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　５ 年　 ６ 月 ２３ 日現在

機関番号：２３６０４
研究種目：基盤研究(C)（一般）
研究期間：2020～2022
課題番号：２０Ｋ１１８６９

研究課題名（和文）高齢者への音声による効果的な情報伝達のための韻律制御モデルの構築と評価

研究課題名（英文）Construction and evaluation of a prosody control model for effective information transmission by speech to the elderly

研究代表者

水野　秀之（Mizuno, Hideyuki）

公立諏訪東京理科大学・工学部・教授

研究者番号：３０８３３８９２

交付決定額（研究期間全体）：（直接経費）　　3,400,000円

研究成果の概要（和文）：2020年度は高齢者から最も発話が聞き取りやすいと評価された女性話者1名により重要箇所にラベルを付与した136文書の高齢者を意識した発話と読み上げ発話の収集を行った．
2021年度は2種類の音声間の韻律の比較分析を行い，F0の平均値及びレンジの拡大及び重要箇所でのF0最大値の上昇を確認した．
2022年度は韻律制御モデルの構築を行い客観評価ではF0最大値については決定係数0.75と高い精度で制御可能であることがわかったが，分析合成音を用いた主観評価実験からは聞き取りやすさに関する効果は確認できなかった．また，重要箇所を予測する言語モデルを構築し約81%と高い精度で予測可能であることが確認した．

研究成果の学術的意義や社会的意義
1）高齢者にとって聞き取りやすいと評価されている話者が同一内容の文章を高齢者を意識して発話した音声と読み上げた音声をパラレルで収集し，両者の韻律的な差異を統計的に分析することで，高齢者にとって聞き取り易い音声の韻律的な特徴を明らかにした．
2）読み上げ音声から高齢者向け発話の韻律を予測する韻律予測モデルを構築し，高い精度で予測可能であることを示し，通常の読み上げ音声から高齢者にとって聞き取りやすい音声への変換が可能であることを示した．
3）高齢者の情報取得の観点から重要と考えられる文書内での重要な箇所を言語モデルによって高精度に予測することが可能であることを示した．

研究成果の概要（英文）：In 2020, we created 136 documents with labels attached to important parts by a female speaker who was evaluated as being the easiest to hear by elderly, and collected both of utterances with conscious of elderly and reading style utterances.
In 2021, we conducted a comparative analysis of the prosodic differences between the two types of speech, and confirmed the expansion of the average value and range of F0 and the increase of the maximum F0 value at important parts.
In 2022, we constructed a prosody control model and confirmed that the F0 maximum value can be controlled with a high accuracy of 0.75 as a coefficient of determination by objective evaluation, but we did not found an effect on the ease of hearing by a subjective evaluation using analysis-by-synthesis speech. In addition, we constructed a language model that predicts important parts and confirmed that it is possible to predict with a high accuracy of about 81%.

研究分野：音声情報処理

キーワード：高齢者向け発話データの整備　韻律分析　韻律モデル構築　言語モデル構築

１．研究開始当初の背景

近年，情報処理学会においてアクセシビリティ研究会が発足したように，高齢者の情報の取得を支援する情報処理技術の開発が重要な課題となっている．これまで，認知機能が衰え始めた高齢者を支援するための技術として字幕付与等の視覚系や居場所明確化等の行動系だけではなく，聴覚を通した情報把握の支援のための補聴器などの音響系の研究も行われてきた．しかし，音声言語系では話速を一律に伸長した音声での短文の理解への効果を測るといった研究[1]など限定的な条件下での成果は有るものの，理解しやすさに関わる音声や言語の特徴が十分明らかにされているとは言えない．また，会話術といった観点では，高齢者応対者育成のための教科書などが存在するものの“大きな声でゆっくりと話す”といった記述に留まっている．しかし，介護の現場等で高齢者の対応に慣れた人は内容によって大きさや速さに強弱をつけて高齢者に話しかけており，高齢者に対する話し方には一定の方法論があると思われるが明確化されているとはいえない．これまで我々は商品宣伝や高齢者への語りかけなどの様々な場面の音声と読み上げ音声との比較を通して，各場面の音声のどこで特異な現象が起こるかも明らかにしてきた[2]．さらに，聞き手を高齢者として想定した音声において段落境界や箇条書き項目といった文書の構造と文間ポーズの長さとの相関について解明も行なってきた[3].

