様　式　Ｃ－１９、Ｆ－１９－１、Ｚ－１９　（共通）

科学研究費助成事業　　研究成果報告書

令和　６ 年　 ６ 月 １２ 日現在

機関番号：１５３０１
研究種目：基盤研究(C)（一般）
研究期間：2021～2023
課題番号：２１Ｋ１１９６３

研究課題名（和文）感情や個人性を高品質に表現可能なDNNに基づく音声合成方式の研究

研究課題名（英文）A Study on DNN-based speech synthesis to improve expressiveness of emotion and speaker individuality

研究代表者

阿部　匡伸（Abe, Masanobu）

岡山大学・ヘルスシステム統合科学学域・教授

研究者番号：７０５９５４７０

交付決定額（研究期間全体）：（直接経費）　　3,200,000円

研究成果の概要（和文）：DNN音声合成において、非言語情報の感情や個人性を表現する方式を２つのアプローチで検討した。アプローチ１は、言語情報の伝達は二の次とし、感情の伝達を優先する。数時間程度の通常発話データを用いて学習した後、１０分程度の感情音声で再学習する。人間の発声した音声らしさを保ちつつ、感情音声を合成できること、感情の強度も制御できることが明らかとなった。アプローチ２は、入力の音声に含まれる感情を合成音声に反映する方式である。声質変換で話者性を変更した後に、感情情報をx-vectorで条件付けしてTTSする。評価実験の結果、入力音声の感情ばかりでなく、感情の強弱も反映させて合成できることが明らかとなった。

研究成果の学術的意義や社会的意義
テキストからの音声合成（Text-to-Speech：TTS）は，近年の深層学習（Deep Leaning）に基づくDeep Neural Networks(DNN）を利用することで飛躍的に性能が向上し，音声対話システムに導入されるに至った．音声は人間にとって便利なコミュニケーション手段であることから、今後もさらに合成音声の用途が増えていくと考えられる。その際、非言語情報である感情や個人性を含めた多様性を十分に表現することが肝要である。本研究は少量のデータによって非言語情報を制御する方式を提案しており、今後の音声対話システムを広く展開する上で必要な要素技術となる。

研究成果の概要（英文）：In DNN-based speech synthesis, we examined methods for expressing non-verbal information such as emotions and individuality through two approaches. Approach 1 prioritizes the transmission of emotions over linguistic information. After training with several hours of normal speech data, the model is retrained with about 10 minutes of emotional speech. It was found that this approach can synthesize emotional speech while maintaining the naturalness of human speech, and it also allows for controlling the intensity of the emotion. Approach 2 reflects the emotions contained in the input speech into the synthesized speech. After modifying the speaker's identity through voice conversion, the emotional information is conditioned with x-vectors for TTS. Evaluation experiments revealed that this approach can synthesize speech that reflects not only the emotions of the input speech but also the intensity of those emotions.

研究分野：音声情報処理

キーワード：音声合成　感情　話者性　音声対話システム　DNN

１．研究開始当初の背景

テキストからの音声合成（Text-to-Speech：TTS）は、近年の深層学習（Deep Leaning）に基づく Deep Neural Networks(DNN）を利用することで飛躍的に性能が向上し、音声対話システムに導入されるに至った。しかしながら、非言語情報である感情や個人性を含めた多様性を表現するには至っていない。非言語情報は音声が担う重要な情報であり、これを欠く現状の音声対話システムの応答は無味乾燥となっており、ユーザからの不満が多いばかりか、音声対話システムの応用領域を狭めていることは否めない。本研究では、DNN の利用による非言語情報表現の向上を実現する方式を検討する。

２．研究の目的

DNN に基づく TTS においては、その学習に大量な音声データを必要とすることが課題である。特に、感情音声では明確に喜怒哀楽を分類することは困難であり、本質的にその境界は曖昧であるため、感情データベースの設計自体が困難である。また、感情を保ちつつ長時間の音声を発声することも難しいため、音声データの収集もままならない。一方、話者性は多様なため学習用の音声データは多量になる。また、全ての話者の音声データを集めることは不可能であるため、話者性の補完や推定が必要となる。本研究では、非言語情報を表現できる音声合成方式を、少量の学習音声データで実現することを検討する。

