CG-002

# 心音に対する説明文の自動生成
# Automatic Captioning for Cardiac Sounds

柏野 邦夫† 中野 允裕† 渋江 遼平† 塚田 信吾† 友池 仁暢†,‡
Kunio Kashino Masahiro Nakano Ryohei Shibue Shingo Tsukada Hitonobu Tomoike

## 1. はじめに

生体音には身体の機能や状態に関する様々な情報が含まれている。このため、生体音を取得する装置、取得した音響信号の解析や診断支援技術等に関して数多くの研究開発が行われており、その活用への期待が高まっている [1,2]。

生体音の取得は特殊な検査装置によることなくマイクロホンで行うことができるため [3]、医療機関等に限らず日常生活の中でも手軽に行える可能性がある。その一方で、取得した音を聴取しての診断、即ち聴診は、経験を積んだ医療者が個別に行う必要がある。我々はこのギャップを埋めるためのアプローチを試みている。本稿では、特に心音を題材として、とらえた心音がどのような音であるかを言葉で表現する手法について検討する。日常生活の中でとらえた心音に対して、本人をはじめ、専門知識を持たない非医療者にも分かりやすい言葉で説明を自動的に行うことができれば、必要な場合に早期に医療機関の受診を勧奨することなども可能になると考えられる。

近年のメディア情報を対象とする認識技術の進歩の中で、例えば環境音を対象として、何の音であるかを認識する音響イベント検出 [4]、どのような場面かを推定する音響シーン分析 [5]、故障や事件などといった異常を監視するための異常音検出 [6]などの研究が行われている [7]。これらはいずれも音響信号を入力とするクラス分類の問題である。心音に関しても、クラス分類、つまり正常・異常の判定や病名の推定などが重要な課題であることは言うまでもない [8]。そこで本稿では、クラス分類と説明文の生成問題とをマルチタスクの問題として同時に扱う方法を提案する。

画像、動画、音響信号などのメディア情報に対する説明文生成は近年研究が盛んな課題である [9,10,11]。一般に、クラス分類に対比した場合の説明文生成のメリットは、単に含まれる対象が何であるかだけではなく、複数の対象同士の関係性なども含めた高い記述力を持ち得る点や、記述の目的や用途に応じて、有用性の高い柔軟な表現が可能である点などが挙げられる。特に本稿では、詳細度と呼ぶパラメータを説明文生成の際の制御入力とすることにより、生成される説明文の長さや詳しさを調整できることを示す [12]。

## 2. システム構成

図 1 に示すように、提案する説明文生成システムは、音響信号の特徴系列を単語の系列に変換する系列変換モデルに基づくことを基本とし、複雑になることを避け、少数のシンプルな要素で構成することを意図したものである。システムの入力は心音の音響信号と詳細度と呼ぶパラメータ値であり、出力はクラス分類結果と説明文である。入力音響信号は SoundNet [13] の学習済モデル [14] を用いて音響特徴時系列に変換する。これを Bidirectional LSTM により潜在空間に射影し潜在特徴とする。この潜在特徴から、クラス分類と説明文のデコードのマルチタスクでの出力生成を行う。説明文のデコードは 3 層の Layered LSTM で行い，クラス分類は全結合型ニューラルネットワークで行う。

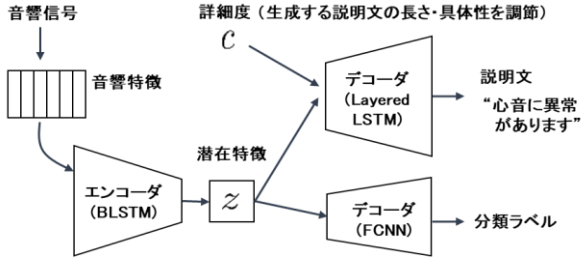


**図 1 提案システムの構成**

### 2.1 クラス分類

本稿では、 (1) 概要 (2) 音の特徴 (3) 病態の 3 つの観点からクラスを設定した。主なものを表 1 に記す。

### 2.2 説明文生成

音に対する説明文生成における大きな問題点の一つは、記述の自由度が一般に非常に大きく、音だけからは、様々な説明のうちのどれが正しくどれが正しくないということを決められない点である。例えば、ある心音に対して「異常あり」とだけ指摘することも出来るし、どのような異常があってどのような病気が考えられるかを詳細に述べることも出来る。無数に可能な説明のうち、どれが最も適切か

