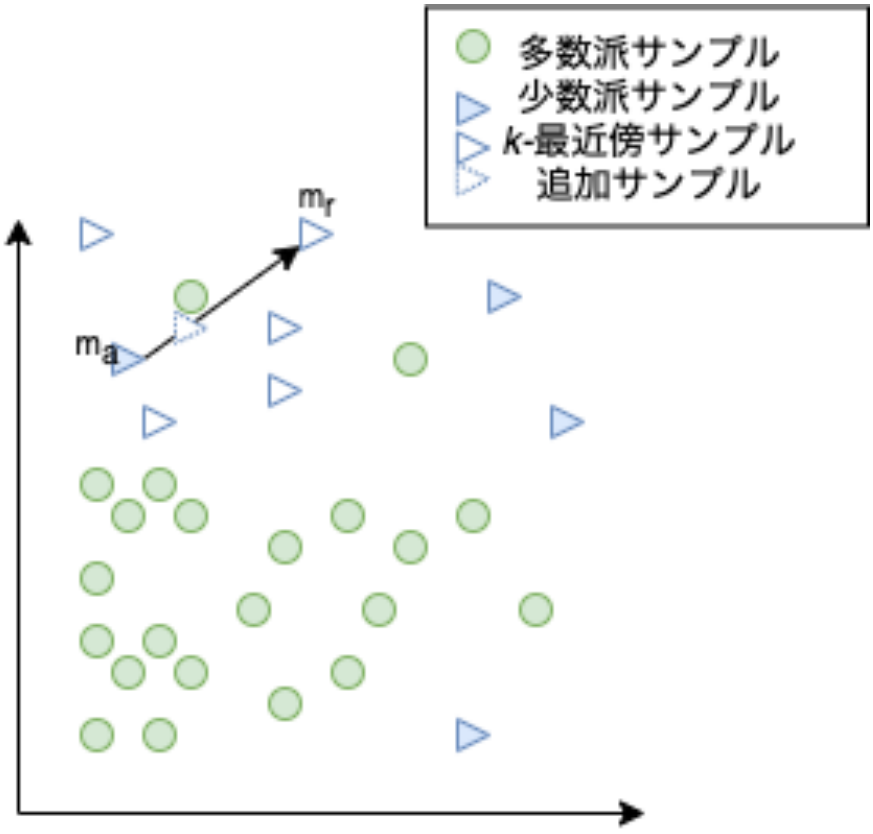

図 3.10: 手順 4 新しいサンプルの生成

第4章　　実験・評価

4.1 概要

本章では，評価実験に関して詳説する．提案手法であるサンプル数の最適化を実験・評価する．4.2 節でデータセットの説明，4.3 節で実験条件，4.4 節で評価実験を述べる．

4.2 データセット

データセットとして OGVC[8]（感情評定値付きオンラインゲーム音声チャットコーパス: Online Gaming Voice Chat corpus with emotional label）を用いる．OGVC[8]は 2 種類の感情音声で構成されており，一つは自然に表出した感情を含んだ自然発話音声で，もうひとつは演技による感情音声である．自然発話音声は，自発的で，活き活きしとした感情を誘発させるためオンラインゲームをプレイしている 2〜3 人程度の音声チャットのやりとりを収録している．それを 6 グループつくり収録することで，男性 9 名，女性 4 名の 13 話者，約 1 万発話が収録されており，3 人の第三者により感情評定値が付与されている．音声データは 48kHZ，16bit，モノラルで記録されている．また，演技による感情音声は，感情表現が得意であるプロの俳優が演じた音声である．これは自然発話音声収録時に転記テキストを作成し，その文章を男性 2 名，女性 2 名のプロの俳優が感情を込めて発声した音声である．

4.2.1　自然発話音声

自然発話音声は，MMORPG (Massively Multiplayer Online Role-Playing Game)と呼ばれるオンラインゲームで遊んでいる最中に 2 名ないしは 3 名の話者間で交わした対話音声 6 対話分，計 9,114 発話を収録している．対話の話者はオンラインゲームの経験がある大学生 13 名（男性 9 名，女性 4 名）で，親近性のある同性同士が参加している．転記テキストには以下の情報が記載されている．

- 発話番号（対話内の通し番号）
- 開始時刻（発話の開始時刻．単位は秒）
- 終了時刻（発話の終了時刻．単位は秒）
- 話者番号（対話内の発話者 ID）
- 発話内容（発話内容の転記）

また，発話単位基準を 400ms 以上のポーズによって挟まれた音声の範囲としている．転

記テキストの一部を以下の図 4.1 に示す． z

```
001,0.0925,1.4850,B:,組めるよね？、ぎりぎり
002,1.9775,2.2100,B:,うん
003,3.4625,3.9325,B:,よく
004,4.8075,7.7075,B:,よくキュージューロクまで上げたよね。{笑}
005,9.3825,10.8675,B:,全部ソロでしょうしかも？
006,14.2375,14.6800,B:,マジ？
007,15.7525,17.1950,B:,びっくりするよね。{笑}
008,18.2500,19.5575,B:,あ、ん別に
009,22.7150,23.2525,B:,{笑}
010,24.0175,24.2100,B:,{笑}
011,26.7825,27.1800,B:,{笑}
012,29.3475,29.7150,B:,うん
013,32.2700,32.6850,B:,ああ
014,34.9900,35.2525,B:,うん
015,40.2075,40.5775,B:,うん
016,46.3875,46.7300,B:,うん
017,50.4525,51.4675,B:,ロクジューナナねー。
018,51.8675,53.5450,B:,それもアコとは組んでるよね。
019,57.6925,58.0700,B:,{笑}
020,65.9075,66.1325,B:,うん
021,67.9075,68.1850,B:,うん
022,70.4925,72.4425,B:,うんまあそれもありっちゃありです。{笑}
023,72.9150,73.0825,B:,{笑}
024.76.3500,78.7475.B:.あー、コンロンでもいいかも。
```

