動画配信サービスのバリアフリー字幕化のための効果音認識
Sound effect recognition for barrier-free captioning of video distribution services

菊島 一樹
Kazuki Kikushima
法政大学情報科学部ディジタルメディア学科
kazuki.kikushima.3r@stu.hosei.ac.jp

abstract

In order to reduce the cost of barrier-free subtitles, we will work on sound effect sound event classification from program audio. In this study, in order to collect training data for classifying sound effects in the program sound, the sound effect part was extracted from the program sound, SIF (spectrogram image feature) was used as a feature, and CNN (convolution) was used as a classifier. Based on the sound effect event classification using neural networks), we will work on the use of power-related features in addition to the SIF of features. The feature quantities related to the power used in combination are power in the time domain and energy entropy. As a classification evaluation of a program audio classifier, the use of both time waveform power and SIF improved the overall recognition rate. Above all, the sound of opening the shutter door was improved by 20 points, and the thunder was improved by less than 50 points. Therefore, it can be said that power is effective as a feature amount of the sound effect, and a feature different from SIF can be extracted. In addition, the recognition rate of 38.2% was shown by applying only the target sound part of the program audio as the learning data to the classification model as the learning data and using it as the feature amount.

1 まえがき

動画に対する字幕付与は、会社の中、電車の中、図書館・博物館など、静かにしなければならない環境で音が出しにくいときや環境雑音の影響で音が聞き取りずらいときに役に立つ技術である。また、動画における単語がどのようなニュアンスで使用されているのか、どのような漢字で書かれるのかを知ることができるので、日本語の勉強がしたいときにも役に立つ。

字幕には、映画・ビデオ・DVD・テレビ番組などのセリフやナレーションの翻訳を字幕化した「翻訳字幕」と高齢者や聴覚に障害のある人のために、セリフや音など耳から得る情報を字幕にして補足する「バリアフリー字幕」というものが存在する。翻訳字幕とは、映像の出演者のセリフやナレーションを文字に起こしたものである。対して、バリアフリー字幕とは、セリフやナレーションに加えて、音楽・効果音の説明や話者表記などが文字に起こされている字幕のことである。[1, 2] また、バリアフリー字幕には、画面外で生じている音・状況を私たちの目に見えるように文字に起こされた場面が多々見られる。

つまり、バリアフリー字幕は、動画を視覚情報として観るだけでは、認識できない聴覚情報を視覚情報として得ることができるので、『視覚情報から得ることのできない情報を補足する』という役割をはたしている。バリアフリー字幕が施されることによって、より作品を深く楽しめることが可能になったのである。

しかし、バリアフリー字幕付与はすべての作品に付与できていないのが現状である。それは、バリアフリー字幕の需要に対し、供給が追い付いていないのが原因である。バリアフリー字幕化実装の現状は事前収録番組に限られている。また、動画に字幕を付与する手法として、番組の台本をベースに、動画の音声からセリフとして字幕付与する音・効果音・字幕付与しない音を分類し、字幕文字として認識する音声認識すること、認識結果である字幕文字から字幕の表示位置やタイミング、文字色など提示変換を施し字幕として表示することが必要である。このような作業がパソコンを用いて、すべて人の手によって施されているのでとても人の手が足らず、供給が追い付かないのである。

図 1. バリアフリー字幕の例

本研究では、この問題を解決し、バリアフリー字幕化にかかるコストの削減を目指す足がかりとして、番組音声からの効果音サウンドイベント分類に取り組む。提案するシステムの大まかな流れは、まず番組音声から効果音部分を検出し、音声データとして抜き出すことで学習データを収集する。次に、集めたデータの様々な特徴量を用いて各効果音サウンドイベントに分類を行う。このようにして、番組音声からの効果音サウンドイベント分類に取り組む。

2 関連研究

従来から、効果音の認識は難しい問題とされている。その理由としてまず効果音は人の声と比べて突発的な音であることがあげられる。つまり、人の声よりも短い範囲に特徴が現れたり、非定常な特徴であるために認識が難しいのである。また、目的の効果音のなっている環境下では、目的音以外の音が混じってしまっていることも認識が難しい理由の一つである。

