

レシピ情報に基づく調理行動認識における 音響識別器の中間特徴量利用に関する検討

宮澤 要二¹ 齋藤 大輔² 峯松 信明²

概要：調理において必要不可欠であるレシピの情報は、現代では様々な媒体を通してアクセス可能である。しかし、調理中に媒体を通してレシピを参照する際、調理が中断されてしまう。この問題を解消するには、調理を妨げない調理支援システムが求められる。そのようなシステムの実現を考えると、システムが自動的に調理状況を判断する必要がある。本研究では、調理者の調理行動に着目して、音響的情報を用いて調理行動認識を検討した。調理行動認識において、従来の研究とは異なり、レシピによる時系列情報を考慮することで認識を行った。その上で、本研究では2つの検討を行った。その結果、特徴量として新たに導入したニューラルネットワークによる音響識別器の中間特徴量を用いることが有効であることが確認された。また、調理行動以外の調理環境音を認識のクラスに加えることが有効であることが示唆された。

1. はじめに

調理を行う際に参照するレシピの情報は様々な媒体を通してアクセスすることができる。レシピ本はもちろん、クックパッド^{*1}をはじめとするレシピ共有のウェブサイトや、YouTube^{*2}のような動画投稿サイトの存在が大きな影響となっているだろう。しかし、調理初心者にとって、レシピ通りに調理を行うことや、動画通りに調理を行うことは簡単ではない。調理をしながらそれらの情報を活用することを考えると、レシピ本を参考にする場合はレシピ本をめくることが必要であり、ウェブサイトを参考にする場合はタブレットを操作する必要がある。調理中は手に油が付着していたり、水で濡れていたいたりするため、これらの操作は調理者にとって煩雑であり、調理が中断されてしまう。

この問題の解消には調理を妨げない調理支援システム、すなわち、調理状況に合わせて次の調理行動を自動的に提示するような調理支援システムが考えられる。その実現には、システムが何らかの情報から調理者の調理状況を判断する必要があり、どのようにして判断を行うかが問題となる。そこで、システムが判断すべき調理状況は大きく2つによって決定されるものとする。1つ目が調理者の行動で、2つ目が調理者が扱う対象である。調理者の行動は「切る」「焼く」「煮る」などの行動のことで、調理者の扱う対象は「にんじん」「キャベツ」「肉」などの材料のこと

である。すなわち、調理行動と調理対象によって調理状況が決定されるものとする。

本研究では、調理を妨げない調理支援システムを目標とし、調理環境音を用いた調理者の調理行動認識を検討する。調理環境音を用いて認識を行う利点は2つ考えられる。1つ目は、ビデオカメラに比べセンサの設置箇所が比較的自由であることがあげられる。2つ目は、視覚的に判別が難しいものでも、調理環境音ならば判別できる可能性があることがあげられる。その代表的なものとして、加熱処理の完了の判別などが考えられる。

具体的には、焼きそばを調理した際の調理環境音を用いて、「切る」と「焼く」の2つの調理行動について認識を行う。目標はオンラインで調理行動の開始と終了時刻を推定することだが、本研究では、オフラインで調理行動の時間区間を推定するシステムを考える。システムには焼きそばのレシピが予め与えられているとし、調理行動の前後関係は「切る」の後に「焼く」となることを考慮して認識を行う。この条件のもとで、音響特徴量として音響識別器による中間特徴量を用いる検討と、「切る」と「焼く」以外のクラスに関する検討を行う。実験を通して、中間特徴量の導入や、調理行動以外のクラスの導入の有効性をF値を指標として検討する。また、調理行動は並列しないものとする。

2. 音響特徴量を用いた認識に関する先行研究

2.1 音響イベント検出

本研究では調理環境音を用いて調理行動を認識することが目的であり、類似の研究として音響イベント検出(Acoustic

¹ 東京大学大学院情報理工学系研究科

² 東京大学大学院工学系研究科

^{*1} <https://cookpad.com>

^{*2} <https://youtube.com>

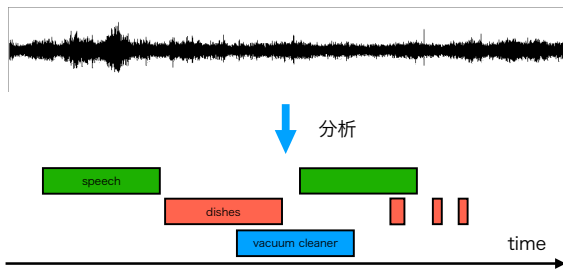


図 1 音響イベント検出の模式図

tic Event Detection; AED) があげられる。音響イベントとは、人の話し声・掃除機の音・動物の鳴き声などのあらゆる音のことを指す。AED では、図 1 のように音響の情報から音響イベントの種類と区間を推定することが目的である。