図 4.1 OGVC[8]中の転記テキストの一部（左から「,」区切りで，発話番号，発話開始時間，発話終了時間，話者番号，発話内容を示している.）

4.2.2 感情ラベル

収録した 9,114 発話のうち，振幅レベルが小さく評定に使用できないと判断した発話を除外した 6,578 発話に対して，第三者である 3 人の評価者が感情のアノテーションを施している．感情の分類として Pluchik ら[6]の基本 8 感情と「平静」と「その他」の 10 個のラベルを用いている．図 4.2 に感情分類のラベルを OGVC[8]より抜粋する．

分類	記号	説明
喜び	JOY	良いことに出会って非常に満足し，うれしい，ありがたいと思う感情
受容	ACC	心がひきつけられ，積極的に受け入れよう，接し続けようとする感情
恐れ	FEA	危害が及ぶことを心配してびくびくし，その人やその物と接することを避けたがる感情
驚き	SUR	意外なことを見聞きして心が強く動揺し，平静を失う，どう判断すべきか戸惑う感情
悲しみ	SAD	不幸なことに出会った時など，取り返しのつかない事を思い続けて泣きたくなる感情
嫌悪	DIS	その状態・行為をすんなりと受け入れることができず，避けようとする感情
怒り	ANG	許しがたい事柄に接し，不快感を抑えきれず，いらだった状態の感情
期待	ANT	望ましい事態の実現，好機の到来を心から待つ感情
平静	NEU	まったく感情が表れていない
その他	OTH	ノイズが大きい場合など 8 種の感情に分類不能のもの

図 4.2: OGVC[8]の感情分類ラベル

表 4.1: OGVC[8]の感情種別ごとの発話数

感情	記号	発話数[発話]
喜び	JOY	627
受容	ACC	777
恐れ	FEA	361
驚き	SUR	697
悲しみ	SAD	402
嫌悪	DIS	737
怒り	ANG	321
期待	ANT	831
平静	NEU	1322
その他	OTH	503
合計	-	6578

4.3 実験条件

本実験では CNN と LSTM による学習のミニバッチ数は 25 とし, 損失関数をクロスエントロピー誤差，オプティマイザ（最適化アルゴリズム）には Adam を使用した．学習率は 0.001，β_1は 0.9，β_2は 0.999 とした．最大エポック数は 50 とした．実験は Google Colaboratory[4]を用いた．ハードウェアアクセラレータに GPU を使用し，python3.6.9 にて実装した．

訓練・評価に使用するのは OGVC[8]の自然発話音声のみである．使用する感情は阿部ら[9]に倣って，「怒り」「喜び」「悲しみ」「驚き」「平静」の 5 つの感情とした．5 つそれぞれの感情音声について，3.4 節で示したようにメルスペクトラムに変換した際の時間軸方向の大きさが，200 より大きくなる発話を除いたサンプルを訓練・評価に用いる．各感情の総発話数，訓練・評価用発話数のまとめを表 4.2 に示す．訓練用・評価用のデータには，各話者はランダムで存在している．また，各話者は訓練・評価データ中にそれぞれ含まれる．自然発話音声を訓練データとテストデータの比率を 3:1 になるようランダムに分割した. 訓練データは全 2512 発話，テストデータは 838 発話となった．しかし，評価データの感情によってサンプル数が大きく異なってしまうため，評価用データの各感情分類において 80 発話ずつになるようにランダムにダウンサンプリングした．評価データの感情分類によりサンプル数が異なる場合，ある特定のサンプル数が多い感情分類を推測できるモデルの精度が高く評価されてしまう．そのため評価データの感情分類ごとのサンプル数は等しくなるようにするためである．これにより評価データは計 400 発話となった．OGVC を訓練データと評価データにランダムに分割する際には scikit-learn[5]の train_test_split により shuffle=True で用いた．サンプル数を減らす際，scikit-learn[6]の resample 関数を用いた．

[4] https://colab.research.google.com/notebooks/welcome.ipynb?hl=ja

[5] https://scikit-learn.org/stable/

[6] https://scikit-learn.org/stable/

表 4.2: OGVC の自然発話音声サンプル数

感情	総発話数	訓練用発話数	評価用発話数	評価用発話数（調整後）
怒り	318	238	80	80
喜び	618	463	155	80
悲しみ	401	301	100	80
驚き	697	523	174	80
平静	1316	987	329	80
総発話数	3350	2512	838	400

4.4 評価実験

本実験では，4.4.1 節にて提案手法であるアップサンプリングによる実験を行う．その後 4.4.2 節にて LSTM のハイパーパラメータ（隠れ層のサイズ，出力サイズ）の調整を実施する．