そこで、本研究では突発的な効果音の特徴を抽出するために、12 種類の特徴ベクトルを用いて暴力的な映画から確率推論モデルを用いて音楽、セリフ、銃声を分類する研究 [9] を参考にしようと考えている。ここで取り上げられているパワーに関する特徴量であるエネルギーエントロピーと、時間波形のパワーは突発的な音の認識に有用であると考えるので、特徴量の一つとして用いる。

また、目的の効果音のなっている環境下では、目的音以外の音が混じってしまっている問題に取り組んでいる従来研究として、SNR0dB の環境ノイズを付与した RWCP-DB の環境音 50 ラベルの分類を行っている研究 [3] が存在し、80% 弱の認識率を誇っている。この研究は、特徴量として平滑化・低周波

Supervisor: Prof. Katunobu Itou

成分除去を施したスペクトログラム画像特徴（SIF）を分類器としては隠れ層 2 層からなる畳み込みニューラルネットワーク (CNN) を用いていて、メル周波数ケプストラム係数（MFCC) を用いた隠れマルコフモデル（HMM）やサポートベクターマシーン（SVM)、40 次のメルフィルタバンク（MelFb) を用いた CNN よりも雑音成分が大きい状況下（SNR0dB/10dB）での認識で、より良い性能を示している。この研究から、特徴量の SIF、分類器 CNN は目的音以外の音が混じった混合音に対する分類に強いことがいえるので、番組音声の効果音認識に有用だと考える。しかし今回、入力として用いる番組音声に含まれる目的音以外の音が従来研究の環境ノイズ 0dB よりも大きな音になるのでより認識の難しさが高いことが考えられるので、モデルに改良を加えることを試みる。

また、複数のニューラルネットワークを用いて環境音を単音/長く続く音/繰り返しのある音に分類する分類部と分類した音声を特定のラベルに認識する識別部を用いて環境音を認識する研究 [4]、2 つの CNN の出力結果を DS 理論を用いて最終的な分類結果を示す研究 [5] がある。この研究から、番組音声を効果音/BGM/セリフ/無音に分類する分類部と分類した音声を特定のラベルに認識する識別部として CNN を用いるは番組音声の効果音認識に有用だと考える。

他にもマルチメディアイベントの検出を行っている研究として、ディープニューラルネットワーク（DNN）を用いた研究 [6]・リカレントニューラルネットワーク（RNN）を用いた研究 [7]・畳み込みニューラルネットワーク（CNN）と RNN の併用を用いている研究 [8] など多くの従来研究がみられた。

3　システムの構成・処理方法

まず、番組音声での効果音分類を行うための学習データを収集するために、番組音声から効果音部分の抜き出しを行った。処理手順としては、まず、バリアフリー字幕化が施されている番組動画から字幕ファイルを抽出した。抽出した字幕ファイルには表示されている字幕のラベル・表示開始時間・終了時間などの情報を用いて、機械学習を行う上で必要になる学習データを作成する。

次に、番組音声の効果音サウンドイベント分類を行うにおいて前述で用意した音声データを学習データに対して、特徴量として平滑化とノイズ除去が施されたスペクトログラム画像特徴（SIF）を抽出し、分類器として畳み込みニューラルネットワーク（CNN）を利用したものをベースとして用いて効果音サウンドイベント分類に取り組んだ。構築したモデルは、特徴量として SIF（52 × 40)、分類器として CNN を用いたモデル [3]・特徴量として SIF（52 × 40）とパワーに関する特徴量を併用し、分類器として CNN とディープニューラルネットワーク（DNN）を併用したものを用いたモデルの 2 つのモデルを構成した。それぞれのモデルを用いて効果音 11 種類の分類を行った。

3.1　スペクトル画像特徴（SIF）

3.2　パワーに関する特徴量

パワーに関する特徴量として挙げられる、エネルギーエントロピーとパワーが突発的な音の認識に有用であると考える。

3.2.1　エネルギーエントロピー

エネルギーエントロピーは音声信号のエネルギーレベルの急激な変化を測る特徴量で

SIF とは平滑化/ダウンサンプリング・ノイズ除去を施したスペクトログラム画像特徴である。具体的な抽出手法を下記に示す。

入力される環境音をサンプリング周波数 16kHz、FFT 解析窓長 1024、オーバーラップ 64 点として離散フーリエ変換（FFT）を施すことで、1 フレーム 64ms、1 フレーム間の時間差 4ms の短時間スペクトル表現を得る。次に周波数方向に平均化することでダウンサンプリングを施す。（B = 52 点（最適））次は、低周波成分の雑音を除去するために任意のフレームの最小周波数成分を全体のスペクトルから減算する。このとき、平滑化された l 番目のフレーム $f_b(l、k)$ とする；