AED では、隠れマルコフモデル (Hidden Markov Model; HMM) を用いた手法や非負値行列因子分解 (Non-negative Matrix Factorization; NMF) を用いた手法が報告されている [1], [2]。ニューラルネットワークを用いた研究も行われていて、深層ニューラルネットワークを用いる手法や畳み込みニューラルネットワークと回帰型ニューラルネットワークを用いた手法が報告されている [3], [4]。

2.2 音響特徴量を用いた調理行動認識

調理環境音による行動認識をした後に、レシピの時系列情報を考慮することで、最終的な行動認識を行う手法が報告されている [5]。この手法では、特徴量としてメル周波数ケプストラル係数 (Mel Frequency Cepstral Coefficients; MFCC)、モデルに混合ガウス分布 (Gaussian Mixture Model; GMM) を用いて調理行動を「切る」・「焼く」・「その他」の 3 つのクラスに分類する。その後、レシピに基づいた階層型隠れマルコフモデル (Hierarchical Hidden Markov Model; HHMM) に入力することで、最終的な調理行動認識の結果を得る。この手法ではレシピの時系列情報を考慮することで、認識結果の平滑化を行っているといえらる。NMF を用いた調理行動認識手法も報告されている [6]。この手法では、NMF を用いて調理行動を、「切る」・「焼く」・「その他」の 3 つのクラスに分類する。また、1 フレームの特徴量を評価時に 1 秒の長さのセグメントにまとめてから認識を行う。後処理として、調理行動の遷移回数が与えられていると仮定し、認識結果の遷移回数が仮定した回数以下になるまで平滑化を行うことで、最終的な認識結果を得る。HMM を用いた手法より、NMF を用いた手法が優位であるという結果が報告されている。

3. 提案手法

3.1 レシピ情報を加味した調理行動認識

これまでの調理環境音を用いた調理行動認識の研究では

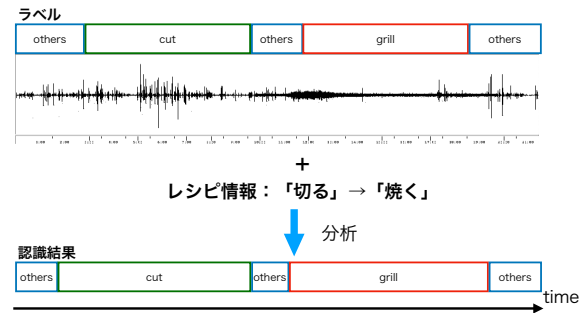


図 2 本研究の模式図

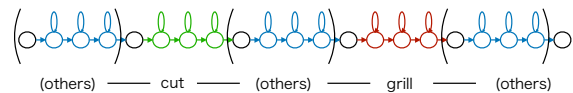


図 3 レシピ情報を加味した HMM ネットワーク

調理特有の時系列情報は基本的に考慮されていない。[5] による手法では、認識の後処理としてレシピを利用しているが、調理環境音を用いた認識の処理ではレシピを考慮せず認識を行っている。[6] による手法では、1 秒というセグメント単位で認識を行い、後処理に調理行動の遷移回数を利用しているが、レシピによる時系列情報は考慮されていない。そこで、本研究では、レシピによる時系列情報を予め考慮した状態で調理行動認識の検討を行う。焼きそばの調理行動認識を考えると、調理行動はレシピ情報により「切る」後に「焼く」と決定するため、本研究の模式図は図 2 のようになる。

音響特徴量生成モデルとして HMM を用いる。HMM を用いる利点として、レシピ情報を組み込むことが容易であることが挙げられる。調理行動を「切る」・「焼く」・「その他」、つまり cut・grill・others の 3 クラスとして認識を行う。各調理行動に対応する 3 状態 HMM を構築し、それらの HMM をレシピ情報に基づいて連結することで、レシピ情報を加味した HMM ネットワークを構築し認識を行う。レシピ情報を考慮すると HMM ネットワークは cut の後に grill というネットワークになり、各調理行動の前後に others の挿入・脱落が考えられるため図 3 のような HMM ネットワークとなる。

