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　２ 年　 ６ 月 １９ 日現在

機関番号：８２４０１
研究種目：基盤研究(C)（一般）
研究期間：2017～2019
課題番号：１７Ｋ０２９２５

研究課題名（和文）Developing a learner-adaptive captioning system to improve second language listening

研究課題名（英文）Developing a learner-adaptive captioning system to improve second language listening

研究代表者
ＭＥＳＨＧＩ　Ｋｏｕｒｏｓｈ（MESHGI, Kourosh）

国立研究開発法人理化学研究所・革新知能統合研究センター・研究員

研究者番号：８０７７４８３５

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）：本研究では、字幕への依存度を減らしながら聞き取り能力を向上させるために、 部分的かつ同期された新たな字幕提示法を提案する。対象に合致して学習され た高精度の音声認識システムによって、発話中の各単語の音声区間に正確にタ イミングがあった字幕を生成する。その上で、学習者の聞き取りを阻害する可 能性の高い単語やフレーズを自動的に選択する。最終的な字幕は、学習者のレベルに応じた部分的な単語列から構成され、音声に同期して順次提示される。

研究成果の学術的意義や社会的意義
Improved listening skill of Japanese learners of English; Popular among university students and teachers; Facilitated training listening skill and using authentic material such as TED talks; Introduced a novel type of captioning method in Japan.

研究成果の概要（英文）：This project aimed at developing a user-adaptive caption system to realize an individualized platform for training second language listening skills. In this caption, words are synchronized by the utterance, and only difficult words are shown in the caption. The difficulty of the words is determined using word frequency and specificity, speech rate, lexical and syntactic complexity, and automatic speech recognition (ASR) errors as the indicators of acoustic/speech difficulty. ASR errors have been analyzed, and identified patterns were evaluated by experiments to ensure that they cause difficulties for L2 listeners. A listening difficulty index has been proposed using these factors. Additionally, a data-driven difficulty detector is built on top of these factors. In the proposed caption, learners can improve the word selection by feedback (adaptation), and timely and useful cues are embedded in the caption for ambiguous and difficult words/phrases (hints).

研究分野：Machine Learning and NLP

キーワード：PSC　Surprisal Model　L2 listening　ASR error analysis　Listening Difficulty Idx　Instantaneous Hints　Adaptive Captions　Individualized Cues

様　式　Ｃ－１９、Ｆ－１９－１、Ｚ－１９（共通）

１．研究開始当初の背景

Despite many studies focusing on L2 listening difficulties, analyzing the nature of speech and identifying problematic speech segments for individual L2 listeners has not been systematically investigated. This highlights the need to create a tool that can predict the speech-related difficulties for language learners and provide a learner-specific scaffold to alleviate such difficulties. L2 listening entails constant effort as the speech stream is transient and listeners need to process each part of the input quickly before receiving the next part, without having the option to return to the earlier points. During this process, learners need to go through perception, recognition, comprehension, meaning construction, ambiguity resolution, inferencing, and many more in a short time. Thus, listeners are continuously subjected to handle a number of cognitive tasks on-the-fly, which imposes large working memory loads. To foster L2 listening, captioning is used as a popular tool that facilitates the comprehension of the input by allowing for reading the text along with listening to the audio. The use of captions, however, is subjected to some limitations as it promotes more reading than listening, thus inhibits listening skill development. and increases the cognitive load by providing too much text. Many learners, especially beginners, struggle with cognitive load and split attention, while attending to caption text together with other modes of input.

While L2 listeners need to process input attentively and utilize skills adeptly to gain adequate comprehension, some factors associated with the speech input itself can impede their listening. The lexical units used in speech and the ambiguities involved in articulation, such as uncertain word boundaries, can lead to difficulties for many learners. Despite many studies focusing on L2 listening difficulties, analyzing the nature of speech and identifying problematic speech segments for individual L2 listeners has not been systematically investigated. This highlights the need for a tool that can predict speech-related difficulties for language learners and provide a learner-specific scaffold to assist in learner comprehension.

Another element that significantly affects learner’s misrecognition of the words and results in difficulty for L2 listening is the lexical and syntactic surprisal. This notion is based on the surprisal theory, which assumes that difficulty can be determined by a word's predictability. Surprisal is intended as high-level theory grounded in the principles of information theory, providing a possible explanation for difficulty. In this view, the cognitive effort it takes for the learner to process a word is proportional to its surprisal. Speech is transient, and when a learner encounters a word that is different from what she/he expects to hear, her/his attention is confined, leading to confusion, cognitive overload, and misrecognition. A similar situation happens when a learner tries to match a preferred sentence structure to an input speech and finds a mismatch.