３．研究の方法

本研究では２つのアプローチで非言語情報の合成を試みる。アプローチ１は、言語情報の伝達は二の次とし、感情の伝達を優先する。これにより、小量なデータで感情の制御に関する知見を得ることができる。また、応用の面では、感情表出だけを担う機能として利用することが考えられる。アプローチ２は、入力の音声に含まれる感情を合成音声に反映する方式である。感情の表出は個人毎に異なることから、汎用的に精度の高い感情認識器の実現は困難である。この方式では、カテゴリカルな感情認識は不要となる点に特徴がある。

４．研究成果

（１）言語情報は含まず、感情情報だけを表出する音声合成方式（アプローチ１）

概要を図１に示す。WaveNet を利用した合成方式である。学習では、Step1 で言語情報を含む音声を生成できるように学習を行う。ここでは、大規模（数時間程度）な “Neutral”データを用い、補助特徴量はメルスペクトログラムと感情 ID である。感情 ID は one-of-K 表現を用いる。その後、Step2 で少量（１０分程度）な感情音声（“Angry”“Happy”）とそれに対応する感情 ID だけを用いて再学習する。メルスペクトログラムは使用しない。Step1 で学習済みの重み係数を初期値とするため、人間の発声した音声らしさを保ちつつ、感情表現を学習することができる。音声の合成は、学習済み WaveNet モデルに感情ラベルと最初のデータ点を与えることで WaveNet が連続的に音声を生成する。感情の強さは、感情 ID の one-of-K に「負の重み（α）」をかけて調整する。「負の重み（α）」とは、α が０の時に最も感情が強く、α が１の時に最も感情が弱くなるような重みである。

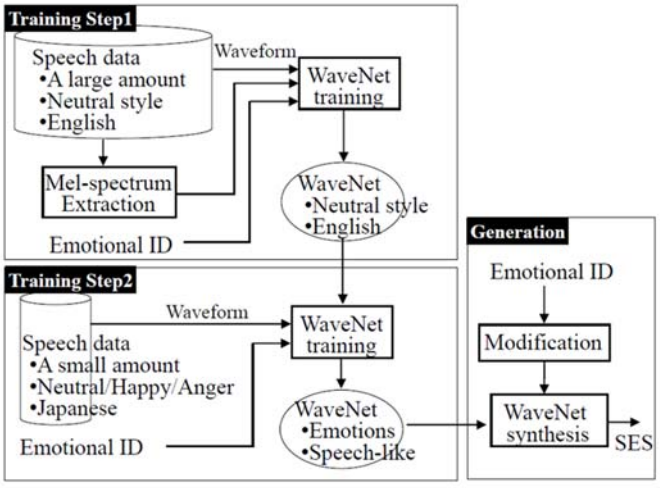


図１　感情情報だけを合成する合成方式

① 感情の強さ制御に関する実験

提案方式によって感情表出ができることは、既に確認済みであるため、α操作による感情の強さ表現の可能性を検討する。感情の強さの評価は Mean Opinion Score (MOS) テストで行った。実験に使用した音声は、感情毎に５発話×６種類=３０発話であり、３回繰り返した。また、実験に使用するすべての合成音声の長さは４秒である。各発話に対して５段階（５：とても感情が強い～１：とても感情

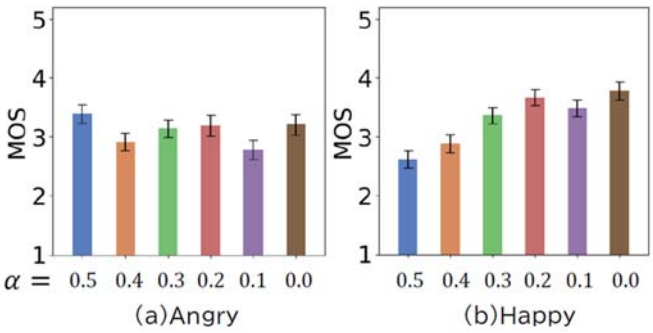


図２　感情の強さの制御性

が弱い）で評価をおこなった。実験参加者は２０代の男女１２ 名である。結果を図２に示す。なお、エラーバーは 95%信頼区間を示す。図２(a) から、“Angry” ではαの違いによって知覚される感情の強さの変化が小さい。また、αの大きさに応じて感情の強さも上下するような相関がなく、αの操作によって感情の強さの制御が可能とは言い難い。一方、図２(b) から、“Happy” ではαが０の際に最も強い感情として知覚されている。さらに、αを０に近づけていくと、概ね知覚される感情が強くなっており、感情の強さを制御できているといえる。表１ はαを“Neutral”“Angry”“Happy”のいずれかを０、他は１として音声を合成し、被検者に合成音声の感情を識別させた実験結果（混同行列）である。表１ から“Angry”は“Neutral” との混同が多く、差が小さい。このため、“Angry”では、αによって感情ラベルによる制御が難しくなったと考えられる。