２．研究の目的

本研究では高齢者が聞いて分かりやすいと評価する発話が，非高齢者に対する発話と“どこがどのように異なるのか”を韻律的特徴の比較分析によって明らかにし，さらに韻律的特徴と文書の言語的特徴との関連を定量化することによって，合成対象の文章の特徴に基づいた高齢者にとって聞き取りやすい韻律を有する音声を合成する方法の確立を目指す.

３．研究の方法

本研究は大きく分けて 4 ステップで進めた．まず，高齢者に対する情報提供といった観点から発話用文書の作成を行い，聞き取りやすいと高齢者から評価される話者の選択および聞き手としての高齢者を想定して発話した音声(Elderly)と読み上げ発話(Reading)の 2 種類の音声の収録を行い，それらの音声に対し基本周波数，継続時間長等の韻律的特徴を抽出し，文章中の重要箇所ラベルを含めて音声コーパスとして整備した．次に，前記 2 種類の音声の韻律的特徴の統計的な差異について分析を行うことで，Elderly の韻律的特徴に影響を与えている可能性が高い素性を明らかにするとともに，Reading の韻律的特徴から Elderly の韻律的特徴を予測する韻律の予測モデルの構築を行った． また，文書内の言語的な重要箇所を予測するため深層学習に基づく言語モデルを構築し評価を行った．最後に，韻律予測モデルに基づいて韻律を予測した合成音声を用いた高齢者による主観評価実験によって韻律予測モデルの評価を行った.

以下に各ステップの具体的な方法について示す.

(1) 高齢者向け発話と読み上げ発話のパラレル音声コーパスの整備

業務などにおける高齢者との対話経験者か，高齢者応対の資格を有する複数の話者が発話した音声を高齢者が聴取・評価し高い評価を得た話者を選定した．発話文章としては高齢者への情報提供の観点から．高齢者の日常生活面で重要な情報元となる自治体の広報文を模擬した文章を作成した．また，発話文章には話者自ら高齢者への情報提供の観点から重要だと判断した部分にたいしてラベル付けを行った．さらに，呼気段落，アクセント境界をラベル付けするとともに，形態素等の言語的特徴を付与した．音声としては聞き手として高齢者を想定した発話(Elderly)と特に聞き手を意識せず単に読み上げた音声(Reading)を収録した. その後, それらの音声から話速，基本周波数，ポーズ長，フレーズ長の各特徴を音声データから抽出した.

(2) 韻律的特徴の統計的な対比分析及び韻律的特徴の予測モデルの構築

Elderly と Reading の韻律的特徴の差異に関して，呼気段落及びアクセント句単位でそれらの統計量を求め比較分析を行った．また，特に文章中の重要箇所の韻律へ影響を調べるため重要箇所における韻律的特徴の差異に関し比較分析を行った．Elderly の韻律予測モデルの構築については，Reading の韻律的特徴及び言語的素性から読み上げ音声の韻律からの差分を予測するモデルを構築し客観評価を行った.

(3) 言語的な重要箇所の予測モデルの構築

情報伝達の観点から重要となる文章内の言語的な重要箇所を予測するため，深層学習を用いて言語素性から重要箇所を予測する言語モデルを構築し客観評価を行った.

(4) 高齢者による韻律制御音声の評価

韻律予測モデルを用いて Reading 発話の韻律から変換し分析合成技術を用いて音声として再構成した音声と Reading 発話をそのまま分析合成した音声とを高齢者による対比較での主観評価実験を行うことで，韻律制御による高齢者の聞き取りやすさに対する有効性を確認した.