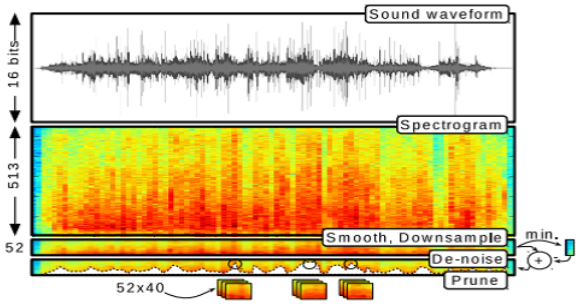

図 2. SIF 抽出の手順

$$f_{dn}(l,b) = f_b(l,b) - min f_b(l,b) \qquad for \quad b = 1...B$$

雑音除去が施されたスペクトログラムの各フレームごとのエネルギーを計算する；

$$e(l) = \sum_{b=1}^{B} f_{dn}(l,b)$$

計算された時間領域のエネルギーのピークを中心に時間成分 L（= 40（最適））×周波数成分 B が SIF である。本研究では従来法と同様に最大の 3 つのピークの前 2 フレーム～後 3 フレームの計 18 フレームの SIF を特徴量とする。また、SIF を抽出するとき、エネルギーのピークから前後のフレームがとってこれないという問題を解決するために、下記の処理の前処理として入力音声の前後に番組音声の無音成分を $SNR := 20dB$ で前 100 フレーム・後 140 フレーム分付け加えている [10]。エネルギーエントロピーの抽出手法は、各フレーム（本研究では 1024 点）をさらに固定長の K 個のサブフレームに分割し、それぞれのサブフレームにおいて正規化されたエネルギー σ^2 を計算する。ここにおける正規化されたエネルギーはサブフレームのエネルギーをフレーム全体のエネルギーで割ったものである。そのあと、エネルギーエントロピーは次の式で計算されます。

$$H = -\sum_{i=0}^{K-1} \sigma^2 * \log 2(\sigma^2)$$

本研究では、フレーム長が 1024 点であるのでサブフレームは 128 点を固定長として 8 個のサブフレームを形成している。

3.2.2　パワー

本研究では、分析フレーム (1024 点) ごとの時間領域の平均パワーを求めている。

$$\mathrm{P} = 1 \frac{}{\sum_{m=0}^{1023} 1024 y_m^2} \tag{1}$$

3.3　分類器

3.3.1　畳み込みニューラルネットワーク（CNN）

CNN とは深層学習の一つでありニューラルネットワークに「畳み込み」という操作を導入したものであり、『入力層』・『出力層』・隠れ層に当たる『畳み込み層』と『プーリング層』、『全結合層』から構成される。畳み込み（convolution）とは、画像処理でよく利用される手法で、カーネル（またはフィルター）と呼ばれる格子状の数値データと、カーネルと同サイズの部分画像（ウィンドウと呼ぶ）の数値データについて、要素ごとの積の和を計算することで、1 つの数値に変換する処理のことである。この変換処理を、ウィンドウを少しずつずらして処理を行うことで、小さい格子状の数値データに変換する。つまり、入力画像のどの個所に特徴があるのかを計算するという役割を持つ。プーリングとは、大きな画像から重要な情報は残しつつ縮小する役割をになっている処理である。ウィンドウ中の最大値を選択する最大値プーリングや、ウィンドウ中の平均値を選択する平均値プーリングなどがある。本研究では最大値プーリングを用いている。全結合層とは、CNN では出力の手前で使用されることが多く、畳み込み層とプーリング層によって検出