To cover various sources of listening difficulties, it is possible to incorporate more features to the PSC system: lexical features, acoustic/speaker features, content attributes, and perceptual hindering factors. these features may be correlated with each other and finding their footprint in listening difficulty is not straightforward. Furthermore, not all the features are equally useful for detecting difficult words in the speech. To gauge the difficulty of a word, readability scores (e.g., Dale-Chall score) consider the frequency of the words in corpus but discard the speaker/acoustic information.

The baseline method provides caption for a coarse-grained proficiency categorization of the learners, while ignoring individual differences within each proficiency group, the limitation of the tests to measure the different listening difficulty features, and the effect of learners’ background on their listening comprehension (e.g. engineers listening to medical talks). Moreover, the pre-set settings do not reflect the gradual improvement of learner’s listening skills. Previous analyses revealed that some learners need additional factors to be considered when generating PSC (e.g. speech disfluencies) and others gradually adapt to the listening material (e.g. getting used to vocabulary and speech rate of the speaker), hence no longer needing some words in the caption.

Through the preliminary experiments, we found that PSC can assist L2 listeners by successfully detecting and presenting difficult segments. However, further observations suggested that merely showing the words in the captions may not provide the optimal assistance, especially for words out of the learner’s vocabulary reservoir. In such cases, learners’ attention is confined to the ambiguous segment, which inhibits them from moving on to process the next input. This happens particularly for those who overemphasize on using bottom-up strategies and word-by-word decoding.

２．研究の目的

The main focus of this research is developing a user adaptive CALL system to realize an individualized platform for training second language listening skill. Therefore, we developed a system, which detects learners’ difficulties in recognizing the speech and provides a caption that presents the detected difficult words/phrases in the caption and hides easy ones to promote listening over reading, adapt to the individual learners, and scaffolding them when they encounter listening difficulties.

Figure 1: Screenshot of a TED talk with full caption (*left*), with PSC (*middle*), and with PSC+Hints (*right*)

３．研究の方法

We developed a novel method to generate learner-adaptive captions to train L2 listening skill into a CALL system. To this end, advanced machine learning and sophisticated features gauging speech difficulty will be incorporated into the CALL system. We created a baseline system and enhanced it using the following modules:

(1) We propose the use of ASR systems as a model of L2 listeners and hypothesize that ASR errors can predict challenging speech segments for these learners. We compared ASR erroneous and PSC hidden cases in order to discover the useful candidates for PSC. In this view, we conducted a root-cause analysis on the ASR errors not shown by PSC, which are classified into the following categories: homophones, minimal pairs, negatives, breached boundaries, verb inflections, determiners, interjections, derivational suffixes, stop words, and unknown sources. We annotated the ASR substitution errors on 70 TED talks to distinguish between useful and useless ASR erroneous cases, regardless of their categories. Among different cases of ASR errors, annotation results suggest the usefulness of four categories of homophones, minimal pairs, negatives, and breached boundaries for L2 listeners. We mark a word as (i) a homophone (e.g. *"rain" /R EY N/* and *"reign" /R EY N/*) if the Levenshtein distance between the phonetic sequences of the ASR erroneous phrase and its corresponding transcript is zero, and (ii) a minimal pair (e.g. *"pin" /P IH N/* and *"bin" /B IH N/*) if the distance is one. Negative cases are detected by considering the negative particle *"not"*, plus prefixes and suffixes that form negation (e.g. *"legal"* and *"illegal"*). To detect breached boundaries, we checked for the four prominent patterns based on L2 studies: strong-syllable strategy: (e.g. *"disguise"* heard as *"the skies"*, *"ten-to-two"* heard as *"twenty to"*), assimilation rule (*"right you are"* heard as *"rye chew are"*), frequency rule (e.g. *"achieve her"* heard as *"a cheaper"*), and resyllabification (e.g. *"made out"* heard as *"may doubt"*).

(2) In another extension, we measure lexical surprisal using the probability of the next word based on a pre-trained language model. We also compute syntactic surprisal using the structural confusion of a sentence recovered by a probabilistic grammar/parser. The words with high surprisal scores are selected to be included in PSC.

