INTERSPEECH 2023
20-24 August 2023, Dublin, Ireland

Speaking Clearly, Understanding Better: Predicting the L2 Narrative Comprehension of Chinese Bilingual Kindergarten Children Based on Speech Intelligibility Using a Machine Learning Approach

Hiuching Hung[1], *Paula Andrea Pérez-Toro*[2], *Tomás Arias Vergara*[2],*Andreas Maier*[2], *Elmar Nöth*[2]

[1]Friedrich-Alexander Universität Erlangen-Nürnberg, Germany
[2]Pattern Recognition Lab, Friedrich-Alexander Universität Erlangen-Nürnberg, Germany

hiuching.hung@fau.de, paula.andrea.perez@fau.de, tomas.arias@fau.de,
andreas.maier@fau.de,elmar.noeth@fau.de

Abstract

This study investigated the relationship of speech intelligibility and the narrative comprehension among bilingual kindergarten children and how well the speech intelligibility of second language (L2) predicted the L2 narrative comprehension using a machine learning approach. Fifty Chinese-English bilingual children aged 5-6 years old participated in this study by taking a narrative comprehension test. Their L2 narrative comprehension was assessed using the MAIN test. The speech intelligibility was assessed in terms of twenty-four features that encode confidence levels with respect to phoneme and word classifiers trained on native speaker speech data.

Our hypothesis posits that it is possible to predict L2 narrative comprehension based on speech intelligibility features. By using seven out of the twenty-four considered features we were able to make predictions of the MAIN test scores with an RMSE of 2.13 and a Pearson correlation coefficient of 0.468 based on a data set of 50 bilingual kindergarten children. We conclude the paper by providing pedagogical implications for second language teaching as well as suggestions for future work.
Index Terms: bilingual kindergarten children, speech intelligibility, L2 narrative comprehension, speech recognition, human-computer interaction, computational paralinguistics

1. Introduction

Ever since the beginning of the language learning journey, children are introduced to narrative stories. By 3 years of age, a typically developing preschooler has already developed a battery of cognitive and language skills to comprehend and answer questions of oral narratives that often accompanied by wordless pictures [1]. These skills are reported closely linked to the literacy and academic performance of children at an older age.

In the realm of second language education, examining the narrative comprehension of bilingual children is a more complex endeavor due to a multitude of combinations of L1 and L2 and the asynchronous trajectory of these languages. Although influencing factors such as age and input-output have been reported in the recent years, a more comprehensive understanding of the bilingual narrative abilities is imperative given the rapidly growing population of bilingual children and the resulting demand for the enhanced bilingual educational quality. Accordingly, further research that promotes the development of narrative comprehension for bilingual children is much needed.

Several factors that affect children's L2 narrative comprehension, apart from the above-mentioned age and input-output, have also been found to impact their speech intelligibility. Speech intelligibility provides essential information on how well the speech is understood by others and therefore has been a well-established component of language assessments and language therapies. Despite extensive research on both L2 narrative comprehension and L2 speech intelligibility in young bilingual children respectively, to date, no clear correlation between the two has been identified.

Given the significant overlap in the factors that influence both, we hypothesize that there is a correlation between narrative comprehension and intelligibility.

The contribution of this study is twofold. Firstly, we introduce a phonological dimension to the bilingual language assessment with respect to narrative comprehension for young children. Secondly, we explore the potential of a machine learning approach to measure speech intelligibility of children, which provides implications for the ASR-based technologies, particularly in the nonnative related topics.

2. Related work

Children's level of narrative comprehension is generally defined as the ability to comprehend and construct meanings of the oral narratives, usually in the form of stories with a sequence of wordless pictures [1, 2]. Narrative comprehension has been reported in multiple studies to be a robust predictor of children's academic achievement, especially in literacy [3, 4]. Comprehending narratives is a complex meaning-making process that requires linguistic, cognitive and other skills [5, 6]. In the field of second language acquisition, measuring narrative comprehension is more complicated because this process often involves factors such as cultural backgrounds, contextual differences, and children's L2 language competency, which, in most cases, is lower than that of the L1.