された特徴の組み合わせから、予測結果に分類するための識別部になる。また、全結合層は 1 次元のデータを入力として 1 次元が出力される。

CNN のモデルの構成・流れは、SIF が入力として入力層に与えられる。次に、2 層の隠れ層を用いて特徴が検出する。順に、カーネルサイズ 5 × 5 の 6 つのカーネルを持つ畳み込み層とカーネルサイズ 2 × 2、スライド幅 2 × 2 のプーリング層から成る隠れ層、カーネルサイズ 5 × 5 の 12 つのカーネルを持つ畳み込み層とカーネルサイズ 2 × 2、スライド幅 2 × 2 のプーリング層から成る隠れ層の 2 層となる。隠れ層による検出の結果 12 個の小さい格子状のデータ（入力の大きさが 40 × 52 なので、7 × 10 の大きさになる）が得られる。次に、検出された特徴を組み合わせ、分類するために全結合層を用いる。全結合層の入力は 1 次元でないといけないので、検出された特徴（7 × 10 × 12 個＝ 840 個）を一次元配列に変換することが必要である。また、全結合層の出力数は最終的に分類したい効果音の個数になるので、11 個の出力数になっている。

3.3.2 ディープニューラルネットワーク（DNN)

DNN とは、人間が自然に行うタスクをコンピューターに学習させる機械学習の手法の一つであり、分類を行う際に、データに法則性を見つけ、関数近似を行なっている。また、DNN はニューラルネットワークの隠れ層を多層にして用いているので、より複雑な関数近似を行う事ができるようにしたものである。ニューラルネットワークとは、人間の脳内にある神経細胞（ニューロン）とそのつながり、つまり神経回路網を人工ニューロンという数式的なモデルで表現したものである。また、ニューラルネットワークは、入力層、出力層、隠れ層から構成され、層と層の間には、ニューロン同士のつながりの強さを示す重み「W」が存在し、全結合が用いられている。ニューロン同士を組み合わせることで複雑な関数近似を行っている。

DNN のモデルの構成・流れは、時間成分 100 点分のフレームが入力として入力層に与えられる。今回は入力音声の最もパワーが大きな点を中心に 100 フレーム分とする。次に、ニューロン数 2000 個の 3 層の隠れ層を通して関数の近似を行い、出力数は最終的に分類したい効果音の個数になるので、出力として 11 個の出力数を得る。

3.3.3 CNN と DNN の併用モデル

CNN と DNN の併用したモデルの構成・流れは、SIF とパワーに関する特徴量の時間成分 100 点分のフレームの 2 つの入力を用いている。まず、スペクトログラム特徴を CNN モデルの 2 つの隠れ層・全結合層を用いて特徴検出を行い、840 個の一次元配列の特徴量を得る。ここで得た 840 個の特徴量と 100 個のパワーに関する特徴量を結合することで新しく一次元配列を作成する。作成した一次元配列を DNN モデルの入力として用いることで、出力として 11 個の出力数を得る。

4 評価

本研究では、番組音声の認識に有用だと考えたパワーに関する特徴量の評価実験と従来研究のモデルである CNN をベースに構築したモデルの評価実験を行った。

まず、作成した番組音声のデータセットのラベルとデータ量をグラフ（表 1）に示す。各効果音ラベルのデータセットから 5 個の音声データをテストデータとし、残りを訓練データとして用いる。

4.1 パワーに関する特徴量の評価実験

併用に用いたパワーに関する評価実験として、RWCP-DB のクリーンな環境音 50 ラベルの音声データと番組から抽出したの音声データの 2 つのデータセットを用いてそれぞれ場合において、評価を行った。

入力としては音声の一番パワーの大きなフレーム中心に 100 フレーム分を特徴量として抽出し、上記で構成された DNN を用いて分類評価を行った。

4.1.1 RWCP-DB のクリーンなデータセットを用いた分類評価

パワーに対する分類評価としては、全体での分類率が 30% ほどであった。また、各効果音ラベルの混合行列を見てみると、衝撃音のような単発の音・長く続く音に対して、パワーの違いによって細かい分類ができている結果を示しました。

表 1. データセット

効果音ラベル	ファイル数
電話	79
ノック	60
拍手	50
チャイム	34
ドアの開く音	33
シャッター	32
戸の開く音	31
通知音	18
雷鳴	12
物音	11
風鈴の音	10