3.2 音響識別器を用いた音響特徴量の抽出

従来の研究では、特徴量に関する検討は MFCC や NMF という信号処理的な特徴量に留まっており、調理行動の認識という目的に適した特徴量について検討の余地がある。1 つの調理行動の継続時間を考慮すると、音声認識とは異なり、数秒の特徴を抽出した特徴量を用いることが求められる。そのため調理行動認識においては、MFCC の窓幅を大きくすることや、数秒の特徴を抽出できる特徴量を用いることが有効であると考えられる。そこで、本研究では、MFCC の他に、音響識別器から得られる中間特徴量を新た

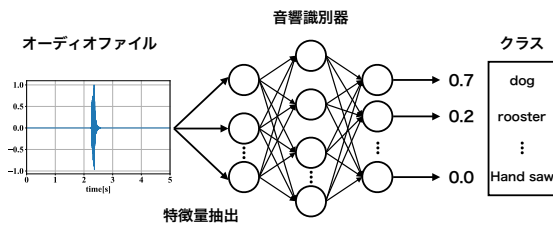


図 4 音響分類の模式図

表 1 中間特徴量の抽出に用いる音響識別器

Layer	kernel_size	stride	filter	Data Shape
Input				(1, 1, 24014)
conv1	(1, 8)	(1, 1)	40	
conv2	(1, 8)	(1, 1)	40	
pool2	(1, 160)	(1, 160)		(40, 1, 150)
trans				(1, 40, 150)
conv3	(8, 13)	(1, 1)	50	
pool3	(3, 3)	(3, 3)		(50, 11, 46)
conv4	(1, 5)	(1, 1)	50	
pool4	(1, 3)	(1, 3)		(50, 11, 14)
flatten				(50 × 11 × 14,)
fc5				(4096,)
bn				(N,)
fc6				(4096,)
fc7				(50,)

な特徴量として検討を行う。

特徴量の抽出に音響分類 (Audio Classification) に用いられる音響識別器を用いる。音響分類は、数秒から十数秒のオーディオファイルにどのような音響イベントが含まれるかを分類する研究である。本研究における音響識別器としては、図 4 のように、波形を入力しクラスの確率を出力する音響識別器を用いる。音響識別器として、[7] で提案された EnvNet-v2 のニューラルネットワークにボトルネック層を挿入した、表 1 のニューラルネットワークを用いる。trans 層はデータの転置を行う層で bn 層がボトルネック層である。このニューラルネットワークは約 1.5 秒の波形を入力としクラスの確率を出力するニューラルネットワークである。学習には ESC-50[8] のデータセットを用いる。学習後の音響識別器に調理環境音を入力した時のボトルネック層の出力を新たな特徴量として調理行動認識を行う。音響識別器の中間特徴量を用いることで、1.5 秒という調理行動認識に適した時間幅で、かつ、クラス識別に適した特徴量を得られることが期待できる。

4. 提案手法の検討

4.1 実験条件

本実験で用いたデータセットは 4 人の調理者による、焼きそばを調理した際の 6 つの調理環境音のデータである。4 人のうち 2 人が 2 回調理を行っている。なお、各音声ファイルは iPhone8 で録音されたサンプリング周波数が

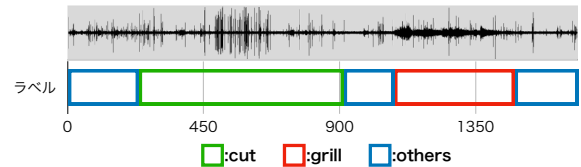


図 5 実験で用いたデータのラベル

16 kHz の wav ファイルである。録音した音声と調理者に対し聴取して得た調理手順をもとに図 5 のようにラベルを付した。cut と grill についてそれぞれ 1 度、開始時刻と終了時刻でラベルを付した。

モデルとして、図 3 の HMM ネットワークを用いて、Viterbi アルゴリズムによって調理行動認識を行った。特徴量として MFCC および音響識別器のボトルネック層の出力を用いた。MFCC は窓幅 16・160・1600ms の 3 条件で、MFCC13 次元とその $\Delta \cdot \Delta\Delta$ の合計 39 次元を特徴量として用いた。中間特徴量について、ボトルネック層の次元は 25・50・100 の 3 条件を用いた。HMM の出力は GMM に従うとして、混合数 1・2 の 2 条件を用いた。また、データ数が 6 つと小規模であるため、6-fold Cross Validation によって評価を行い、評価指標として F 値を用いた。