4.4.1 感情音声のアップサンプリング評価

まず，提案手法のサンプル数のアップサンプリングによる効果を測定するため，表 4.2 に示すアップサンプリングを何も行わないサンプル数で訓練を行った．評価用データとして，「怒り」「喜び」「平静」をダウンサンプリングした表 4.2 中の『評価用発話数（調整後)』を用いた．以下の図では，怒りを ANG，喜びを JOY，悲しみを SAD，驚きを SUR，平静を NEU としている．また，入力がメルスペクトラムの場合，BiLSTM を 3 層，隠れ層の出力サイズを 512 とした．入力が 3 次元メルスペクトラムの場合，BiLSTM を 2 層，隠れ層の出力サイズを 256 とした．図 4.3，図 4.4 に入力特徴量をそれぞれメルスペクトラムとしたとき，3 次元メルスペクトラムとしたときの実験結果を示す．

表 4.3: 訓練用・評価用データサンプル数

感情	訓練用発話数	評価用発話数（調整後）
怒り	238	80
喜び	463	80
悲しみ	301	80
驚き	523	80
平静	987	80
総発話数	2512	400

図 4.3: 表 4.2 のサンプル数による実験結果 Accuracy=25.8%（入力：メルスペクトラム）（単位[%]）

図 4.4: 表 4.2 のサンプル数による実験結果 Accuracy=25.3%（入力：3 次元メルスペクトラム）（単位[%]）

次に，提案手法である 3.4 節のアップサンプリング手法の評価実験を行う．アップサンプリングを実施するのは訓練用のデータのみとする．実験①・実験②・実験③において訓練用データのサンプル数を表 4.3 に示す．計 673 発話の評価用データを，評価用データセットとして，すべての実験において用いる．

- 実験①は 3.4 節のアップサンプリングにより，「怒り」と「悲しみ」の音声特徴量の前・後それぞれに 0 パティングを施したものを追加し，「喜び」と「驚き」の音声特徴量の前・後をランダムに選び 0 パティングを施したものを追加した．その結果「怒り」と「悲しみ」のサンプル数が 3 倍に，「喜び」と「驚き」のサンプル数は 2 倍となった．「平静」はそのままのサンプル数とした．

- 実験②は，実験③において比較実験を行うため，提案手法である 3.4 節のアップサンプリング手法のみを用いて各感情 714 発話ずつで訓練した．具体的には，サンプル数最適化前の「怒り」「喜び」「悲しみ」「驚き」の訓練データをランダムに取り出し，音声特徴量の前・後をランダムに選び 0 パティングを施したものを追加する．ただし，「怒り」「悲しみ」については音声特徴量の前・後をランダムに選び 0 パティングを施したものを追加した後さらに，訓練データをランダムに取り出し 0 パティングが施されていない箇所（前もしくは後）に 0 パティングを施した．また，「平静」は 987 発話あったものを，ランダムにダウンサンプリングし 714 発話とした．

- 実験③は一般的にアップサンプリング手法として用いられる Chawla ら[28]の提案した SMOTE(Synthetic Minority Over-sampling Technique)を用いて，サンプル数最適化前（「怒り」238 発話，「喜び」463 発話，「悲しみ」301 発話，「驚き」523 発話，「平静」987 発話）の音声感情のアップサンプリングを行い各感情 714 発話とした．

表 4.4:訓練を行うサンプル数（単位は発話数）

	怒り	喜び	悲しみ	驚き	平静	合計
サンプル数 最適化前	238	463	301	523	987	2512
実験①	714	926	903	1046	987	4576
実験②	714	714	714	714	714	3570
実験③	714	714	714	714	714	3570