**表 1 クラス分類におけるクラスの設定**

| 大分類 | カテゴリ | クラス | クラス数 |
|---|---|---|---|
| 概要 | 正常性 | 正常、異常 | 2 |
| | 病態種類 | 正常、弁膜症、先天性心奇形、その他の病態 | 4 |
| 音の特徴 | I音 | 正常、亢進、減弱、分裂 | 4 |
| | II音 | 正常、分裂、A亢進、A減弱、P亢進、P減弱 | 6 |
| | 雑音 | なし、収縮期、拡張期、連続性 | 4 |
| 病態 | 僧帽弁閉鎖不全 | あり、なし | 2 |
| | 僧帽弁狭窄 | あり、なし | 2 |
| | 大動脈弁閉鎖不全 | あり、なし | 2 |
| | 大動脈弁狭窄 | あり、なし | 2 |
| | 心房中隔欠損 | あり、なし | 2 |
| | 心室中隔欠損 | あり、なし | 2 |
| | 動脈管開存 | あり、なし | 2 |
| | 肺動脈弁狭窄 | あり、なし | 2 |
| | 心筋症 | あり、なし | 2 |
| | 心房細動 | あり、なし | 2 |
| | 期外収縮 | あり、なし | 2 |
| | 虚血性心疾患 | あり、なし | 2 |

† 日本電信電話株式会社 バイオメディカル情報科学研究センタ NTT Biomedical Informatics Research Center
‡ NTT Research, Inc.

は目的や場合によって異なり、一概に正誤や適切さを評価することは困難である。

この問題に対処するため、本稿では詳細度と呼ぶパラメータを導入する。詳細度とは、説明文に含まれる単語毎に計算された情報量を文全体で足し合わせたものであり、

$$I_s = \sum_{t=1}^{n} (-\log p_{w_t}) \quad (1)$$

で定義される。ここで $p_{w_t}$ は説明文中の $t$ 番目の単語の出現確率、$n$ は説明文中の単語数である。ありふれていない、つまり具体性の高い単語を数多く用いるほど、詳細度の値は大きくなる。

提案システムにおいては、音響特徴から得られる潜在特徴に加え、詳細度の目標値 $c$ を補助入力として、条件付のデコードを行うことにより説明文生成を行う。これにより、短く端的な説明が望ましい場合は $c$ を小さく、長く詳細な説明が望ましい場合は $c$ を大きくすることで、目的に適った説明文となるよう制御することが可能となる。

## 3. 学習

このようにして構成されたニューラルネットワークの学習は、学習用データとして、音響信号と、クラスラベルと、説明文とが組になったものを多数用いて、誤差逆伝搬法により行うことが出来る。誤差は、クラス分類に関する誤差と、説明文に関する誤差と、詳細度に関する誤差の重み付き和とする。

クラス分類に関する誤差 $Lc$ は、表 1 におけるカテゴリ毎に Softmax 関数による正規化を行い、クロスエントロピー誤差のカテゴリについての総和をとることにより算出する。