4.2 F 値の導出

あらゆる音響イベントを対象とした AED のタスクでは評価指標として F 値が用いられており [9]、本研究でも同様の F 値を用いる。AED における F 値はそれぞれのクラスに対して導出する。各クラスの F 値は次の手順で計算する。まず、一定のセグメント (本研究では 1 秒) ごとに、結果を以下の 4 つに分類する。

- (1) True Positive (TP): 認識結果が A で正解が A である
- (2) False Positive (FP): 認識結果が A で正解が A 以外である
- (3) True Negative (TN): 認識結果が A 以外で正解が A 以外である
- (4) False Negative (FN): 認識結果が A 以外で正解が A である

この 4 つに分類されたセグメントの数をそれぞれ、 $TP \cdot FP \cdot TN \cdot FN$ とする。適合率 P と再現率 R を以下のよう求める。適合率は認識結果が A のセグメントのうち正しく認識したセグメントの割合、再現率は A と認識すべきセグメントのうち正しく認識したセグメントの割合である。

$$P = \frac{TP}{TP + FP} \quad (1)$$

$$R = \frac{TP}{TP + FN} \quad (2)$$

クラス A に関する F 値は P と R を用いて以下のように導出される。

$$F = \frac{2 \cdot P \cdot R}{P + R} \quad (3)$$

導出される F 値は 0 から 1 の値であり，1 に近いほど性能が高いことを示す．本研究では認識対象である調理行動の cut・grill の F 値を用いて性能を検討する．

4.3 実験結果

調理行動認識の各条件における，全ての validation set の F 値の平均を表 2 に示す．MFCC による結果と中間特徴量による結果を比較すると，中間特徴量の導入による認識性能の向上が示唆される．特に grill についての認識は中間特徴量を用いることで明確に F 値が向上している．cut については中間特徴量の導入による向上は確認できない．しかし，cut と grill の F 値の平均値をモデルの性能とすると，中間特徴量・混合数 1・次元数 100 の条件が最も性能が高いことがわかる．

cut・grill の認識における F 値について，MFCC・混合数 2・窓幅 1600ms 以外の条件で，grill の方が F 値が高い．grill 音は定常的な音であるため，HMM で特徴を表現することが比較的有効であると考えられる．それに対して，cut はストロークの音が断続的であるため，HMM では表現が難しいと考えられる．HMM は図 3 のように状態が遷移するモデルであるため，cut のストロークが一定時間行われないと，状態が遷移して，others と認識されてしまうと考えられる．

GMM の混合数については，GMM の混合数の増加が性能の向上に寄与するとは言えない．これについては，GMM の混合数をあげることで cut・grill の HMM と others の HMM がより近い出力分布を持つ可能性が考えられる．others とラベルが付与された音には，食器を置く音・ビニール袋から食材を取り出す音・水の音など様々な音が含まれる．実際の波形を見ると，食器を置く音などは cut 音のパルス的な音に類似して，ビニール袋の音は grill 音のような定常的な音に類似している．混合数を増やすことで表現力の向上を図ると，others の確率分布が cut・grill に類似した音を出力する確率が上がる可能性がある．以上の議論より，others に様々な音が含まれていることが問題だと考えられる．よって，改善すべき点として，others をより細分化することがあげられる．others を細分化することで，cut・grill の誤認識が減少し認識精度が向上することが期待される．

5. 詳細なラベル付けによる提案手法の検討

5.1 実験条件

焼きそばを調理した際の調理環境音を新たに録音して実験を行った．データセットは 6 人の調理者による，焼きそばを調理した際の 6 つの調理環境音のデータであり，それ以外の録音条件は 4 章と同様である．「野菜を切る音」・「食

表 2 MFCC と中間特徴量を用いた認識における F 値

特徴量	混合数	窓幅 [ms]	次元数	cut	grill
MFCC	1	16	39	0.43	0.89
		160		0.38	0.82
		1600		0.57	0.73
	2	16		0.29	0.78
		160		0.22	0.60
		1600		0.69	0.55
中間特徴量	1	1500	25	0.58	0.93
			50	0.45	0.94
			100	0.68	0.93
	2		25	0.59	0.95
			50	0.29	0.91
			100	0.35	0.91