表 4.3 の訓練サンプル数で訓練を行い，表 4.2 に示す計 673 発話の評価用データを用いた評価実験の結果を図 4.5 から図 4.10 に示す．

図 4.5:実験①の結果平均 Accuracy=25.8%（入力：メルスペクトラム）（単位[%]）

図 4.6:実験①の結果平均 Accuracy=25.8%（入力：3 次元メルスペクトラム）（単位[%]）

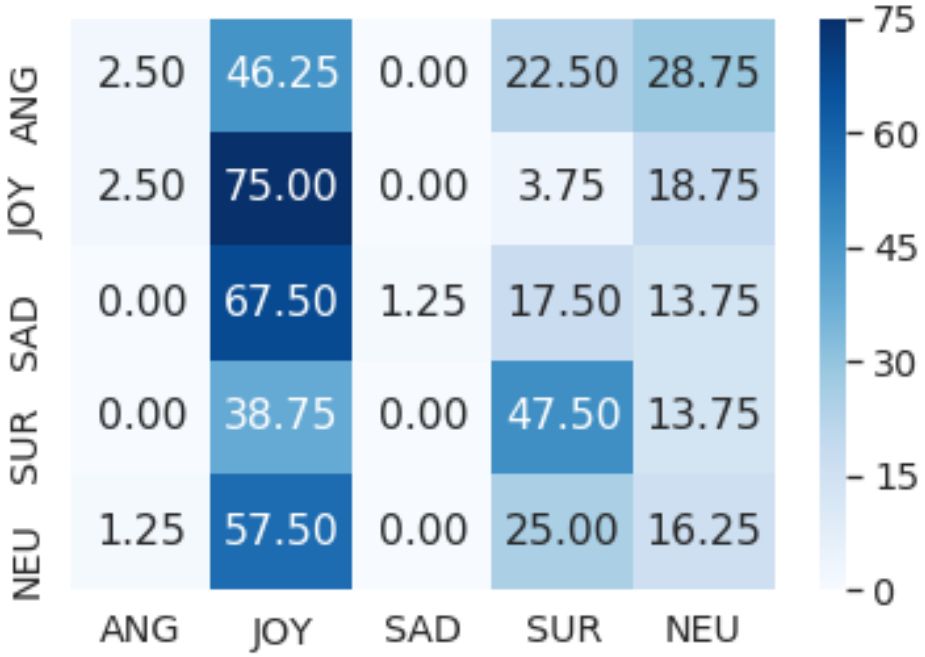

図 4.7:実験②の結果 平均 Accuracy=28.5%（入力：メルスペクトラム）（単位[%]）

図 4.8 実験②の結果 平均 Accuracy=31.5%（入力：3 次元メルスペクトラム）（単位[%]）

図 4.9: 実験③の結果 平均 Accuracy=25.0%（入力：メルスペクトラム）（単位[%]）

図 4.10: 実験③の結果 平均 Accuracy=31.0%（入力：三次元メルスペクトラム）（単位[%]）

提案手法によるアップサンプリングとの比較を表 4.5，表 4.6 に示す．表 4.7 にアップサンプリング前，提案手法によるアップサンプリング，SMOTE（Synthetic Minority Over-sampling Technique）によるアップサンプリングの 3 つの Accuracy と，感情認識率の標準偏差を示す．提案手法のアップサンプリングにより，入力をメルスペクトラムとしたときに，Accuracy は 25.8%から 28.5%となった．標準偏差は 30.7 から 28.6 となった．入力を 3 次元メルスペクトラムとしたとき，Accuracy は 25.8%から 31.5%となった．標準偏差は 24.0 から 19.4 となった．SMOTE によるアップサンプリングでは，入力をメルスペクトラムとしたとき Accuracy25.0%，標準偏差は 34.5，入力を 3 次元メルスペクトラムとしたとき Accuracy は 31.0%，標準偏差は 21.4 なった．入力をメルスペクトラムとしたとき提案手法は，アップサンプリング前と比較して Accuracy を 2.7%（アップサンプリング前の Accuracy の 10.5%）上げ，標準偏差を 2.1(アップサンプリング前の標準偏差の 6.8%)下げた．入力を 3 次元メルスペクトラムとしたとき提案手法は，アップサンプリング前と比較して Accuracy を 5.7%(アップサンプリング前の Accuracy の 22%)上げ, 標準偏差を 4.6(アップサンプリング前の標準偏差の 19%)下げた．SMOTE によるアップサンプリングは入力がメルスペクトラムの場合, Accuracy を 0.8%（アップサンプリング前の Accuracy の 3%）下げ，標準偏差を 3.8(アップサンプリング前の標準偏差の 14.7%)上げた．入力を 3 次元メルスペクトラムとしたとき, Accuracy を 5.2%(アップサンプリング前の Accuracy の 20.2%)上げ，標準偏差を 3.6(アップサンプリング前の標準偏差の 15%)下げた．以上の実験結果から，提案手法によるアップサンプリングにより，感情認識精度が向上し，感情認識率のばらつき（標準偏差）を低下させることが分かった．要因として，サンプル数の向上以外変化していないため「サンプル数の向上」が考えられる．本提案手法の仮説である「徐々に他人の感情を理解できるようになること」が，提案手法のアップサンプリングによる「サンプル数の向上」により実現できることが確かめられたと言える．

表 4.5: アップサンプリング前と提案手法の実験結果の比較（入力：メルスペクトラム）

	アップサンプリング前		実験②		評価発話数
感情	訓練発話数 [発話]	精度[%]	訓練発話数 [発話]	精度[%]	[発話]
怒り	238	5.00	714	2.50	80
喜び	463	8.75	714	75.0	80
悲しみ	301	1.25	714	1.25	80
驚き	523	30.0	714	47.5	80
平静	987	83.75	714	16.3	80
Accuracy	−	25.8	−	28.5	−

表 4.6: アップサンプリング前と提案手法の実験結果の比較（入力：三次元メルスペクトラム）（単位[%]）

	アップサンプリング前		実験②		評価発話数
感情	訓練発話数［発話］	精度[%]	訓練発話数［発話］	精度[%]	［発話］
怒り	238	0.00	714	15.0	80
喜び	463	30.0	714	13.8	80
悲しみ	301	0.00	714	30.0	80
驚き	523	35.0	714	31.3	80
平静	987	63.8	714	67.5	80
Accuracy	−	25.8	−	31.5	−