(3) We propose a word listening difficulty score, as a linear combination of several complementary features. A dataset of expert-annotated partial and synchronized captions for TED talks is prepared for a target language proficiency, in which only the difficult words are shown. A classifier was trained on this dataset, and the learned features/weights were automatically transferred to the proposed score. Compared to the rule-based one, this data-driven score demonstrates higher accuracy on the annotated dataset and facilitates model and feature expansion.

(4) We provide a pool of multimodal features from word, including lexical (e.g. frequency, specificity, length, syllables), syntactic (e.g., part-of-speech, sentence structure), semantic (e.g. polysomic words, co-references, idiomacity) and acoustic or perceptual complexities (e.g., speech rate, hesitation and pauses, noise and distortion). Selecting the difficult words for a target language proficiency can be reformulated as a binary classification problem using these features. A classifier is then trained on the annotated dataset to automate the detection of difficult words/phrases for a target proficiency level of L2 learners. Using this classifier, the most informative features to make partial captions are recognized and analyzed. This data-driven PSC, is compared with rule-based version and annotated dataset.

(5) To handle individual learner demands, PSC should adapt its word selection criteria. We proposed an adaptive PSC, which improves its word selection and retrains itself on-the-fly by applying learner feedback on the generated caption to provide individualized and effective assistance that satisfies the learners' requirements. We developed the adaptive PSC, in which an online machine learning module receives the feedback from the learners and adjusts the parameters of the system on-the-fly. The feedback includes user clicks either on a masked word they wish to see or on a shown word that is too easy. The system reacts by showing/hiding the word and learns to intelligently classify words with similar features in the future. Rather than defining rules, our classifier is trained by giving several

examples for each category of words in context. Therefore, it can easily expand to support other types of listening materials (e.g., daily conversations, news) that require different rules, features, and thresholds. Additionally, the system can detect and learn the discriminative features of the learners' feedback. Such feedback serves as a bag of examples for retraining the system, which can be easily obtained from the learners and used to their advantage. The learner feedback acts as new labels for words that the system misclassified, and the classifier is retrained with feedbacks to learn about individuals' problems, backgrounds, vocabulary reservoirs, and sources of listening difficulties.

(6) We provide cues for ambiguous and difficult words/phrases in the caption while filtering out the easy words. The hints are generated in the form of short explanations/definitions of the words to allow for meaning construction and resolving difficulties on-the-fly. With the use of NLP tools and word sense disambiguation, we tried to generate appropriate hints for the selected words to provide instantaneous and minimally intrusive assistance. The process of generating hints involves: (i) determining the words/phrases that need supplementary description or clarification, and (ii) providing useful description for them to assist learners. First, we focus on the problematic categories that were specified by L2 listeners including: low-frequency, technical, ambiguous, and polysemous words, proper nouns, named entities, ambiguous references, and uncommon/multi-purpose abbreviations. Next, the proper hints for each category are retrieved from in-house and online resources. A synonym is selected as the hint for low-frequency words (e.g. *cobble* → *put* together). Wikipedia and glosses are consulted for word definitions and abbreviation expansions (e.g. *neocortex* → *part* of mammalian brain). Short descriptions are retrieved for proper nouns, named-entities, and symbolic names from Wikipedia and the Google search engine (e.g. *Basel* → *city* in Switzerland, *big apple* → *New York City*). The referent of references is displayed as a reminder hint if their co-reference was distant. For words with different meanings (e.g. polysemous words), we employed word sense disambiguation to find the most probable meanings from available synonyms/descriptions. The hints provided to the learners should be short, helpful, and relevant. To this end, we seek the shortest description for the word or generate one by searching for the keywords in the retrieved description. Along with this, a filtering process assures that the final hint includes high-frequency words that are familiar for learner.

Figure 2 depicts the proposed methods and how they fit into the framework of Partial and Synchronized Caption.