Recent research on narrative comprehension of young bilingual children has mainly focused on two aspects: comparing monolingual and bilingual children and comparing L1 and L2. However, although influencing factors such as age [7, 8], language exposure and narrative task [9, 10] have been continuously identified, as yet, little is known about the relationships between speech intelligibility and L2 narrative comprehension. The causes could be due to: (i) Lack of bilingual language measurements. There are only a limited number of standardized language assessment tools available for bilingual children, and in some languages, there are none at all [11]; and (ii) the high cost of human raters. Assessing the narrative comprehension or speech intelligibility typically involves human raters. As many advantages as human evaluation process has, such measurement process is not only labor-intensive, but also easily prone to bias due to various accents of the raters or the sound quality caused by the different recording settings [12, 13].

10.21437/Interspeech.2023-2057

3. Materials and methods

3.1. Data Set

The data set used in this article comes from the first author's dissertation study on the compilation of a nonnative children speech corpora [14]. The data was collected in 2022 from an English-immersion bilingual kindergarten in China. Fifty sequential bilingual children (26 boys and 24 girls), aged between 5-to 7-years-old (mean age = 5 years, 10 months [5;8], SD = 6 months, range =5;0–6;8) and five teachers participated in this study. All participants are native Chinese. Due to the restriction policies during the pandemic, the data was collected remotely with ZOOM and OBS (Open Broadcaster Software).

3.2. Material

This study used MAIN (Multilingual Instrument for Narratives, 2019 version) [11] to test children's L2 narrative comprehension. MAIN is a test instrument designed to assess the narrative skills of bilingual children. MAIN is suitable for our test because it has a uniform set of testing procedure, including modal stories, questions and quasi-standardized answers. Additionally, it is considered well-tested globally and culturally appropriate.

3.3. Data processing

Traditionally, phoneme assessments are conducted by human raters. Human raters can easily discern and track the voice of tested speakers, even in a slightly noisy environment or situations where speakers are talking simultaneously. In our study, it is imperative to eliminate or reduce crosstalk between different microphone signals to ensure that the speech signals solely come from one speaker in each recording. Our solution to tackle this issue involves an adaptive noise gate effect that is able to distinguish between desired speech and undesirable crosstalk. This is achieved by deriving speech activity thresholds dynamically based on short-term averaged powers of both signals. For example, if a segment of the teacher's audio recording exhibits high power, it would raise the child speech activity threshold and vice versa. This approach ensures that even stronger crosstalk will be effectively suppressed, while maintaining optimal sensitivity towards the desired speech signal.

3.4. Speech intelligibility

Phonemic-level features:

A multilabel recurrent network with Long-Short Term Memory cells (LSTM) is used for the automatic recognition of phonemes [**AnonRef**]. The model was trained to detect three main phonemic dimensions: (1) manner of articulation (stop, nasal, lateral, trill, fricative, approximant, and vowels), (2) place of articulation (labial, alveolar, velar, palatal, postalveolar), and (3) voicing (voiced, and voiceless). The architecture of the network is as follows: Two convolution layers process the input tensors (Mel-spectrograms) with ReLU activation functions, two max-pooling layers, and dropout. The resulting feature maps are concatenated to form the sequence of feature vectors processed by two stacked bidirectional LSTMs. Then, a sigmoid activation function is used to compute the sequence of phoneme posterior probabilities. The network was trained with the TIMIT corpus, a dataset with time-aligned phonetic transcriptions of speech recordings from 630 American English native speakers [15].

We use this network to measure the phonemic precision by computing the average maximum posterior probability (MaxPost) of each phonemic class. MaxPost can take values between "0" and "1", where a value of "1" represents high "confidence" from the network to predict a phoneme class from the recording. Furthermore, since the network was trained with English native speakers, we hypothesized that MaxPost can quantify the phonemic precision of L2 children. Figure 1 shows the values of MaxPost computed for two groups: children with a MAIN score less than 5 and the rest.