表 4.7: 提案手法と SMOTE(Synthetic Minority Over-sampling Technique)との比較

	Accuracy[%]	感情認識率の標準偏差
アップサンプリング前（入力：メルスペクトラム）	25.8	30.7
アップサンプリング前（入力：3 次元メルスペクトラム）	25.8	24.0
SMOTE によるアップサンプリング（入力：メルスペクトラム）	25.0	34.5
SMOTE によるアップサンプリング（入力：3 次元メルスペクトラム）	31.0	21.4
提案手法 実験②（入力：メルスペクトラム）	**28.5**	**28.6**
提案手法 実験②（入力：3 次元メルスペクトラム）	**31.5**	**19.4**

4.4.2 LSTM のハイパーパラメータ最適化実験

4.4.2 節では交差エントロピー誤差を最小化する LSTM の層数, 隠れ層への出力サイズの最適化を行う．具体的には，LSTM 層を 1 層，2 層，3 層，出力サイズを 128，256，512 とした実験を行い，Accuracy を比較する．訓練用データには，実験②のサンプル数（各感情）を用いる．感情音声は 1 次元メルスペクトラムと 3 次元メルスペクトラムのそれぞれについて行う. 実験結果を表 4.7 に示す. ただし, LSTM 層の層数と出力サイズを$\mathrm{LSTM}_{\mathrm{j}}^{i}$と表す. iは層数，jは出力サイズを示す．数値は Accuracy(%)，括弧内の数値は交差エントロピー誤差である．

表 4.8 LSTM のハイパーパラメータ最適化実験結果（単位：Accuracy[%](交差エントロピー誤差)）

LSTM 設定	LSTM_{128}^{1}	LSTM_{128}^{2}	LSTM_{128}^{3}	LSTM_{256}^{1}	LSTM_{256}^{2}	LSTM_{256}^{3}	LSTM_{512}^{1}	LSTM_{512}^{2}	LSTM_{512}^{3}
入力：メルスペクトラム	30.8 (1.62)	25.0 (1.56)	**27.8 (1.51)**	29.3 (1.56)	28.0 (1.62)	22.0 (1.57)	28.25 (1.56)	26.0 (1.60)	30.5 (1.53)
入力：3 次元メルスペクトラム	30.0 (1.59)	31.3 (1.57)	24.5 (1.59)	25.0 (1.63)	21.3 (1.70)	28.3 (1.61)	**31.8 (1.56)**	26.5 (1.58)	26.5 (1.65)

表 4.8 より，交差エントロピー誤差を基準にすると，入力をメルスペクトラムにした場合LSTM_{128}^{3}が，入力を 3 次元メルスペクトラムにした場合LSTM_{512}^{1}が LSTM のハイパーパラメータとして最適であると分かった. このときの実験結果をそれぞれ図 4.11, 図 4.12 に，最終結果を表 4.9 に示す．この際の Accuracy と感情認識率の標準偏差を表 4.9 に合わせて示す.

図 4.11: LSTM_{128}^{3}の実験結果 Accuracy=27.8%（入力：メルスペクトラム）（単位[%]）

図 4.12: LSTM_{512}^{1}の実験結果 Accuracy=31.8%（入力：3 次元メルスペクトラム）（単位[%]）

表 4.9: LSTM のハイパーパラメータ調整後の実験結果

	Accuracy[%]	感情認識率の標準偏差
入力：メルスペクトラム	27.8	21.7
入力：3 次元メルスペクトラム	31.8	18.5

4.4.3 感情音声サンプル数の最適化実験

4.4.3 節では，感情音声サンプル数の最適化実験を行う．具体的には，入力を 3 次元メルスペクトラム，4.4.2 節で最適だと分かったLSTM_{512}^{1}を用いて，表 4.10 に示すサンプル数で訓練する．図 4.12 より訓練用サンプル数が各感情音声 712 発話での結果から以下のことが分かる．

- 「怒り」の 62.5%が「平静」，15.0%が「悲しみ」，20%が「驚き」と誤分類されてしまう．
- 「喜び」の 20.0%が「悲しみ」に，57.5%が「平静」と誤分類されてしまう．
- 「悲しみ」の 15.00%が「驚き」に，57.5%が「平静」と誤分類されてしまう．
- 「驚き」の 17.5%が「悲しみ」に，32.5%が「平静」と誤分類されてしまう．
- 「平静」の 42.5%が正しく推測でき，25.0%が「悲しみ」に 26.25%が「驚き」と誤分類されてしまう．

以上のことから，「平静」に誤分類されることが最も多いことが分かる．したがってまず，「平静」のサンプル数の最適化を検討する．その後，「平静」のサンプル数最適化後に誤分類が最も多い感情に着目して 5 感情のサンプル数最適化を図る．訓練用サンプルには 4.4.1 節にて使用したアップサンプリング前の音声感情をランダムに抽出し，提案手法によりアップサンプリングにより表 4.10 に示すサンプル数までアップサンプリングする．評価には表 4.2 に示す計 400 発話の評価用発話を用いる．サンプル数①～サンプル数⑩の実験意図を以下に示す．

サンプル数①. 誤分類の多かった「平静」のサンプル数を 642 発話とする．その他を 714 発話（図 4.12 と同じサンプル数）とする．「平静」の最適なサンプル数調査のためである．