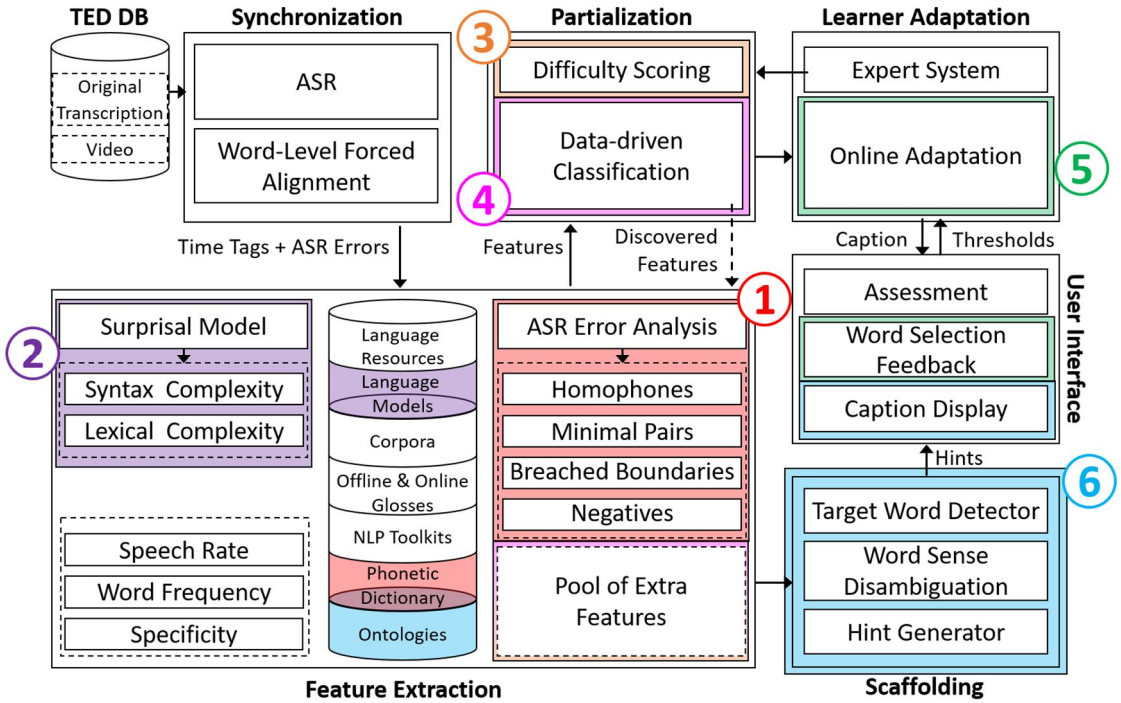

Figure 2: Partial and Synchronized Caption Framework. The proposed methods in this study is highlighted in the figure: (1) using ASR errors as the indicators for acoustic difficulties, (2) using lexical and syntax surprisal model to predict learner's cognitive difficulties, (3) calculating listening difficulty score for words, (4) data-driven partialization for the caption, (5) online adaptation of the captions based on learner's feedback, (6) providing / generating hints for the learner embedded in the caption to facilitate comprehension

４．研究成果

We evaluated different extensions of our system, with participants ranging from 30~58 Japanese and Chinese students in two classes who enrolled in CALL courses at Kyoto University. The learners were engineering majors ranging from 19 to 22 years old. The participants' scores on CASEC test ranged from 560 to 850. A pre-study questionnaire on the learners' language history background revealed that most of the learners have started studying English from the age of 10-13 with 7-9 years of experience.

(1) An experiment with L2 learners focusing on these four categories of the ASR errors revealed that these cases highlight the problematic speech regions for L2 listeners.

(2) In an experiment with pre-intermediate L2 learners of English, we asked the learners to transcribe those segments of the audio that included a surprisal word in a cloze test format. Results revealed that the majority of the learners could not transcribe the segments correctly, and found those words appearing in the caption useful for fostering listening.

(3) An experiment with L2 learners was conducted to check the effectiveness of data-driven and rule-based PSC on the recognition of specific points, where the two versions showed different words/phrases. Results suggest that the enhanced version included better clues to foster L2 listening recognition for specific segments.

(4) Experimental results revealed that the system was relatively successful to adapt itself to the demand of the L2 learner, which raised learner satisfaction on the resultant captions.

(5) Experimental results revealed that learners' scores significantly increased when they used these cues compared to having no hint.

The results were published in international journals that focuses on computer assisted language learning (CALL), language and speech processing, and in top domestic and international CALL conferences.