表１　感情合成音の聞取り実験による感情認識率

	Subject-perceived emotions		
Correct emotions	Neutral	Angry	Happy
Neutral	**0.927**	0.050	0.023
Angry	0.300	**0.600**	0.100
Happy	0.200	0.077	**0.723**

（２）声質変換と参照音声を用いた感情制御 TTS の音声対話システムへの応用（アプローチ２）

音声対話システムでは、感情や強調などの非言語的な情報を含む音声を合成し、状況に応じて様々な応答を生成することが重要である。一方で先に述べたように、DNN を学習するために、大量な感情音声や多数の話者の音声を収集することは困難である。そこで、入力の音声に含まれる感情を合成音声に反映する方式を検討した。検討では、教師である音声対話システムが生徒の発話に応答するタスクを設定した。音声の感情表現という観点では、図３のようにユーザの発話に含まれる感情とその強さを合成音声に反映させるというアプローチである。提案方式では、音声対話システムはユーザの感情と同じ感情で応答する。

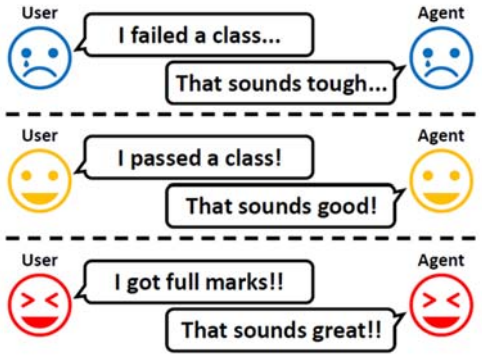


図３　発話者の感情で応答する音声対話システム

①提案方式

提案方式は２つのステップから構成される（図４）。VC（声質変換） Step では、入力音声の話者性を音声対話システムの TTS モデルの話者性へ変換する。TTS（テキスト音声合成） Step では、VC Step で変換された音声から抽出された x-vector を TTS モデルに埋め込んで感情音声合成をおこなう。

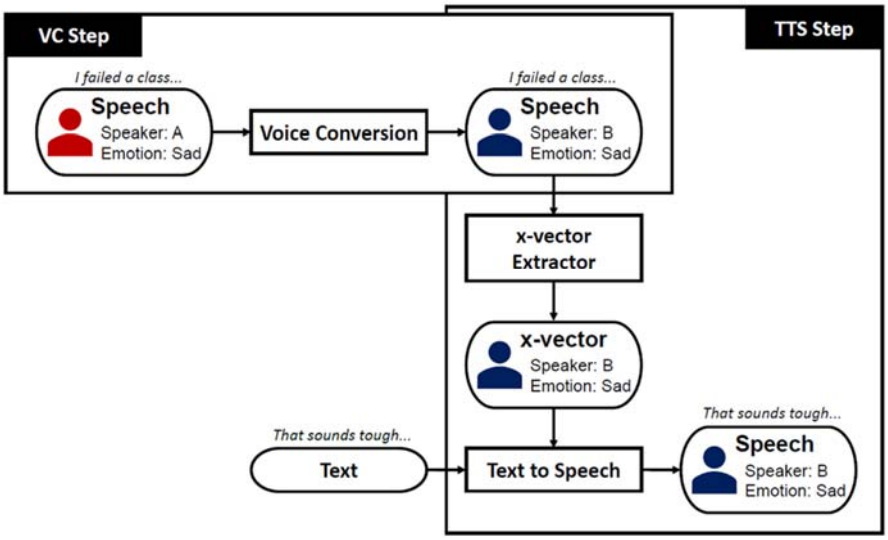


図４　声質変換と x-vector による発話者の感情を付与する TTS

x-vector とは、話者認識に利用される特徴量であり音韻情報以外の話者毎の特徴である。TTS において x-vector で条件づけして学習することで、話者性を制御できることが知られている。そこで、特定話者の感情音声から x-vector を抽出し、これで条件づけして TTS を学習しておけば、感情性を制御できると考えられる。提案方式では、x-vector は話者情報と感情情報を含んでいると考え、前段の VC Step で話者情報を特定話者に変換しておき、変換音声から x-vector を抽出して感情情報を制御する。