エネルギーエントロピーに対する分類評価としては、全体での分類率が 10% ほどであった。また、各効果音ラベルの混合行列を見てみると、単発の音とそれ以外の音のグループに大まかな分類はできている結果を示しました。エネルギーエントロピーに関しては本研究の特徴量としてあまり効果がみられないと考察する。

4.1.2 番組音声を用いた分類評価

各特徴量の分類結果を結果の良かったものを抜粋して表 2 に示す。

番組音声の分類結果としては、パワーを特徴量として用いた時には、電話・ノック・シャッター・戸の開く音に対して比較的良い性能を示した。また、エネルギーエントロピーを特徴量として用いた時には、ノックに対してよい性能を示した。しかし、他の効果音ラベルの認識の精度はほとんど認識できていなかった。認識率が悪かった原因としては、番組音声に含まれる bgm やセリフが入力として用いている 100 フレームに抽出されてしまい、効果音が鳴っている部分が入っていない可能性があるということ。また、番組音声は常に何かしらの音（BGM・セリフなど）がなっている状況が多いためにエネルギーの変化での判断が難しいことが考えられる。

4.2 分類器の評価実験

分類器の評価実験としては番組音声を入力としたときの CNN モデルと CNN と DNN を併用したモデルの分類性能を比較する。また、従来研究である RWCP-DB を入力としたときの性能と比べることで評価を行う。

番組音声を入力としたときの CNN モデルと CNN と DNN を併用したモデルの分類性能を比較を行った結果を表 3 に示す。ここで SIF_power は SIF とパワーを併用した特徴量のことを指し、SIF_entropy は SIF とエネルギーエントロピーを併用した特徴量のことを指している。

分類結果から、パワーを併用した場合、全体的に認識率の向上がみられた。中でも、シャッター・ドアの開く音に対しては 20 ポイントの向上がみられ、雷鳴に至っては 50 ポイント弱の向上がみられた。よって、パワーは効果音の特徴量として有効であり、SIF とは異なる特徴を抽出できていることが確認できた。エネルギーエントロピーを併用した場合、拍手に対して 15 ポイントの向上がみられたが、全体的な認識率の向上はみられなかった。これは、入力の番組音声には常に何かしらの音成分が含まれているために目的音のエネルギー変化をうまく抽出できないことがあげられる。よって、エネルギーエントロピーは本研究の特徴量に適さないといえる。

また、従来研究である RWCP-DB を入力としたときの性能と番組音声を入力としたときの性能を比べた結果、RWCP(従来研究)：76.8% 番組音声：14.5% という認識結果が得られた。この結果より、これだけ認識率に大きな差が生まれてしまっているのは、番組音声を入力としたとき、目的音部分の SIF がとれていないことが原因であるのではないかと考え、データセッ

表 2. 番組音声を用いた分類結果（%）

特徴量	電話 (%)	ノック (%)	シャッター (%)	戸の開く音 (%)
パワー	40	40	53.3	40
エネルギーエントロピー	6.7	80	13.3	0

表 3. 分類器の分類結果（%）

	SIF	SIF_power	SIF_entropy
Open-set	93.5	98.2	83.9
Close-set	14.5	22.4	14.5
シャッター	20	40	6.7
チャイム	13.3	20	6.7
ドアの開く音	0	20	0
ノック	46,7	46.7	40
戸の開く音	13.3	13.3	20
通知音	0	0	0
電話	33.3	13.3	33.3
拍手	26.7	26.7	40
風鈴の音	0	0	0
物音	0	13.3	0
雷鳴	6.7	53.3	13.3

トの音声データから目的音が鳴っている部分のみを切り出し、それを学習データとすることで認識率の向上がみられるのか評価実験を試みた。結果としては表 4 に示したとおりになった。20 ポイント以上認識率の向上がみられたラベルが複数存在することから、切り取り前の番組音声から目的音部分のフレームの SIF がうまく選択できていないことが確認することができた。また、全体の認識率の割合のから 50% の確率で目的音部分のフレームの SIF を選択することに失敗している。