説明文に関する誤差 $L_d$ は、学習データに含まれる説明文中の各単語を教師ラベルとし、音響信号の入力に対する説明文デコーダの各ステップにおける出力層についてクロスエントロピー誤差を計算し、その総和をとることにより算出する。この時、教師ラベルは one-hot ベクトルで表されるため、

$$L_d = \sum_{t=1}^{n} (-\log q_{w_t}) \quad (2)$$

となる。ここで $q_{w_t}$ は、$t$ 番目のステップにおいて、教師ラベルと一致する単語について出力された確率値である。

詳細度に関する誤差 $Ls$ を求めるためには、まず生成文の詳細度の推定値を次式により求める。

$$\hat{I}_s = \sum_{t=1}^{n} \sum_{w} p_{w_t} I_w \quad (3)$$

(3) 式において、$w$ に関する総和は、$t$ 番目のステップにおいて選択される単語 $w_t$ のもつ情報量の期待値を、出力層における各単語の確率値と、その単語のもつ情報量との積和で計算することを表している。次に、

$$L_s = (\hat{I}_s - c)^2 \quad (4)$$

により $Ls$ を求める。

なお、説明文に関する過学習や詳細度への過剰適応を抑制するための経験則として、2 段階で学習を進めることが効果的である。1 段階めでは、学習データセットを用いて通常の学習を行う。この時 $c$ の値は、教師となる説明文から計算される。2 段階めでは、詳細度の再現を目的として学習を行う。つまり、潜在空間内の点をサンプリングし（実際の学習データの中から任意に 1 つを選び、その時点でのエンコーダを用いて射影してもよい）、これとランダムに選択した $c$ の値とを入力とする。この時説明文の教師データとするのは、教師データ中で $c$ に最も近い詳細度をもつ説明文である。この操作は、対応する説明文が教師データ中に存在していない潜在変数に対しても、詳細度を反映した説明文生成が行えるようにデコーダの学習を誘導することを意図している。このため、2 段階めの学習においては、説明文用のデコーダのみを更新し、エンコーダとクラス分類用のネットワークの更新は行わない。

## 4. 実験

### 4.1 データ

評価実験のための心音の音響信号として、市販の参考書 [15] に付属する CD 音源を用いた。この CD には、55 の症例について、1 症例あたり 1 トラックずつ、合計約 65 分の聴診音の音響信号が収められている。音響信号の一部には解説音声が重畳されている。本実験では、これら 55 件の音響信号に対して、8-fold のクロスバリデーションを行うこととし、以下の 2 条件でデータを準備した。

(条件 1) 拍同期切り出し
人手によりマーキングした拍位置を基準とし、かつ解説音声の重畳区間は除外して 1 拍ごとに 4 拍分を切り出す。

(条件 2) 定周期切り出し
音響信号全体にわたって、6 秒の時間窓を 3 秒ごとにずらしながら、音声重畳部分も含む全区間の音響信号を切り出す。

このようにして切り出された音響信号の総数は、
・拍同期切り出し 1,403 個
・定周期切り出し 1,302 個
であった。

これらの音響信号に対し、人手でクラスラベルと説明文を与えた。クラスラベルは、表 1 に例示したカテゴリのそれぞれについて、同書の本文の記述に即して与えた。説明文はクラスラベルを参照し、切り出す前の音響信号を単位として、音響信号 1 件（症例 1 件）につき、説明文を 7 件ずつ与えた。これにより、用意したデータの総数は、
・拍同期切り出し 9,821 組
・定周期切り出し 9,114 組
となった。

### 4.2 学習

表 2 に、本実験におけるパラメータを列挙する。

**表 2　実験におけるパラメータ**

| | |
|---|---|
| 説明文デコーダ レイヤ数 | 3 |
| デコーダ LSTM セル数 | 328 |
| 潜在空間 次元数 | 4096 |
| 語彙数 | 82 |
| バッチサイズ | 64 |
| エポック数 | 100 |
| 最適化 | Adam |

### 4.3 方法

本実験では、クロスバリデーションにより、クラス分類の精度、説明文生成における詳細度制御の動作、及び説明文生成の動作確認を目的とする。このため、テストデータとするデータに対してペアリングされているクラス、説明文、及び説明文から生成される詳細度、をそれぞれ正解として評価を行うこととした。このため、テストにおいて補助入力とする詳細度は、正解の詳細度（7 通り）とした。