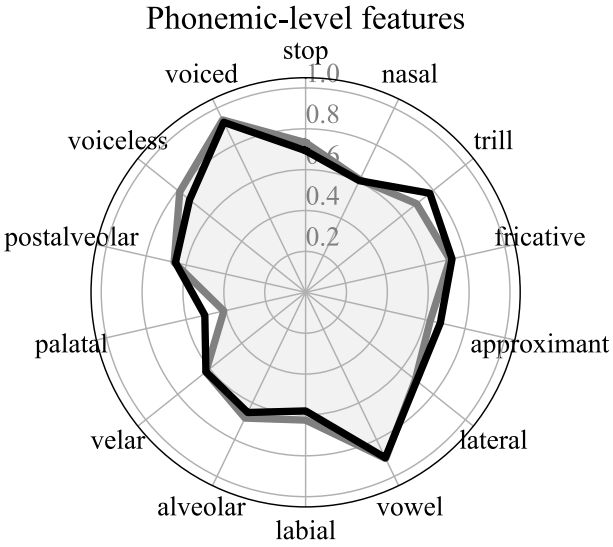

Figure 1: *Radar plots of the phonemic-level features computed for L2 children. The shaded region represents children with a MAIN score less than 5 and the black solid line represents the rest.*

Word-level features:

A grammatical tagging algorithm was applied to the transcriptions to extract the corresponding Part of Speech (POS) categories. Then, the categories were standardized from *PennTreebank tagset* [16] to the framework *Universal Dependencies* [17]. In this study, we considered a subset of these categories (see Table 1) in order to simplify the analysis of the distribution of words in the text. Finally, the confidence values produced by the Automatic Speech Recognition (ASR) system were averaged based on individual word categories and forming a 10-dimensional vector per participant. Subsequently, this was used for regression analysis.

Figure 2 shows the confident values for the different word-level features computed for two groups: children with a MAIN score less than 5 and the rest.

3.5. Regression analysis

We trained a linear Support Vector Regressor (SVR) to predict the MAIN score (Section 3.1) using the speech intelligibility features described in Section 3.4. The margin C and the ϵ-insensitive tube parameters were optimized through a grid search with $2^{-7} < C < 2^{3}$ and $2^{-7} < \epsilon < 2^{3}$ using a combination of Leave-One-Spekaer-Out (LOSO) and nested 10-fold cross-validation strategies, i.e, the cross-validation is performed in every iteration of LOSO. For testing the SVR, we compute the median of the best C and ϵ obtained during training and perform LOSO again, this time with fixed parameters. The performance of the regressor is evaluated with Root Mean Squared

Table 1: *Word categories used in this study*

Category	Description ¡
Nouns	Denote a person, animal, place or object
Verbs	Describe an event or action. Modal verbs are included in this category
Adverbs	Modify verbs in terms of place, time or direction
Interjections	Used as an exclamation or part of an exclamation
Proper nouns	Denote a specific individual, place, or object
Adjectives	Typically modify nouns
Coordinating conjunction	Connects words, phrases, and clauses that are coordinates, or equal to each other
Adposition	It encloses prepositions and postpositions
Determiner	Modify a noun or a noun phrase and convey the meaning of the noun phrase within the given context
Pronouns	Words that substitute for nouns, whose meaning is recoverable from the linguistic

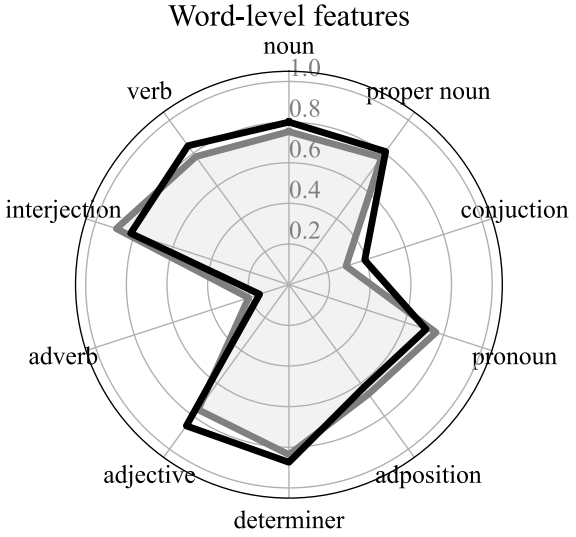

Figure 2: *Radar plots of the word-level features computed for L2 children. The shaded region represents children with a MAIN score less than 5 and the black solid line represents the rest.*

Error (RMSE) and Pearson's correlation coefficient (ρ). We opted to use a linear SVR to avoid overfitting due to the limited amount of subjects available.

3.6. Feature importance

We ranked the feature intelligibility features (14 phonemic/10 word-level) using *permutation feature importance*, an inspection technique used to evaluate the dependence of a model (e.g., the SVR) on a set of features[1]. In this study, the importance of a feature is defined as the increase in the RMSE when a single feature is randomly shuffled between samples; thus, breaking the relationship between feature and target.