４．研究成果

(1) 高齢者向け発話と読み上げ発話のパラレル音声コーパスの構築

発話用の文セットとして，日常的に高齢者が接する内容として自然であると考えられる自治体の広報文に類似した内容のテキストを 136 文書作成した．話者としては，高齢者が最も聞き取りやすいと評価した女性話者[4]を選択した．また収録前に当該話者に文書を提示して，高齢者に

とって重要と話者が判断した箇所にラベル(IL)を付与させた．発話時には，高齢者を聞き手として意識して発話するとともに，上記ラベルを参考にして発話するよう指示を行った．また，同一文書を特に聞き手を意識せず読み上げるよう指示して発話された音声の収録も行った．以降，高齢者を意識した発話を高齢者向け発話(Elderly)，単に文書を読み上げた発話を読み上げ発話(Reading)とする．各発話に対し手作業で呼気段落境界及びアクセント句境界のラベリングを行い WORLD [5] (D4C edition[6])を用いて基本周波数を抽出した．標本化周波数は 48 kHz，フレームシフトは 5 msec である．ただし基本周波数が 100Hz 以下のフレームは抽出誤りとして除去するとともに，明らかに抽出誤りと思われる場合は手作業で修正を行った．話速抽出のため MeCab を用いて単語境界抽出及び読み付与を行ない解析誤りは手作業で修正した．Table.1 に構築した音声コーパスの文書数や呼気段落数等の諸元を示す．

Table.1 Quantities of language related and intonation related elements

| | Total Number |
|---|---|
| Documents | 136 |
| with title | 136 |
| with contact information | 32 |
| with listing | 31 |
| Sentences | 1,291 |
| Word | 28,290 |
| With IL | 5,813 |
| Without IL | 22,477 |
| Breath group including accent phrase with IL | 1326 |
| Breath group not including accent phrase with IL | 2658 |
| Accent Phrase | 12,679 |
| With IL | 3,483 |
| Without IL | 9,196 |

(2) 高齢者向け発話と読み上げ発話間の韻律的特徴の統計的な差異の確認

Elderly と Reading のフレーム単位での基本周波数(F0)と ΔF0 の平均値と標準偏差を Table 2 に示す．また Figure 1 に F0 の分布図を示す．分布が重なっている部分を■色で示している．表より高齢者向け発話は F0 と ΔF0 の両方とも平均値と分散が大きいことから，Elderly では平均的には声が高くかつ抑揚も強めに発話されているが値のばらつきも大きいことがわかる．Figure 1 から Elderly と Reading ともに 200Hz 付近で頻度が最大となる点は同じだが、高齢者向け発話では 150Hz 前後の値が減少し 250Hz 以上の値が顕著に増加していることがわかる．また，それぞれの発話に対しアクセント句単位での韻律的特徴の統計量を求めた．ただし前述のとおり各発話の F0 に差異があるため，発話全体の F0 の平均値を基準に cent 単位で求め相対値での比較を行った．話速についてはアクセント句のモーラ数と時間長の比をアクセント句単位での話速(mora/sec)とした．結果を Table 3 に示す．Elderly では読み上げ発話と比べて最小値が下降し、最大値が上昇していることから F0 レンジが広がっていること、標準偏差が増大していることからアクセント句内での変動が大きいことがわかる．話速については，Elderly では平均値が若干上昇し，標準偏差が減少していることから少し早めの話速かつ一定速度での発話であると思われる．これらの値についてマンホイットニーの U 検定により統計的有意差検定を行ったところ全ての値で有意差($p<0.01$)が認められたことから，高齢者を意識した発話と単に読み上げた発話間においては F0 と話速において統計的に有意な差異があることが確認できた．

Table.2 Fundamental statistics of $F_0$ and $\Delta F_0$

| | Elderly | | Reading | |
|---|---|---|---|---|
| | Avg. | SD | Avg. | SD |
| $F_0$ (Hz) | 236.1 | 42.7 | 236.1 | 42.7 |
| $\Delta F_0$ (Hz/s) | 13.2 | 32.1 | 13.2 | 32.1 |