### 4.4 結果

#### 4.4.1 クラス分類

表 1 に挙げたクラスに対応する分類精度を表 3 に示す。この値は、教師ラベルと分類結果とが一致したデータの、全数に対する割合である。全般に拍同期切り出しの方が高精度であり、異常有無の 2 値分類では 99.7%、病態に関する 2 値分類では 97% 以上の精度となっている。定周期切り出しでは、拍同期切り出しと同等の場合もあるが、数ポイント程度精度が低下する場合が見られる。特に音の特徴のカテゴリでは精度低下が大きく、およそ 5〜10 ポイントの差となっている。これは重畳された解説音声など心音以外の音の音量が大きい部分を除外することなく、機械的に切り出したことによる影響と考えられる。

音の特徴に関するクラスの混同行列を図 2〜4 に示す。図 2 では、定周期切り出しの場合に、I 音が正常なデータを亢進あるいは減弱と誤る場合や、I 音が減弱しているデータを正常あるいは亢進と誤る場合が多いことが分かる。図 3 の II 音の場合にも I 音と概ね同様の傾向が観察できる。図 4 では、定周期切り出しの場合に、心雑音なしとされるべきデータを雑音ありとする誤りや、その逆の誤りが生じる割合が、拍同期切り出しの場合に比べ多くなっている。

心音の取得では様々な雑音が重畳しやすい。上記を踏まえ、もしロバストな拍時刻検出が可能であれば、拍同期切り出しの方が定周期切り出しより高精度を期待できると考えられる。一方で、異常有無や病態の 2 値分類に限れば、精度の違いは本実験では数ポイント程度に収まった。

**表 3 クラス分類精度**

| 大分類 | カテゴリ | 拍同期切り出し | 定周期切り出し |
|---|---|---|---|
| 概要 | 正常性 | 0.997 | 0.979 |
| | 病態種類 | 0.965 | 0.916 |
| 音の特徴 | I音 | 0.984 | 0.887 |
| | II音 | 0.946 | 0.898 |
| | 雑音 | 0.948 | 0.902 |
| 病態 | 僧帽弁閉鎖不全 | 1.000 | 0.980 |
| | 僧帽弁狭窄 | 0.998 | 0.982 |
| | 大動脈弁閉鎖不全 | 0.976 | 0.930 |
| | 大動脈弁狭窄 | 0.996 | 0.965 |
| | 心房中隔欠損 | 0.996 | 0.992 |
| | 心室中隔欠損 | 0.994 | 0.992 |
| | 動脈管開存 | 0.997 | 0.994 |
| | 肺動脈弁狭窄 | 1.000 | 1.000 |
| | 心筋症 | 0.996 | 0.984 |
| | 心房細動 | 0.998 | 0.970 |
| | 期外収縮 | 1.000 | 1.000 |
| | 虚血性心疾患 | 1.000 | 0.997 |