②実験データ　実験に使用したデータセットは、個別指導塾の女性講師役と男子生徒役、女子生徒役の３名の話者の日本語音声コーパスであり、女性講師（Teacher）と男子生徒（M-student）、あるいは、女性講師と女子生徒（F-student）との模擬対話音声である。各発話には感情ラベル

が振られており、平静（Neutral）、喜び（Happy）、悲しみ（Sad）の３種類に分類される。それぞれ、300～400 文の音声である。

③VC Step の客観評価　図５に３話者の Ground Truth、男子・女子生徒音声を女性講師に変換した音声とから抽出した x-vector の散布図を示す。なお、散布図にあたって、x-vector の次元圧縮を t-SNE によって行っている。図から、女性講師、女子学生、男子学生の声質は明らかに異なっていることが見て取れる。一方、男子・女子生徒音声を女性講師に変換した音声は、女性講師に近づいていることがわかる。この結果から、声質変換により話者性が制御できていると考えられる。

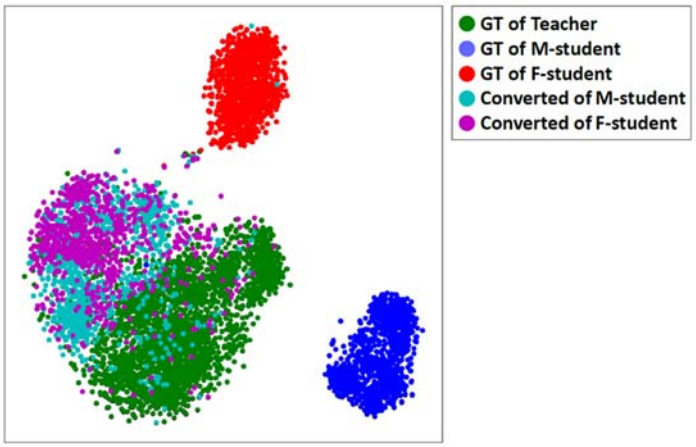


図５　原音声と変換音声の x-vector の分布

次に、図６に男子・女子生徒の変換前後の x-vector の散布図を示す。図５と同様に t-SNE で次元圧縮している。変換前の散布図より、Happy と Sad は明確に分離されているが、Neutral と Sad は似通っていると言える。これは、そもそもの発声において、Neutral と Sad が似通っていることを意味している。また、変換後の散布図を見ると、この２つの傾向は保存されていると言える。以上より、声質変換によって話者性は変換されるものの、感情性は保存されることが明らかとなった

④主観評価　感情認識テストとして、実験参加者に合成音声を聴かせ、Neutral、Happy、Sad 中から知覚した感情を選択させた。実験参加者は 20 代の日本語母語話者 10 名である。音声は以下の３種類の条件で合成した。

１. 女性講師の Ground Truth の x-vector（感情毎）を埋め込む合成音（Teacher-GT）
２. 男子生徒の変換音声の x-vector（感情毎）を埋め込む合成音（MStudent-Converted）
３. 女子生徒の変換音声の x-vector（感情毎）を埋め込む合成音（FStudent-Converted）。

合成音声のテキストは、ニュース原稿とし、発話内容による感情知覚への影響を除いた。情認識テストの結果の混合行列を表２に示す。表から、提案方式は Neutral と Happy の認識率は Teacher-GT とほぼ同等のため、十分に制御できているといえるが、Sad は大きく劣る結果となった。この結果は、図６に示したように、そもそもの発声において、Neutral と Sad が似通っていたためと考えられる。特に図６(b)の女子生徒では、Neutral と Sad の差が小さいため、表２(d)の Sad の結果が悪い。