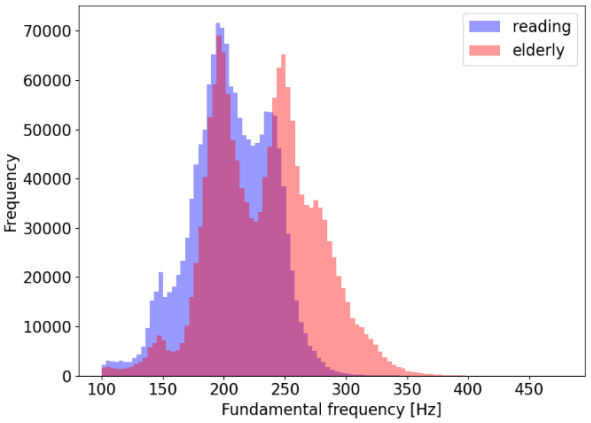


Figure 1 Histogram of $F_0$

(3) 高齢者向け発話の重要箇所における韻律的特徴の統計量の差異の確認

内容的な重要度に伴う，韻律的特徴の差異を確認するため，文書毎に重要ラベルが付与された呼気段落（以降，重要文）と付与されてない呼気段落(以降，通常文)間で比較を行った．また呼気段落の一部のみに重要ラベルが付与されている場合も呼気段落全体を重要文とした．ただし比較を行ったのは，内容を考慮した発話である高齢者向け発話のみである．文書毎に当該文書に含まれる重要文 (key sentence)，通常文(normal sentence)の F0 最大値，F0 最小値，話速，平均パワーに対し平均を求めた．Table 4 にそれぞれの値の全文書の平均値を示す．さらに，それらの値について重要文と通常文間の相関を求めた．Table 5 に重要文と通常文の相関を示す．F0 最大値にはある程度相関がみられるのに対し F0 最小値には

Table.3 Statistics of maximum, minimum standard deviation of F0 (cent) and speech rate over accent

| | Elderly | | Reading | |
|---|---|---|---|---|
| | Avg. | SD | Avg. | SD |
| Min. of $F_{0\,[cent]}$ | -627.8 | 318.8 | -540.6 | 327.0 |
| Max. of $F_{0\,[cent]}$ | 389.4 | 222.2 | 377.3 | 213.5 |
| SD. of $F_0$ | 292.2 | 218.9 | 254.9 | 175.2 |
| Speech Rate [mora/s] | 6.5 | 1.0 | 6.4 | 1.3 |

全く相関がないことがわかる. これらの結果から, F0 最小値については重要文において単に値が通常文より減少するのではなく, 文書内の他の通常文とは無関係に文章の内容に応じて細かく制御されているものと推測される. 一方 F0 最大値や F0 平均, 話速は概ね通常文と相関があることから単純に値が上下にシフトする傾向があること, またパワーにはほぼ差が無いことがわかった.

Table 4 Averages of prosodic features over breath group

| | Key sentence | Normal sentence |
|---|---|---|
| Min. of $F_{0\,[cent]}$ | 323.7 | 319.7 |
| Max. of $F_{0\,[cent]}$ | 137.9 | 146.2 |
| SD. of $F_0$ | 6.4 | 6.3 |
| Speech Rate [mora/s] | 5.0 | 5.0 |

Table 5 Correlation coefficients of $F_0$ maximum, minimum, average, speech rate and power over breath group between key sentences and normal sentences

| $F_0$ | | | Speech rate | Average Power |
|---|---|---|---|---|
| Max. | Min. | Avg. | | |
| 0.53 | 0.02 | 0.72 | 0.63 | 0.87 |