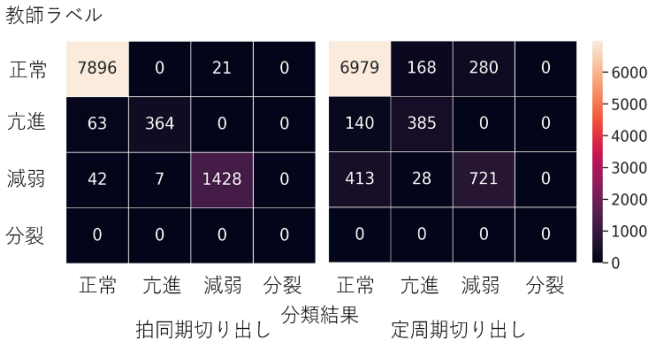


**図 2 「I 音」に対する混同行列**


**図 3 「II 音」に対する混同行列**

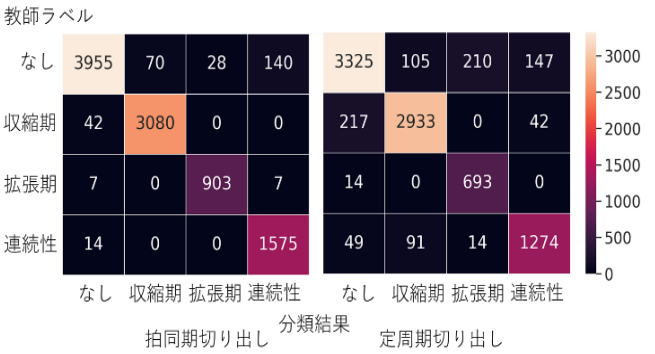


**図 4 「雑音」に対する混同行列**

#### 4.4.2 説明文生成

本実験において生成された説明文の例を表 4 と表 5 に示す。これらは拍同期切り出しに対する結果からサンプリングしたものである。表 4 は説明文が適切と判断される例、表 5 は誤りを含む不適切な例である。表 4 では、詳細度の値が 20 前後であれば異常の有無を端的に述べ、詳細度が増すにつれて言葉を補って長い説明を与える動作が確認できる。表 5 に示されるように、本システムの生成文に含まれる誤りは診断内容に関わるものと言語的なものとに大別される。これらへの対処は今後の課題である。なお、生成された説明文が、正解とする説明文と文字列として完全に一致する例は全体の 19.5% であった（一致しないものが全て適切でない説明文というわけではない）。BLEU1 スコアは 0.82 であった。

図 5 に、指定した詳細度と、実際に出力された説明文に対して計算した詳細度の関係を示す。傾き 1 の直線上に分布していることから、詳細度の制御は有効に機能していると考えられる。

**表 4　生成された説明文の例（適切な例）**

| 音源名 | 詳細度 | 生成された説明文 |
|---|---|---|
| 正常な心音 | 19 | 正常な心音です。 |
| | 42 | 正常な心音です。異常は聞き取れません。 |
| II音の幅広い分裂 | 26 | 心音に異常が見られます。 |
| | 69 | 心音に異常が見られます。心臓の弁に問題があるかも知れません。 |
| | 139 | 心音に異常が見られます。心臓の弁に問題があるかも知れません。I音は正常で、II音は分裂が見られます。収縮期雑音があります。 |
| | 233 | 心音に異常が見られます。心臓の弁に問題があるかも知れません。I音は正常で、II音は分裂が見られます。収縮期雑音があります。病気の兆候として大動脈弁狭窄 の可能性が考えられますので、受診をお勧めします。 |
| 冠動脈疾患 | 26 | 心音に異常が見られます。 |
| | 45 | 心臓や血管に問題があるかも知れません。 |
| | 70 | 心音に異常が見られます。心臓や血管に問題があるかも知れません。 |

**表 5　生成された説明文の例（誤りを含む例）**

| 誤り種類 | 詳細度 | 生成された説明文 |
|---|---|---|
| 診断の誤り | 23 | 心音は聞き取れません。 |
| | 89 | 心音の兆候は僧帽弁閉鎖不全の可能性が考えられますので、受診をお勧めします。（実際には心房中隔欠損） |
| 言語としての誤り（下線部） | 19 | 心音な心音です。 |
| | 43 | 心音の弁に問題があるかも知れません。 |
| | 65 | 心音に異常が見られます。心臓の弁な病気かも知れません。 |

## 5. おわりに

本稿では、心音に対する説明文の自動生成システムを提案した。実験は、55 症例、約 65 分という限られたサイズのデータセットに基づくものであるが、クラス分類、説明文生成、および説明文生成における詳細度の制御の動作を確認することができた。現時点では、有効な説明文が生成される場合と内容的に誤りを含む説明文が生成される場合とがあるが、誤りは主として学習用データセットのサイズが説明文生成の目的には十分でないことに起因するものと推察される。今後、学習用データの量とバリエーションを増やしながらシステム構成にも検討を加え、更に精度を向上させて、有用なシステムとしたいと考えている。