(Journal Articles)

1. M.S. Mirzaei, K. Meshgi, Y. Akita, T. Kawahara, "Partial and synchronized captioning: A new tool to assist learners in developing second language listening skill," ReCALL Journal, vol. 29, pp. 178-199, 2017. https://doi.org/10.1017/S0958344017000039
2. M.S. Mirzaei, K. Meshgi, T. Kawahara, "Exploiting Automatic Speech Recognition Errors to Enhance Partial and Synchronized Caption for Facilitating Second Language Listening," Computer Speech and Language Journal, vol. 49, pp. 17-36, 2018. https://doi.org/10.1016/j.csl.2017.11.001

(International Conferences)

1. M.S. Mirzaei, K. Meshgi, T. Kawahara, "Listening Difficulty Detection to Foster Second Language Listening with Partial and Synchronized Caption," in Proc. of EuroCALL'17: CALL in a climate of change: adapting to turbulent global conditions, pp. 211-216, Southampton, England, Aug 2017. https://doi.org/10.14705/rpnet.2017.eurocall2017.715
2. M.S. Mirzaei, K. Meshgi, T. Kawahara, "Detecting Listening Difficulty for Second Language Learners using Automatic Speech Recognition Errors," in Proc. of Interspeech 2017 Speech and Language Technology in Education (SLaTE) workshop, Stockholm, Sweden, pp. 156-160, Aug 2017. https://doi.org/10.21437/SLaTE.2017-27
3. K. Meshgi, M.S. Mirzaei, "A Comprehensive Word Difficulty Index for L2 Listening," in A. Botinis (Ed), Proc. of 9th Workshop on Experimental Linguistics (ExLing'18), Paris, France, pp. 81-84, Aug 2018.
4. K. Meshgi, M.S. Mirzaei, "A Data-driven Approach to Generate Partial and Synchronized Caption for Second Language Listeners," in Proc. of WorldCALL'18, Concepcion, Chile, Nov 2018.
5. M.S. Mirzaei, and K. Meshgi, "Automatic Scaffolding for L2 Listeners by Leveraging Natural Language Processing," in Proc. of EuroCALL'18: Future-proof CALL: language learning as exploration and encounters, Jyväskylä, Finland, pp. 200-206, Aug 2018. https://doi.org/10.14705/rpnet.2018.26.837
6. M.S. Mirzaei, and K. Meshgi, "Learner-Adaptive Partial and Synchronized Caption for L2 Listening Skill Development," In Proc. of EuroCALL'19: CALL and Complexity, Louvain-la-Neuve, Belgium, pp. 291-296, Aug 2019. https://doi.org/10.14705/rpnet.2019.38.1025
7. M.S. Mirzaei, K. Meshgi, and T. Nishida "Sentence Complexity as an Indicator of L2 Learner's Listening Difficulty," In Proc. of EuroCALL'20: CALL for Widening participation, Copenhagen, Denmark, Aug 2020.

(Domestic Conferences)

1. M.S. Mirzaei, and K. Meshgi, "Toward Adaptive Partial and Synchronized Caption to Facilitate L2 Listening," in Proc. of FLEAT VII, Tokyo, Japan, Aug 2019.

５．主な発表論文等

〔雑誌論文〕　計10件（うち査読付論文　10件／うち国際共著　10件／うちオープンアクセス　10件）

１．著者名 M.S. Mirzaei, K. Meshgi, Y. Akita, T. Kawahara	４．巻 29
２．論文標題 Partial and synchronized captioning: A new tool to assist learners in developing second language listening skill	５．発行年 2017年
３．雑誌名 ReCALL Journal	６．最初と最後の頁 178-199
掲載論文のDOI（デジタルオブジェクト識別子） https://doi.org/10.1017/S0958344017000039	査読の有無 有
オープンアクセス オープンアクセスとしている（また、その予定である）	国際共著 該当する

１．著者名 M.S. Mirzaei, K. Meshgi, T. Kawahara	４．巻 49
２．論文標題 Exploiting Automatic Speech Recognition Errors to Enhance Partial and Synchronized Caption for Facilitating Second Language Listening	５．発行年 2018年
３．雑誌名 Computer Speech and Language Journal	６．最初と最後の頁 17-36
掲載論文のDOI（デジタルオブジェクト識別子） https://doi.org/10.1016/j.csl.2017.11.001	査読の有無 有
オープンアクセス オープンアクセスとしている（また、その予定である）	国際共著 該当する