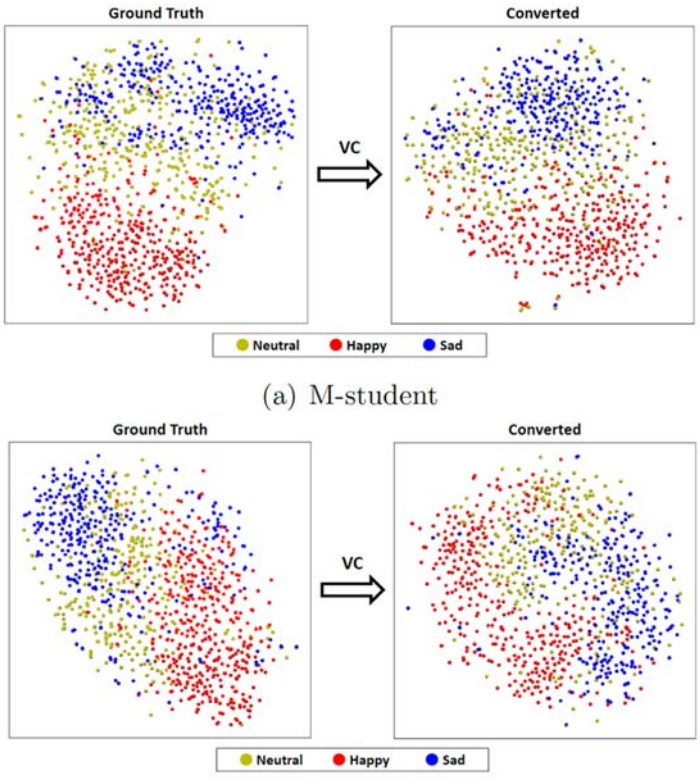


図６　原音声を声質変換した後の感情の分布

次に、感情の強さの表現性を確認するために、一対比較実験を行った。２音声を連続して聴いてもらい、どちらの音声がより喜んで（悲しんで）感じられたかを選択してもらった。喜びと悲しみに関して、以下の５種類の音声を実験に用いた。

１．強い感情の変換音声の x-vector で合成された音声
２．弱い感情の変換音声の x-vector で合成された音声
３．OneHot-EmoID で合成された音声（感情だけを指定する方式、強弱は制御できず一定である）
４．男子生徒の強い感情の自然音声

５．男子生徒の弱い感情の自然音声

図７ に Thurstone’s Case-V 尺度値を示す。値は一対比較実験の結果から算出され、値が大きいほど感情表現が強いことを意味する。図７(a)から、原音声の強い Happy と弱い Happy が明確に知覚されていることがわかる。さらに強弱の Happy から抽出した x-vector で合成された音声も、明確に強弱が知覚されていると言える。一方、OneHot-EmoID の合成音声は強弱の中間にプロットされていることから、中間的な Happy として知覚していると言える。以上のように、Happy については、感情の強弱が制御できていると言える。一方、図７(b)では、そもそも原音声の強い sad と弱い sad があまり明確に知覚されていない。強い sad は適切に制御されておらず、OneHot-EmoID との区別ができていない。これは、図６、表２の実験結果でも述べたように、Neutral と Sad が似通っていたためと考えられる。

表２　聞き取り実験による感情認識

(a) OneHot-EmoID (accuracy: 0.903)

Correct emorion	Subject-perceived emotions		
	Neutral	Happy	Sad
Neutral	**0.940**	0.000	0.060
Happy	0.120	**0.880**	0.000
Sad	0.110	0.000	**0.890**

(b) Teacher-GT (accuracy: 0.963)

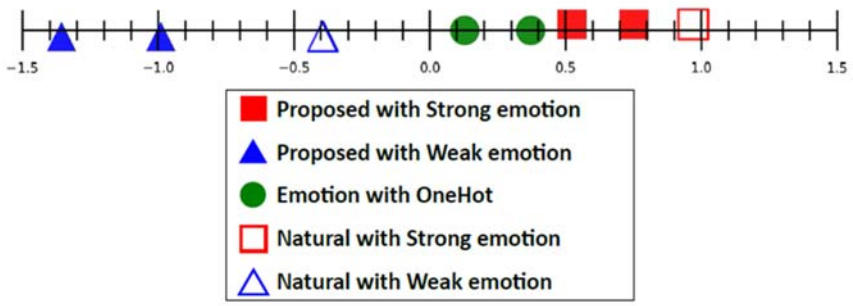
Correct emorion	Subject-perceived emotions		
	Neutral	Happy	Sad
Neutral	**0.970**	0.020	0.010
Happy	0.000	**1.000**	0.000
Sad	0.080	0.000	**0.920**

(c) M-student-Converted (accuracy: 0.830)