[1] https://scikit-learn.org/stable/modules/permutation_importance.html

4. Experiments and results

To analyze the L2 speech intelligibility of the bilingual children, we ranked the phonemic- and word-level features using the linear SVR and feature importance analysis. The features were ranked during training according to the RMSE between predicted and target MAIN score. Figure 3 shows the final ranking of features. Table 2 reports the performance results of the

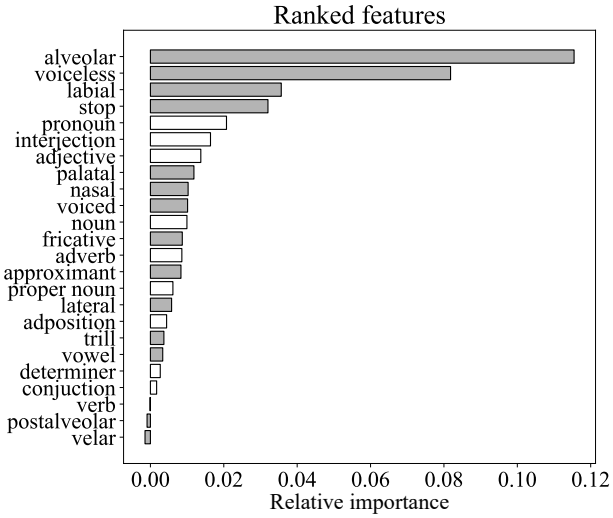

Figure 3: *Ranking of speech intelligibility features at the phonemic (grey bars) and word (white bars) levels.*

SVR tested with different ranked feature sets. The highest performance was achieved when the top seven features (RMSE: 2.13, ρ : 0.468) namely alveolar, voiceless, labial, stop, pronoun, interjection, and adjective, were used for the prediction.

5. Discussion

According to the feature importance analysis, the four highest-ranked features are all phoneme-level features: alveolar, voiceless, labial and stop. This suggests that the phonemic features play a more crucial role than word-level features in predicting bilingual children's L2 narrative comprehension.

We evaluated the performance of different feature subsets based on the feature ranking and found that using the top seven features in the ranking yielded the best prediction of the L2 narrative comprehension with an RMSE of 2.13 and a ρ of 0.468. But even the first three phonemic-level features would yield a slightly higher RMSE of 2.19 and a slightly lower ρ of 0.415.

The results of the phonemic-level and word-level feature extraction and analysis in section 3.4 show slight differences in intelligibility between two groups of children (MAIN scores less than 5 and MAIN scores equal or even higher than 5). Specifically, there are intelligibility differences in the palatal, voiceless, and trill of the phonemic-level features, and the adjective, conjunction, and interjection of the word-level features. This suggests that these features may be particularly important in predicting a child's L2 narrative comprehension.

The different significance of the features were also analyzed using permutation feature importance which generated a ranked

Table 2: *Regression performance with different subsets of the ranked features.* ***RMSE****: Root Mean Squared Error.* $\boldsymbol{\rho}$*:* ***Pearson's correlation coefficient.*** ***p-value****: p-value from* ρ*.*

# Features	RMSE	ρ	p-value
1	2.29	0.348	0.014
2	2.24	0.350	0.014
3	2.19	0.415	0.003
4	2.19	0.402	0.004
5	2.18	0.419	0.003
6	2.19	0.416	0.003
7	**2.13**	**0.468**	**0.001**
8	2.16	0.429	0.002
9	2.18	0.411	0.003
10	2.15	0.450	0.001
11	2.17	0.418	0.003
12	2.18	0.411	0.003
13	2.18	0.407	0.004
14	2.19	0.393	0.005
15	2.21	0.367	0.010
16	2.23	0.347	0.015
17	2.23	0.340	0.017
18	2.26	0.306	0.032
19	2.27	0.292	0.042
20	2.29	0.268	0.063
21	2.31	0.240	0.097
22	2.32	0.226	0.118
23	2.31	0.249	0.085
24	2.31	0.241	0.095

feature list (see Figure 3). The results confirmed our previous hypothesis about the importance of palatal and voiceless phonemic-level features and interjection word-level features.