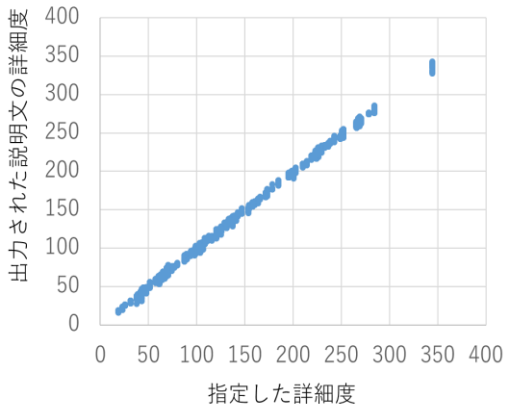


**図 5 指定した詳細度と出力された詳細度の関係**

### 謝辞

本研究に対し多大な支援を頂いた NTT の中島寛主幹研究員、川西隆仁主幹研究員に感謝します。

### 参考文献

[1] Supreeya Swarup, and Amgad N Makaryus: Digital stethoscope, Medical Devices: Evidence and Research, 11, 29-36 (2018).
[2] Adam Rao, Emily Huynh, Thomas J Royston, Aaron Kornblith, Shuvo Roy: Acoustic Methods for Pulmonary Diagnosis, IEEE Rev. Biomed Eng, 12, 221-239 (2019).
[3] 箸尾谷健二, 高田信一, 福水洋平, 山内寛紀, 来見良誠, 谷徹: 人体の心拍音・呼吸音・脈音分離手法に基づく異常周期を持った循環器系疾患の検出, 日本音響学会誌, 68, 8, 387-396 (2012).
[4] 林知樹, 戸田智基: 統計的手法による音響イベント検出, 日本音響学会誌, 75, 9, 523-537 (2019).
[5] 井本桂右: 音響イベントと音響シーンの分析, 日本音響学会誌, 74, 4, 198-207 (2018).
[6] 伊藤彰則: 環境音から異常を検知する統計的手法, 日本音響学会誌, 75, 9, 538-543 (2019).
[7] Michael Mandel, Justin Salamon, and Daniel P.W. Ellis (Eds.): Proc. of the Detection and Classification of Acoustic Scenes and Events 2019 Workshop (DCASE). New York University, NY, USA (2019).
[8] Cristhian Potes, Saman Parvaneh, Asif Rahman, and Bryan Conroy: Ensemble of feature-based and deep learning-based classifiers for detection of abnormal heart sounds. In Computing in Cardiology Conference (CinC), 621–624 (2016).
[9] Oriol Vinyals, Alexander Toshev, Samy Bengio, and Dumitru Erhan: Show and Tell: A Neural Image Caption Generator. In IEEE Conf. on Computer Vision and Pattern Recognition (CVPR), (2015).
[10] Mengyue Wu, Heinrigh Dinkel, and Kai Yu: Audio Caption: Listen and Tell, Proc. of the IEEE Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP) (2019).
[11] Konstantinos Drossos, Sharath Adavanne, Tuomas Virtanen: Automated Audio Captioning with Recurrent Neural Networks, Proc. IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA) (2017).
[12] Shota Ikawa and Kunio Kashino: Neural Audio Captioning Based on Conditional Sequence-to-Sequence Model, Proc. of the Detection and Classification of Acoustic Scenes and Events 2019 Workshop (DCASE). 99–103 (2019).
[13] Yusuf Aytar, Carl Vondrick, Antonio Torralba: Soundnet: Learning sound representations from unlabeled video, Advances in Neural Information Processing Systems (NeurIPS) (2016).
[14] https://github.com/keunhong/pytorch-soundnet
[15] 沢山俊民: CD による聴診トレーニング <心音編>, 南江堂, 改訂第 2 版 (1994).