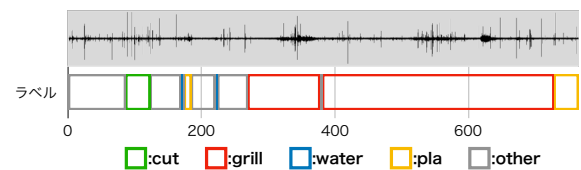


図 6 詳細なラベル付けによるデータのラベル

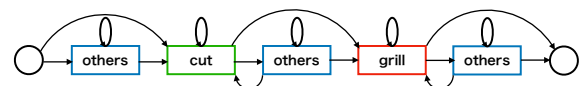


図 7 詳細なラベル付けによる HMM ネットワーク

材を焼く音」・「水が流れる音」・「ビニールが擦れる音」・「その他の音」の 5 つの音を，cut・grill・water・pla・other として，同時に録画した動画ファイルを確認しながらラベルを付した．また 4 章より詳細なラベルを付したことにより，図 6 のように，cut や grill の間に water・pla・other の挿入や，cut や grill の繰り返しの許したラベル付与となる．ラベルの細分化の有効性を検証するため，water・pla・other を others としてまとめた cut・grill・others の 3 クラスの coarse label による認識と，cut・grill・water・pla・other の 5 クラスの fine label による認識の 2 条件で実験を行った．

4 章において，MFCC と中間特徴量のそれぞれで，認識対象の調理行動である cut・grill の F 値の平均が最も高い条件であった，MFCC・窓幅 1600ms と中間特徴量・次元数 100 を用いて実験を行った．

4 章と同様に，調理行動に対応する HMM として，3 状態の HMM を用いて実験を行った．4 章に比べ詳細なラベルを付したことにより，レシピ HMM は cut の後に grill がくるとい順序は保ちつつ，water・pla・other の 0 回以上の挿入，cut・grill の 1 回以上の繰り返しが起こるネットワークを用いた．図 7 に HMM ネットワークの模式図を示す．others には water・pla・other の HMM が並列に接続されている．その他の実験条件については 4 章と同様の条件で行った．

表 3 fine label と coarse label を用いた認識における F 値

ラベル	特徴量	混合数	cut	grill
coarse	MFCC	1	0.52	0.81
		2	0.38	0.81
	中間特徴量	1	0.60	0.87
		2	0.50	0.86
fine	MFCC	1	0.50	0.82
		2	0.36	0.75
	中間特徴量	1	0.55	0.85
		2	0.62	0.90

表 4 fine label を用いた認識における混同行列
推定ラベル

		cut	grill	water	pla	other
正解ラベル	cut	581	0	3	0	156
	grill	0	2884	189	115	232
	water	14	4	104	69	123
	pla	11	0	26	56	134
	other	233	104	255	57	1393

5.2 実験結果

各条件における F 値を表 3 に示す．Fine label の場合と coarse label の場合の結果を比較すると，最も性能の高いモデルが，fine label・中間特徴量・混合数 2 の条件であることから，今回用いた fine label が調理行動認識において有効であることが示唆される．しかし，GMM の混合数を増加させることによる認識の性能の向上は確認できず，混合数と性能の関係は考慮の余地がある．

Fine label・中間特徴量・混合数 2 の全ての validation における，1 秒ごとの認識結果の混同行列を表 4 に示す．前述の調理行動認識の結果で示唆されたように，正解ラベルが cut における誤認識に other が多いことや，推定ラベルが cut における正解ラベルに other が多いことから，cut の特徴と other の特徴を分類することが困難であると考えられる．また，other と誤認識した結果が多いことから，fine label を用いた場合においても，other に他のクラスの認識に影響を与えるような環境音が含まれていると考えられる．本実験では，調理行動に対応する HMM を 3 状態 HMM としたことで，各行動が最短でも約 4.5 秒であることが前提になっている．そのためこのモデルでは，フライパンを置く音のような短時間の音響イベントをモデル化することができず，無音区間や短時間の音響イベントを other としてまとめて実験を行った．それにより，other に他のクラスの認識に影響を与える環境音が含まれている可能性がある．しかし，cut と grill の間には誤認識がない．これはレシピを考慮したことによる成果だと考えられる．レシピの時系列情報を考慮しない場合の例として，[6] による認識結果の一例を図 8 に示す．平滑化を行う前の音響情報による認識結果には cut と grill の間に誤認識が存在する．本研究の結果より，レシピ情報を考慮することは，cut と grill の誤認識の減少に有効であることが確認された．