(4) 高齢者向け発話と読み上げ発話間の韻律的特徴のアクセント句単位での相関分析

今回分析に用いた 2 種類の音声データ（Elderly と Reading）は同一話者が同一内容を発声したものであり，かつ Reading の収録において Elderly でのポーズ位置にあわせて発話を行っているため，アクセント句単位で両発話の正確な対比較が可能である点が特徴である．そこで，Elderly と Reading のそれぞれ対応するアクセント句単位で文の意味的な重要性を考慮して韻律的特徴の相関関係を分析した. 重要ラベルが付与されたアクセント句か(With IL.), 否か(W/O IL.)の別に Table 6 に示す．分析結果から，F0 最大値については Elderly と Reading 間で多少相関関係がみられ，また意味的に重要な文において相関性が下がることから，重要な文であるか否かが韻律に影響を与えていると考えられる．一方 F0 最小値と F0 レンジはともに相関性は低く，特に F0 最小値については Reading の値とは無関係に Elderly の値が分布してことがわかる．話速についても F0 最小値と同様に相関が低く，また文の重要性も特に影響を与えていないことが確認できた.

(5) 高齢者向け発話の韻律予測モデルの構築

Elderly の F0 最大値，F0 最小値，F0 レンジ，話速の予測を行った．F0 に関する特徴の予測には Reading の対応するアクセント句における以下に示す素性のうち話速素性以外の素性を用い，話速に対しては，F0 素性以外の素性を用いた.

・言語素性：位置，モーラ数，当該アクセント句を含む文のモーラ数，重要ラベルの有無
・F0 素性：前後の F0 最大値，F0 最小値，F0 平均値
・話速素性：話速

予測には勾配ブースティング回帰木を用い Dropout に基づく正則化を導入した．10 交差検定による推定結果（決定係数）を Table 7 に示す．結果より Reading の韻律特徴等のみでは Elderly の韻律の予測が困難であることがわかった．そこで，言語的な素性が同一にも関わらず Reading と Elderly の F0 最小値に関連性がほとんど見られない原因として, F0 最小値に任意性が高いこと，またアクセント表現の点から F0 レンジは言語的素性に拘束されていると仮定し，Elderly の F0 レンジの予測に Elderly 自体の F0 最小値を素性として加えて予測を行った．その結果，F0 最大値の決定係数は 0.75 となり高い精度が得られた．Figure 2 に Elderly の F0 最小値を用いた場合と用いていない場合の F0 レンジの実測値と予測値の対応関係を示す．このとき学習と予測には同一のデータを用い，全データの 90%を学習，10%を評価に用いている．図より Elderly の F0 最小値を用いた場合，精度が大きく向上していることがわかる．また，各素性の寄与度を示す平均絶対 SHAP 値を Figure 3 に示す．各素性で_e は Elderly の素性，_r は Reading の素性であることを示す．寄与度最大が Elderly の F0 最小値(F0min_e)で，以降 Reading の F0 最大値(F0max_r)，モーラ数(Mora Num)，F0 平均値(F0ave_r)，後続するアクセント句の F0 最大値(s_F0max_r)の順になっている．残る 9 個の素性については SHAP 値の総和を示す．結果から，重要箇所は予測に与える影響は低く，Reading の韻律的特徴に基づいて Elderly の韻律を予測する場合については有効でないことが確認できた.

(6) 文書中の重要箇所の予測モデルの構築

前述のとおり，文章中の重要箇所は韻律的特徴に対し統計的には影響があることが確認されたため，単語系列から重要/非重要の系列への系列ラベリング問題として設定し，重要箇所を予測する言語モデルを構築した．モデルには BERT モデル[7]を用い単語自体とその文書内での位置情報を活用できるように文書単位で BERT の学習を行なった．BERT では NICT から公開のモデル[8]を用いた．また，比較用に文単位で学習した BERT モデルでの予測も行なった．それぞれ，5 分割したデータのうちの 4 つを用いて前記 BERT の fine tuning を行ない，残りの 1 つを評価データとする実験を 5 回行なった.