サンプル数②. 誤分類の多かった「平静」のサンプル数を 499 発話とする．その他をサンプル数①と同様とする．「平静」の最適なサンプル数調査のためである．

サンプル数③. 誤分類の多かった「平静」のサンプル数を 464 発話とする．その他をサンプル数①と同様とする．「平静」の最適なサンプル数調査のためである．

サンプル数④. 誤分類の多かった「平静」のサンプル数を 392 発話とする．その他をサンプル数①と同様とする．「平静」の最適なサンプル数調査のためである．

サンプル数⑤. サンプル数④にて誤分類が最も多い「喜び」のサンプル数を 642 発話とする．その他のサンプル数をサンプル数③と同様とする．

サンプル数⑥. サンプル数④にて誤分類が最も多い「喜び」のサンプル数を 606 発話とする．その他のサンプル数をサンプル数③と同様とする．

サンプル数⑦. サンプル数④にて誤分類が最も多い「喜び」のサンプル数を 571 発話とする．その他のサンプル数をサンプル数③と同様とする．

サンプル数⑧. サンプル数④にて誤分類が最も多い「喜び」のサンプル数を 535 発話とする．その他のサンプル数をサンプル数③と同様とする．

サンプル数⑨. サンプル数⑦にて誤分類が最も多い「悲しみ」のサンプル数を 678 発話とする．その他のサンプル数をサンプル数⑦と同様とする．

サンプル数⑩. サンプル数⑦にて誤分類が最も多い「悲しみ」のサンプル数を 642 発話とする．その他のサンプル数をサンプル数⑦と同様とする．

表 4.10: 訓練に用いる発話数 （単位：発話）

	怒り	喜び	悲しみ	驚き	平静	合計
サンプル数①	714	714	714	714	642	3498
サンプル数②	714	714	714	714	499	3355
サンプル数③	714	714	714	714	464	3320
サンプル数④	714	714	714	714	392	3248
サンプル数⑤	714	642	714	714	464	3248
サンプル数⑥	714	606	714	714	464	3212
サンプル数⑦	714	571	714	714	464	3177
サンプル数⑧	714	535	714	714	464	3141
サンプル数⑨	714	642	678	714	464	3212
サンプル数⑩	714	642	642	714	464	3176

サンプル数①～サンプル数⑩の実験結果を，図 4.13 から図 4.22 に示す．最適なサンプル数を目指すにあたって，アップサンプリング数を変更後に「Accuracy が低下する」「認識率が Accuracy よりも小さくなる」の 2 つの事象が観測された際に，変更前のサンプル数が最適であると判断した．「Accuracy が低下する」ことは誤認識が増えるため，本実験が目指す最適化としてふさわしくない．「認識率が Accuracy よりも小さくなる」ことは，訓練サンプル不足により認識率の低下を招いていると考えられる．以上の評価基準から，サンプル数⑨が OGVC における最適サンプル数だと判断した．サンプル数⑨において，追加でランダムにダウンサンプリングを行い（訓練用発話数と同一とし，訓練用発話が異なる訓練データを用いた），２回の追加実験を行った．この結果を図 4.23，図 4.24 に示す．