表 4. 分類結果（%）

	前	後
Open-set	93.5	100
Close-set	14.5	27.9
シャッター	20	26.7
チャイム	13.3	46.7
ドアの開く音	0	20
ノック	46,7	100
戸の開く音	13.3	0
通知音	0	0
電話	33.3	60
拍手	26.7	46.7
風鈴の音	0	0
物音	0	0
雷鳴	6.7	6.7

5 あとがき

本研究では番組音声をもちいた効果音サウンドイベント分類について取り組んだ。番組音声を入力として従来研究のモデルにかけることで 14.5% の認識率を示した。また、番組音声の目的音部分のみを切り取ったものを学習データとして SIF とパワーの併用したもの特徴量として分類モデルにかけることによって 38.2% の認識率に向上した。認識率が向上した要因としてはパワーの併用・目的音部分のフレームの SIF を選択できるように学習データを切り取ることがあげられる。

パワーを併用した場合、全体的に認識率の向上がみられた。中でも、シャッター音・ドアの開く音に対しては 20 ポイントの向上がみられ、雷鳴に至っては 50 ポイント弱の向上がみられた。よって、パワーは効果音の特徴量として有効であり、SIF とは異なる特徴を抽出できていることが確認できた。エネルギーエントロピーを併用した場合、拍手に対して 15 ポイントの向上がみられたが、全体的な認識率の向上はみられなかった。これは、入力の番組音声には常に何かしらの音成分（BGM・セリフなど）が含まれているために目的音のエネルギー変化をうまく抽出できないことがあげられる。よって、エネルギーエントロピーは本研究の特徴量に適さないといえる。

次に、データセットの音声データから目的音が鳴っている部分のみを切り出し、それを学習データとすることで認識率の向上がみられるのか評価実験を試みた結果を表 4 に示した。20 ポイント以上認識率の向上がみられたラベルが複数存在することから、切り取り前の番組音声から目的音部分のフレームの SIF がうまく選択できていないことが確認することができた。また、全体の認識率の割合のから 50% の確率で目的音部分のフレームの SIF を選択することに失敗している。

番組音声の目的音部分のみを切り取ったものを学習データとして用いた時の認識率が 27.9% とやはり従来研究よりも劣ってしまっている。これは、目的音が他の音と混じっていることが原因の一つに挙げることができる。なので、学習に用いているデータを目的音だけのクリーンなデータに近づけることができれば認識率の向上が見込めると考える。

参考文献

[1] 坂井他. 放送における視聴覚障害者向け情報バリアフリー技術. 映像情報メディア学会誌.2010;64:940-944.

[2] 今井亨：リアルタイム字幕放送のための音声認識、信学技報、SP2009-52, WIT2009-58 (2009).

[3] Zhang, H., et.al. Robust sound event recognition using convolutional neural networks. In Acoustics, Speech and Signal Processing (ICASSP), 2015 IEEE International Conference on (pp. 559-563). IEEE

[4] TOYODA, Y., et.al. 2004. Environmental sound recognition by multilayered neural networks. In Proceedings of the 4th International Conference on Computer and Information Technology (CIT ’ 04). 123-127

[5] Y. Su, et.al, “Environment sound classification using a two-stream cnn based on decision-level fusion,” Sensors, vol. 19, no. 7, p. 1733, 2019.

[6] K. Ashraf, et.al. Audio-based multimedia event detection with DNNs and sparse sampling. In Proceedings of the 5th ACM on International Conference on Multimedia Retrieval, pages 611?614. ACM, 2015.

[7] Y. Wang, L. et.al, “Audio-based multimedia event detection using deep recurrent neural networks,” in Proc. 2016 IEEE Int. Conf. Acoust., Speech, Signal Process., 2016, pp. 2742?2746.

[8] E. Cakir, et.al, “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.

[9] Giannakopoulos, T., et.al.: A multi-class audio classification method with respect to violent content in movies, using bayesian networks. In: IEEE International Workshop on Multimedia Signal Processing, MMSP 2007

[10] Nomura, Y., et.al, Speech enhancement based on the dominant classification between speech and noise using feature data in spectrogram of observation signal, Transactions of the Japan Society of Mechanical Engineers, Series C, Transactions of The Institute of Electrical Engineers of Japan , Series C, Vol.124, No.11(2004),pp.2310-2319(in Japanese).