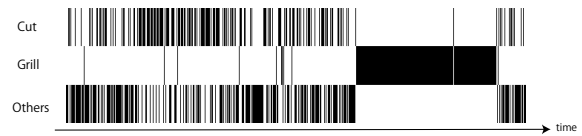


図 8 レシピを考慮しない場合の平滑化前の調理行動認識の結果（[6]より引用）

6. まとめ

調理を妨げない調理支援システムでは，システムが自動的に調理状況を判断する必要がある．本研究では，そのような調理支援システムを目標として，オフラインの調理行動認識に関する検討を行った．提案手法として，レシピを考慮した HMM ネットワークを構成し，音響識別器による中間特徴量を用いて検討を行った．また，調理行動以外のクラスに関する検討も行った．実験の結果から，食材を焼く音の認識においては MFCC と比較して中間特徴量が有効であることが確認され，調理行動の認識における性能の向上にも有効であることが示唆された．また，調理行動以外のクラスについてラベルを細分化することが調理行動認識において有効であることが示唆された．さらに，cut と grill の認識誤りがないことから，レシピを考慮したモデルが有効であることを確認した．しかし，GMM の混合数と認識性能の関係については考慮の余地があった．

今後の展望として，特徴量やモデルの改善，データセットの改善，実際の応用への改善の 3 つが挙げられる．本研究では約 1.5 秒の音響特徴量を用いて，1 つの行動に対する HMM を 3 状態 HMM として検討したことで，各行動が最短でも約 4.5 秒であることが前提となっている．そのため，このモデルではフライパンを置く音のような短時間の音響イベントをモデル化することができないので，無音区間や短時間の音響イベントを other としてまとめた．したがって，短時間の音響イベントが認識可能となるような特徴量やモデルを導入することで性能が向上すると考えられる．また，本研究でのデータセットは小規模であるため，大規模なデータセットを用いて，認識対象の調理行動のクラスを増やすことや，認識の頑健性を向上させることが求められる．さらに本研究では，予め録音された音源に対してオフラインで認識を行った．しかし，調理を妨げない調理支援システムの実現には，オンラインの調理行動認識が求められる．そのため，本研究の手法は直接利用することはできず，時間の長さが不定である調理環境音に対して調理行動認識が可能なモデルを検討することが必要となる．

謝辞 本研究は，JSPS 科研費 JP18K11369『音環境理解と対話管理技術に基づく大規模レシピデータを用いた調理支援に関する研究』の一部として行われた．

参考文献

- [1] Heittola, T., Mesaros, A., Eronen, A. and Virtanen, T.: Context-dependent sound event detection, *Eurasip Journal on Audio, Speech, and Music Processing*, Vol. 2013, No. 1, pp. 1–13 (2013).
- [2] Mesaros, A., Heittola, T., Dikmen, O. and Virtanen, T.: Sound event detection in real life recordings using coupled matrix factorization of spectral representations and class activity annotations, *IEEE International Conference on Acoustics, Speech and Signal Processing*, pp. 151–155 (2015).
- [3] Cakir, E., Heittola, T., Huttunen, H. and Virtanen, T.: Polyphonic sound event detection using multi label deep neural networks, *International Joint Conference on Neural Networks*, pp. 1–7 (2015).
- [4] Cakir, E., Parascandolo, G., Heittola, T., Huttunen, H. and Virtanen, T.: Convolutional Recurrent Neural Networks for Polyphonic Sound Event Detection, *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, Vol. 25, No. 6, pp. 1291–1303 (2017).
- [5] Kojima, R., Sugiyama, O. and Nakadai, K.: Scene Understanding Based on Sound and Text Information for a Cooking Support Robot, *Current Approaches in Applied Artificial Intelligence*, pp. 665–674 (2015).
- [6] Korematsu, Y., Saito, D. and Minematsu, N.: Cooking State Recognition based on Acoustic Event Detection, *The 11th Workshop on Multimedia for Cooking and Eating Activities* (2019).
- [7] Tokozume, Y., Ushiku, Y. and Harada, T.: Learning from Between-class Examples for Deep Sound Recognition, *The 6th International Conference on Learning Representations* (2018).
- [8] Piczak, K. J.: ESC: Dataset for Environmental Sound Classification, *The 23rd ACM International Conference on Multimedia*, pp. 1015–1018 (2015).
- [9] Mesaros, A., Heittola, T. and Virtanen, T.: Metrics for polyphonic sound event detection, *Applied Sciences*, Vol. 6, No. 6, pp. 162:1–17 (2016).