１．著者名 M.S. Mirzaei, K. Meshgi, T. Kawahara	４．巻 -
２．論文標題 Listening Difficulty Detection to Foster Second Language Listening with Partial and Synchronized Caption	５．発行年 2017年
３．雑誌名 EuroCALL'17: CALL in a climate of change: adapting to turbulent global conditions	６．最初と最後の頁 211-216
掲載論文のDOI（デジタルオブジェクト識別子） 10.14705/rpnet.2017.eurocall2017.715	査読の有無 有
オープンアクセス オープンアクセスとしている（また、その予定である）	国際共著 該当する

１．著者名 M.S. Mirzaei, K. Meshgi, T. Kawahara	４．巻 -
２．論文標題 Detecting Listening Difficulty for Second Language Learners using Automatic Speech Recognition Errors	５．発行年 2017年
３．雑誌名 Interspeech 2017 Speech and Language Technology in Education (SLaTE) workshop	６．最初と最後の頁 156-160
掲載論文のDOI（デジタルオブジェクト識別子） https://doi.org/10.21437/SLaTE.2017-27	査読の有無 有
オープンアクセス オープンアクセスとしている（また、その予定である）	国際共著 該当する

1．著者名 K. Meshgi, M.S. Mirzaei	4．巻 -
2．論文標題 A Comprehensive Word Difficulty Index for L2 Listening	5．発行年 2018年
3．雑誌名 9th Tutorial and Research Workshop on Experimental Linguistics (ExLing2018)	6．最初と最後の頁 81-84
掲載論文のDOI（デジタルオブジェクト識別子） なし	査読の有無 有
オープンアクセス オープンアクセスとしている（また、その予定である）	国際共著 該当する

1．著者名 K. Meshgi, M.S. Mirzaei	4．巻 -
2．論文標題 A Data-driven Approach to Generate Partial and Synchronized Caption for Second Language Listeners	5．発行年 2018年
3．雑誌名 WorldCALL'18	6．最初と最後の頁 TBD
掲載論文のDOI（デジタルオブジェクト識別子） TBD	査読の有無 有
オープンアクセス オープンアクセスとしている（また、その予定である）	国際共著 該当する

1．著者名 M.S. Mirzaei, K. Meshgi	4．巻 -
2．論文標題 Automatic Scaffolding for L2 Listeners by Leveraging Natural Language Processing	5．発行年 2018年
3．雑誌名 EuroCALL'18: Future-proof CALL: language learning as exploration and encounters	6．最初と最後の頁 200-206
掲載論文のDOI（デジタルオブジェクト識別子） https://doi.org/10.14705/rpnet.2018.26.837	査読の有無 有
オープンアクセス オープンアクセスとしている（また、その予定である）	国際共著 該当する

1．著者名 M.S. Mirzaei, K. Meshgi	4．巻 -
2．論文標題 Learner Adaptive Partial and Synchronized Caption for L2 Listening Skill Development	5．発行年 2019年
3．雑誌名 EuroCALL ' 19: Call and Complexity	6．最初と最後の頁 291-296
掲載論文のDOI（デジタルオブジェクト識別子） https://doi.org/10.14705/rpnet.2019.38.1025	査読の有無 有
オープンアクセス オープンアクセスとしている（また、その予定である）	国際共著 該当する

１．著者名 M.S. Mirzaei, K. Meshgi	４．巻 -
２．論文標題 Toward Adaptive Partial and Synchronized Caption to Facilitate L2 Listening	５．発行年 2019年
３．雑誌名 FLEAT VII: Language Learning with Technology Facing the Future	６．最初と最後の頁 TBD
掲載論文のDOI（デジタルオブジェクト識別子） なし	査読の有無 有
オープンアクセス オープンアクセスとしている（また、その予定である）	国際共著 該当する

１．著者名 M.S. Mirzaei, K. Meshgi	４．巻 -
２．論文標題 Sentence Complexity as an Indicator of L2 Learner's Listening Difficulty	５．発行年 2020年
３．雑誌名 EuroCALL'20: CALL for Widening participation	６．最初と最後の頁 TBD
掲載論文のDOI（デジタルオブジェクト識別子） TBD	査読の有無 有
オープンアクセス オープンアクセスとしている（また、その予定である）	国際共著 該当する

〔学会発表〕　計0件

〔図書〕　計0件

〔産業財産権〕

〔その他〕

―

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	ＭＩＲＺＡＥＩ　ＭＡＲＹＡＭ　ＳＡＤＡＴ (Mirzaei Maryam Sadat) (10810509)	国立研究開発法人理化学研究所・革新知能統合研究センター・特別研究員（ポストドクター） (82401)	