図 4.13: サンプル数①の実験結果 Accuracy=24.8%（単位[%]）

図 4.14: サンプル数②の実験結果 Accuracy=26.3%（単位[%]）

図 4.15: サンプル数③の実験結果 Accuracy=31.5%（単位[%]）

図 4.16:サンプル数④の実験結果 Accuracy=25.3%（単位[%]）

図 4.17: サンプル数⑤の実験結果 Accuracy=31.0%（単位[%]）

図 4.18: サンプル数⑥の実験結果 Accuracy=31.3%（単位[%]）

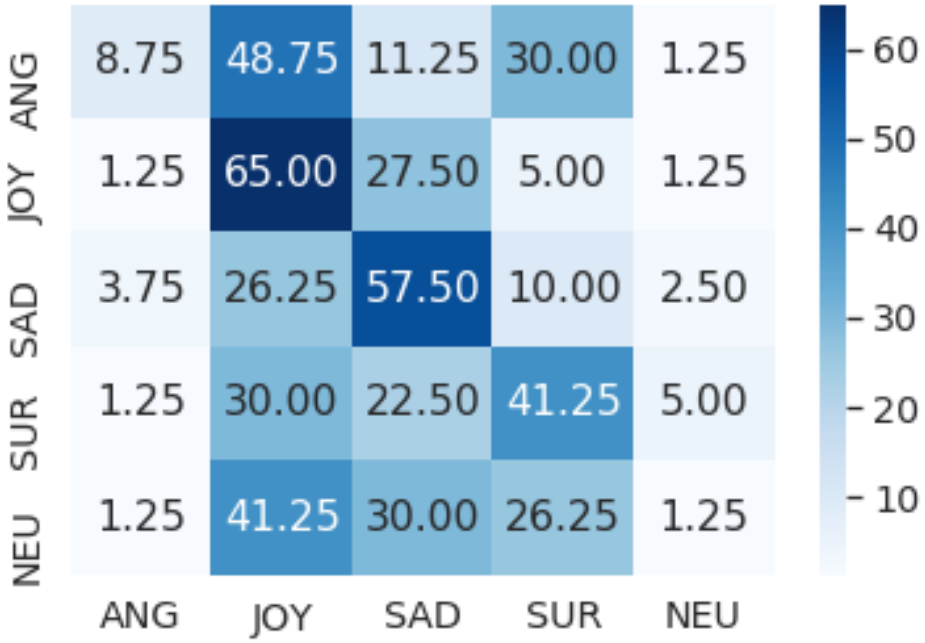

図 4.19: サンプル数⑦の実験結果 Accuracy=34.8%（単位[%]）

図 4.20: サンプル数⑧の実験結果 Accuracy=28.3%（単位[%]）

図 4.21: サンプル数⑨の実験結果 Accuracy=33.0%（単位[%]）

図 4.22: サンプル数⑩の実験結果 Accuracy=27.5%（単位[%]）

図 4.23: サンプル数⑨の追加実験 Accuracy=36.0%（単位[%]）

図 4.24: サンプル数⑨の追加実験 Accuracy=27.0%（単位[%]）

サンプル数⑨における Accuracy と認識率の標準偏差, 4.4.2 節にて実験した最適な LSTM のハイパーパラメータを識別器とし，SMOTE でアップサンプリングした訓練データにより訓練した実験結果を表 4.11 に示す．SMOTE は入力がメルスペクトラム，3 次元メルスペクトラムの際，それぞれ Accuracy が 33.8%，31.5%であるのに対し提案手法は最大 Accuracy36.0%であった．感情認識率の標準偏差は SMOTE の場合入力がメルスペクトラム，3 次元メルスペクトラムの際，それぞれ 20.0，21.7 であるのに対し提案手法は 17.5 であった．アップサンプリング前と比較して，Accuracy は 25.8%から 36.0%に，感情認識率の標準偏差は 24.0 から 17.5 となっていることが分かる．

表 4.11: サンプル数⑤の実験結果と SMOTE を用いた実験結果の比較

手法	Accuracy[%]	感情認識率の標準偏差
アップサンプリング前（入力：メルスペクトラム）	25.8	30.7
アップサンプリング前（入力：3 次元メルスペクトラム）	25.8	24.0
SMOTE によるアップサンプリング（入力：メルスペクトラム，LSTM_{128}^{3}）	33.8	20.0
SMOTE によるアップサンプリング（入力：3 次元メルスペクトラム，LSTM_{512}^{1}）	31.5	21.7
提案手法（LSTM_{512}^{1}，サンプル数⑨，入力：3 次元メルスペクトラム）	**36.0**	**17.5**

第5章　　おわりに

本研究では，日本人の自然対話における音声感情認識の精度向上の一助となることを目的とした．そこで，訓練に用いる感情音声のサンプル数のアップサンプリング手法と感情音声サンプル数の最適化を提案した．入力として 3 次元メルスペクトラムを用いて，感情分類を「怒り」「喜び」「悲しみ」「驚き」「平静」の 5 つの感情とした．学習器には CNN と BiLSTM を組み合わせたモデルを構築し，訓練・評価実験を行った．

今後の展望として，話者独立の音声感情認識モデルの構築する．本実験の仮説として，人間も徐々に他人の感情を理解できるようになること，他人の感情には分かりづらい感情（悲しい，恐れ等）もあることの 2 点に基づいている．本実験では，人物の特定をしなくても感情音声のサンプル数を最適化することにより，音声感情認識の Accuracy が高まり，認識率の標準偏差が小さくなることが分かった．人によって「怒り」「悲しみ」「喜び」「驚き」の表出の度合いは異なるため，話者によりサンプル数の最適化を実施することが理想的と言える．

謝辞

本研究を行うにあたり，大変多くの方のご協力をいただきました．山名早人教授には，ゼミやディスカッション，論文の添削，研究の方針指導など，数多くのご指導をいただきました．心より御礼申し上げます．また，研究が行き詰まったときに様々なアイデアを出していただいた研究室の同期の方々にも大変お世話になりました．深く感謝いたします．また，本学の大学院を通うことを無条件で支えて頂いた両親にただただ感謝に堪えません．

参考文献

[1] 統計数理研究所, “「日本人の国民性 第13次全国調査」の結果のポイント,” *https://www.ism.ac.jp/kokuminsei/resources/KS13print.pdf*, pp. 1–11, 2016.（参照: 2020/01/05）

[2] W. Sato, S. Hyniewska, K. Minemoto, and S. Yoshikawa, “Facial expressions of basic emotions in Japanese laypeople,” *Front. Psychol.*, vol. 10, pp. 1–11, 2019.

[3] A. R. Mohamed, G. E. Dahl, and G. Hinton, “Acoustic modeling using deep belief networks,” *IEEE Trans. Audio, Speech Lang. Process.*, vol. 20, no. 1, pp. 14–22, 2012.

[4] T. N. Sainath, R. J. Weiss, A. Senior, K. W. Wilson, and O. Vinyals, “Learning the speech front-end with raw waveform CLDNNs,” *Proc. Annu. Conf. Int. Speech Commun. Assoc. INTERSPEECH*, pp. 1–5, 2015.

[5] P. Ekman, “An argument for basic emotions,” *Cogn. Emot.*, vol. 6, no. 3–4, pp. 169–200, 1992.

[6] R. Plutchik, *The emotions*. University Press of America, 1991.

[7] J. A. Russell, “Is There Universal Recognition of Emotion From Facial Expression? A Review of the Cross-Cultural Studies,” *Psychol. Bull.*, vol. 115, no. 1, pp. 102–141, 1994.