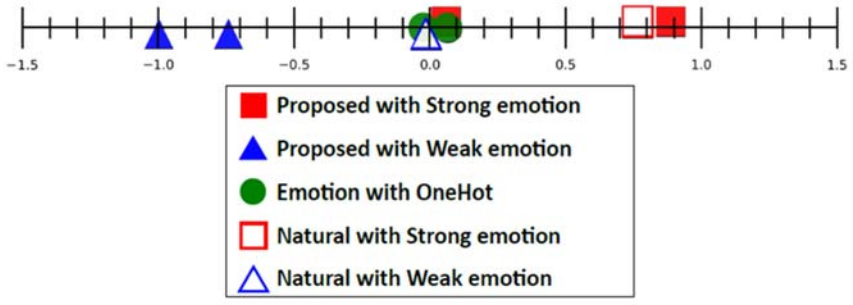
Correct emorion	Subject-perceived emotions		
	Neutral	Happy	Sad
Neutral	**0.860**	0.030	0.110
Happy	0.000	**1.000**	0.000
Sad	0.370	0.000	**0.630**

(d) F-student-Converted (accuracy: 0.690)

Correct emorion	Subject-perceived emotions		
	Neutral	Happy	Sad
Neutral	**0.870**	0.040	0.090
Happy	0.040	**0.960**	0.000
Sad	0.730	0.030	**0.240**



(a) Happy



(b) Sad

図７　感情の強さの主観評価実験結果
(Thurstone’s Case-V 尺度値)

５．主な発表論文等

〔雑誌論文〕　計1件（うち査読付論文　1件／うち国際共著　0件／うちオープンアクセス　0件）

１．著者名 Kento Matsumoto, Sunao Hara, Masanobu Abe	４．巻 Vol.E105-D, No.9
２．論文標題 Speech-like Emotional Sound Generation using WaveNet	５．発行年 2022年
３．雑誌名 IEICE Transactions on Information and Systems	６．最初と最後の頁 1581, 1589
掲載論文のDOI（デジタルオブジェクト識別子） 10.1587/transinf.2021EDP7236	査読の有無 有
オープンアクセス オープンアクセスではない、又はオープンアクセスが困難	国際共著 －

〔学会発表〕　計7件（うち招待講演　0件／うち国際学会　1件）

１．発表者名 青谷直樹，原直，阿部匡伸
２．発表標題 話者特徴量の操作によりシームレスに話者性を制御できるEnd-to-End 音声合成方式の検討
３．学会等名 電子情報通信学会
４．発表年 2022年

１．発表者名 植田遥人，原直，阿部匡伸
２．発表標題 差分メルケプストラムを用いた声質変換による喉締め歌唱音声改善方式の検討
３．学会等名 日本音響学会
４．発表年 2022年

１．発表者名 小原俊一，阿部匡伸，原直
２．発表標題 音声対話システムのための入力音声の感情に同調する声質変換とx-vector 埋め込みを用いたテキストからの音声合成方式の検討
３．学会等名 電子情報通信学会
４．発表年 2023年

１．発表者名
和田拓海，原直，阿部匡伸

２．発表標題
小説オーディオブックの強調部分を学習に用いる抑揚制御可能なEnd-to-End 音声合成方式の検討

３．学会等名
日本音響学会

４．発表年
2023年

１．発表者名
高島和嗣，阿部匡伸，原直

２．発表標題
口唇特徴量を利用した知識蒸留による舌亜全摘出者の音韻明瞭度改善法の検討

３．学会等名
電子情報通信学会技術研究報告

４．発表年
2022年

１．発表者名
小原俊一，阿部匡伸，原直

２．発表標題
音声対話システムのテキスト音声合成における声質変換とx-vector 埋め込みを用いた感情制御方式の検討

３．学会等名
日本音響学会

４．発表年
2023年

１．発表者名
Shunichi Kohara, Masanobu Abe, Sunao Hara

２．発表標題
Speech-Emotion Control for Text-to-Speech in Spoken Dialogue Systems Using Voice Conversion and x-vector Embedding

３．学会等名
APSIPA（国際学会）

４．発表年
2023年

〔図書〕　計0件

〔産業財産権〕

〔その他〕

感情や個人性を高品質に表現可能なDNNに基づく音声合成方式の研究 https://site-330980-4570-3498.mystrikingly.com/ 感情や個人性を高品質に表現可能なDNNに基づく音声合成方式の研究 https://site-330980-4570-3498.mystrikingly.com/

6．研究組織

	氏名 （ローマ字氏名） （研究者番号）	所属研究機関・部局・職 （機関番号）	備考
研究分担者	原　直 (Hara Sunao) (50402467)	岡山大学・ヘルスシステム統合科学学域・助教 (15301)	

7．科研費を使用して開催した国際研究集会

〔国際研究集会〕　計0件

8．本研究に関連して実施した国際共同研究の実施状況

共同研究相手国	相手方研究機関