Table 6 Correlation coefficients of prosodic features

| Feature | With IL | W/O IL |
|---|---|---|
| $F_0$ maximum | 0.46 | 0.51 |
| $F_0$ minimum | 0.36 | 0.37 |
| $F_0$ range | 0.36 | 0.35 |
| Speech rate | 0.33 | 0.35 |

Table 7 Coefficients of determination of features

| Feature | $R^2$ |
|---|---|
| $F_0$ maximum | 0.36 |
| $F_0$ minimum | 0.18 |
| $F_0$ range | 0.18 |
| Speech rate | 0.21 |

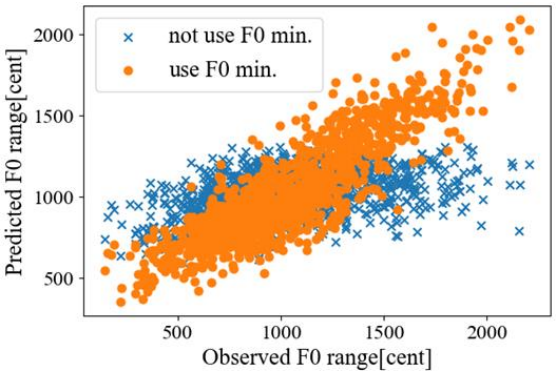


Figure 2 Comparison of prediction results

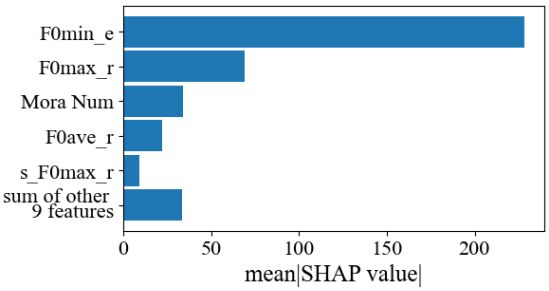


Figure 3 Mean absolute SHAP values of features

5 回の実験の予測性能の平均値を Table 8 に示す．なお，Table 8 への記述を割愛した非重要箇所の再現率と適合率は全て 8 割以上であった．重要/非重要箇所の精度と強調箇所の適合率は文書単位モデルの方が若干高かったが，再現率が低かった．今回の重要箇所のラベルは，模範話者 1 名が 1 度だけ付与したものであるので，揺れが含まれている可能性がある．また，声の高さだけでも強調に伴う変化はさまざまである．高齢者向け発話と読み上げ発話間の差異をより精緻に分析して，確かな予測対象ラベルに絞ることも必要であると考えられる．

(7) 韻律予測モデルに基づく韻律変換音声の主観評価実験による評価

アクセント句単位での言語素性及び Reading の F0 素性に加えて Elderly の F0 最小値を用いた Elderly の F0 レンジ予測モデルを用い，Elderly の F0 最小値の代わりに Reading の F0 最小値を用いることで Reading の韻律的特徴から Elderly の韻律への変換を行った．具体的には，予測モデルを用いて推定した F0 レンジを基に，Elderly 模擬音声の F0 系列を線形変換により生成した．ただし線形変換の際には cent 単位で行った．そして Reading の音声から WORLD により抽出したスペクトル包絡と非周期性指標を基に，前述のとおり変換した F0 系列を用いて Elderly 模擬音声を合成した．実験に用いた Reading の音声データは，136 文からランダムに選択した 17 文の発話音声から 16～20 秒程度の長さの音声区間を抽出したものである。このとき内容的にも当該区間で意味が取れる形で音声の切り出しを行った．また，比較対象として Reading の分析合成音，及び Reading の F0 系列をそのまま Reading の F0 系列に置換した ElderlyF0 置換分析音合成音も作成した．主観評価実験は，前述の三種類の音声の相互の対比較によって行った．評価においては，音質やノイズを気にせず聞いてわかりやすいと思った方を選択するよう指示を行った．被験者は 70 代の男性 20 名女性 20 名である．実験結果を Figure 4 に示す．Reading 分析合成音(Rabs), Elderly 模擬分析合成音(Esim)において Reading 分析合成音が最も評価が高く，次に Elderly 模擬分析合成音となり，最も評価が低かったのは Elderly F0 置換分析合成音となった．本結果は，分析合成の際の F0 変形による音声品質の劣化が評価に大きな影響を与え，韻律面での品質の向上があったとしても評価としては現れなかったものと考えられる．