[8] Y. Arimoto, H. Kawatsu, S. Ohno, and H. Iida, “Emotion recognition in spontaneous emotional speech for anonymity-protected voice chat systems,” *Proc. Annu. Conf. Int. Speech Commun. Assoc. INTERSPEECH*, pp. 322–325, 2008.

[9] 渉阿部, 大介真壁, and 哲夫小坂, “SVMを用いた自然対話音声の認識における週データの検討,” 情報処理学会東北支部研究報告, vol. 2016-7-A3, pp. 1–7, 2016.

[10] Jun, X. Xu, Z. Zhang, S. Fruhholz, and B. Schuller, “Semisupervised Autoencoders for Speech Emotion Recognition,” *IEEE/ACM Trans. Audio Speech Lang. Process.*, vol. 26, no. 1, pp. 31–43, Jan. 2018.

[11] P. Song, “Transfer linear subspace learning for cross-corpus speech emotion recognition,” *IEEE Trans. Affect. Comput.*, vol. 10, no. 2, pp. 265–275, Apr. 2019.

[12] Y. Xie, R. Liang, Z. Liang, C. Huang, C. Zou, and B. Schuller, “Speech Emotion Classification Using Attention-Based LSTM,” *IEEE/ACM Trans. Audio Speech Lang. Process.*, vol. 27, no. 11, pp. 1675–1685, Jul. 2019.

[13] Y. Kim and E. M. Provost, “ISLA: Temporal segmentation and labeling for audio-visual emotion recognition,” *IEEE Trans. Affect. Comput.*, vol. 10, no. 2, pp. 196–208, Apr. 2019.

[14] L. Guo, L. Wang, J. Dang, Z. Liu, and H. Guan, “Exploration of Complementary Features for Speech Emotion Recognition Based on Kernel Extreme Learning Machine,” *IEEE Access*, vol. 7, pp. 75798–75809, 2019.

[15] Z. Zhao *et al.*, “Exploring Deep Spectrum Representations via Attention-Based Recurrent and Convolutional Neural Networks for Speech Emotion Recognition,” *IEEE Access*, vol. 7, pp. 97515–97525, Jul. 2019.

[16] R. Lotfian and C. Busso, “Curriculum learning for speech emotion recognition from crowdsourced labels,” *IEEE/ACM Trans. Audio Speech Lang. Process.*, vol. 27, no. 4, pp. 815–826, Apr. 2019.

[17] H. Meng, T. Yan, F. Yuan, and H. Wei, “Speech Emotion Recognition from 3D Log-Mel Spectrograms with Deep Learning Network,” *IEEE Access*, vol. 7, pp. 125868–125881, Aug. 2019.

[18] 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, Sep. 2019.

[19] A. M. Badshah *et al.*, “Deep features-based speech emotion recognition for smart affective services,” *Multimed. Tools Appl.*, vol. 78, no. 5, pp. 5571–5589, 2019.

[20] I. Shahin, A. B. Nassif, and S. Hamsa, “Emotion Recognition Using Hybrid Gaussian Mixture Model and Deep Neural Network,” *IEEE Access*, vol. 7, pp. 26777–26787, 2019.

[21] E. N. N. Ocquaye, Q. Mao, H. Song, G. Xu, and Y. Xue, “Dual exclusive attentive transfer for unsupervised deep convolutional domain adaptation in speech emotion recognition,” *IEEE Access*, vol. 7, pp. 93847–93857, Jun. 2019.

[22] S. Zhang, X. Zhao, and Q. Tian, “Spontaneous Speech Emotion Recognition Using Multiscale Deep Convolutional LSTM,” *IEEE Trans. Affect. Comput.*, pp. 1–1, 2019.

[23] P. Jiang, H. Fu, H. Tao, P. Lei, and L. Zhao, “Parallelized Convolutional Recurrent Neural Network with Spectral Features for Speech Emotion Recognition,” *IEEE Access*, vol. 7, pp. 90368–90377, Jul. 2019.

[24] A. Dhall, O. V. Ramana Murthy, R. Goecke, J. Joshi, and T. Gedeon, “Video and image based Emotion recognition challenges in the wild: EmotiW 2015,” *ICMI 2015 - Proc. 2015 ACM Int. Conf. Multimodal Interact.*, pp. 423–426, 2015.

[25] S. Zhalehpour, O. Onder, Z. Akhtar, and C. E. Erdem, “BAUM-1: A Spontaneous Audio-Visual Face Database of Affective and Mental States,” *IEEE Trans. Affect. Comput.*, vol. 8, no. 3, pp. 300–313, 2017.

[26] A. Krizhevsky, I. Sutskever, and G. Hinton, “ImageNet Classification with Deep Convolutional Neural Networks,” *Neural Inf. Process. Syst.*, vol. 25, 2012.

[27] Y. LeCun, P. Haffner, L. Bottou, and Y. Bengio, “Object Recognition with Gradient-Based Learning,” in *Shape, Contour and Grouping in Computer Vision, LNCS, vol.1681,* Berlin, Heidelberg: Springer Berlin Heidelberg, pp. 319–345, 1999.

[28] N. V. Chawla, “SMOTE: Synthetic Minority Over-sampling Technique,” *J. Artif. Intell. Res.*, vol. 16, pp. 321–357, 2002.