Table 8 Prediction performance (average [%])

| | Doc-base | Sent-base |
|---|---|---|
| Accuracy | 81.3 | 79.0 |
| Recall of IL | 24.7 | 44.1 |
| Precision of IL | 60.7 | 47.9 |

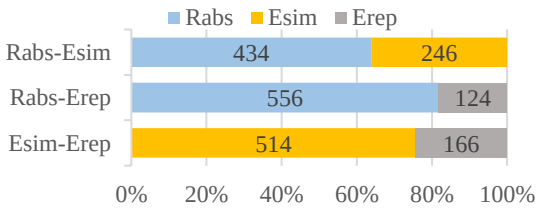


Figure 4 Subjective experiment results

参考文献

[1]成田他,”高齢者を対象とした合成音声の聞き取りやすさに関する研究",ABML2011,pp.O1-3-1,2011.
[2]H.Nakajima., et al., "Creation and analysis of a Japanese speaking style parallel database for expressive speech synthesis." In Proceedings of Oriental COCOSDA, paper id 30, 2010
[3]中嶋他,”高齢者への語りかけ音声におけるポーズ長の分析”,日本音響学会，春季研究発表会講演論文集, 3-Q-49, pp.399-400, 2015
[4]中嶋他,”高齢者にとって聞き取りやすい音声合成の実現に向けた模範話者データの収集と分析”,日本音響学会　春季研究発表会講演論文集,3-P-26，2020.
[5]M.Morise,et al.,”WORLD:avocoder-based high-quality speech synthesis system for real-time applications,” IEICE transactions on information and systems, vol. E99-D, no. 7, pp. 1877-1884, 2016.
[6]M.Morise,”D4C, a band-aperiodicity estimator for high-quality speech synthesis,” Speech Communication, vol. 84, pp. 57-65,2016.
[7] J.Devlin, et al., “Google, KT, Language, AI: BERT: pre-training of deep bidirectional transformers for language understanding.”, In Proceedings of NAACL-HLT, pp. 4171-4186.
[8] https://alaginrc.nict.go.jp/nict-bert/index.html

5．主な発表論文等

〔雑誌論文〕　計0件

〔学会発表〕　計3件（うち招待講演　0件／うち国際学会　0件）

| |
|---|
| 1．発表者名<br>水野秀之，中嶋秀治 |
| 2．発表標題<br>高齢者向け発話と読み上げ発話の韻律的特徴の対比分析 |
| 3．学会等名<br>日本音響学会春季研究発表会 |
| 4．発表年<br>2022年 |

| |
|---|
| 1．発表者名<br>水野秀之，中嶋秀治 |
| 2．発表標題<br>高齢者向け発話の韻律予測 |
| 3．学会等名<br>日本音響学会秋季研究発表会 |
| 4．発表年<br>2022年 |

| |
|---|
| 1．発表者名<br>中嶋 秀治，水野 秀之 |
| 2．発表標題<br>高齢者に対して強調すべきと判断された単語のテキストからの予測と分析 |
| 3．学会等名<br>日本音響学会秋季研究発表会 |
| 4．発表年<br>2022年 |

〔図書〕　計0件

〔産業財産権〕

〔その他〕

―

6．研究組織

| | 氏名<br>（ローマ字氏名）<br>（研究者番号） | 所属研究機関・部局・職<br>（機関番号） | 備考 |
|---|---|---|---|
| 研究分担者 | 中嶋　秀治<br>(Nakajima Hideharu)<br>(90832684) | 日本電信電話株式会社ＮＴＴコミュニケーション科学基礎研究所・協創情報研究部・研究主任<br>(94305) | |

**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 |
| --- | --- |