Discussion of pedagogical implications:

Our feature analysis showed that different phonemes have different predictive powers on children's L2 narrative comprehension. By paying attention to the high-ranking phonemes in children's oral speeches, teachers might be able to gauge a child's narrative comprehension level with greater accuracy. This is especially valuable since narrative assessment tools for bilingual children are sometimes lacking or administrating such tests is too labor-intensive. Another interesting phenomenon is the use of interjections. Although some researchers argue that interjection should be taught at the early stages of language learning, due to its unique linguistic nature, interjection has been poorly researched and are therefore rarely taught in a systematic way [18, 19]. Notably, our feature importance analysis revealed that interjections were not only used by children with elementary English proficiency, but were also related to children's L2 narrative comprehension. This finding may shed some light on the interjection teaching.

It is worth noting that the first four highest-ranking phonemes in Figure 3 shared one common trait: they all have similar counterparts in the Chinese phonological system. Future work could investigate whether teaching these phonemes first facilitates children's narrative skills.

Limitations:

This study is subject to several known limitations. Firstly, due to the scarcity of children's speech corpora, we used tools such as ASR and phoneme detectors that were built based on adults speech data sets. Secondly, the intermittent closures of kindergartens in China during the COVID-19 pandemic, which spanned nearly a year, caused a significant impact on the pedagogical benefits of the English-immersion concepts offered in this kindergarten.

6. Conclusions

This study employed machine learning techniques and automatic speech recognition to make prediction of the language performance of bilingual kindergarten children with respect to L2 narrative comprehension. Our results demonstrate that phonemic-level and word-level intelligibility features can be used to accurately predict the children's performance on the MAIN test, highlighting the potential utility of automatic speech recognition and machine learning in the field of language assessment. These findings have important implications for clinicians and educators seeking to monitor and assess the language development of bilingual children.

Future research should continue to explore additional features, such as prosodic features, and incorporate teacher speech data to further enhance the prediction accuracy. Overall, this study represents an important step towards the development of more objective and reliable tools for assessing language performance in bilingual children.

7. Acknowledgements

We would like to express our appreciation to Professor Thorsten Piske from Fachdidaktiken Department for his valuable suggestions, the Pattern Recognition Lab of Friedrich-Alexander-University, Nürnberg-Erlangen, and the children, families and the Hubin Dadi Kindergarten in Quanzhou, China. We thank Dr. Sebastian Gesemann for preprocessing the audio data to reduce the noise and crosstalk between the microphone signals.

8. References

[1] A. H. Paris and S. G. Paris, "Assessing narrative comprehension in young children," *Reading Research Quarterly*, vol. 38, no. 1, pp. 36–76, 2003.

[2] P. Van Den Broek, P. Kendeou, K. Kremer, J. Lynch, J. Butler, M. J. White, and E. P. Lorch, "Assessment of comprehension abilities in young children," in *Children's reading comprehension and assessment*. Routledge, 2005, pp. 125–148.

[3] M. Silva and K. Cain, "The relations between lower and higher level comprehension skills and their role in prediction of early reading comprehension." *Journal of Educational Psychology*, vol. 107, no. 2, p. 321, 2015.

[4] T. Horowitz-Kraus, K. Eaton, R. Farah, A. Hajinazarian, J. Vannest, and S. K. Holland, "Predicting better performance on a college preparedness test from narrative comprehension at the age of 6 years: An fmri study," *Brain research*, vol. 1629, pp. 54–62, 2015.

[5] Y.-S. G. Kim, "Direct and mediated effects of language and cognitive skills on comprehension of oral narrative texts (listening comprehension) for children," *Journal of experimental child psychology*, vol. 141, pp. 101–120, 2016.

[6] A. Potocki, J. Ecalle, and A. Magnan, "Narrative comprehension skills in 5-year-old children: Correlational analysis and comprehender profiles," *the Journal of Educational research*, vol. 106, no. 1, pp. 14–26, 2013.

[7] U. Bohnacker, "Tell me a story in english or swedish: Narrative production and comprehension in bilingual preschoolers and first graders," *Applied Psycholinguistics*, vol. 37, no. 1, pp. 19–48, 2016.

[8] M. Roch, E. Florit, and C. Levorato, "Narrative competence of italian–english bilingual children between 5 and 7 years," *Applied Psycholinguistics*, vol. 37, no. 1, pp. 49–67, 2016.

[9] S. Kunnari and T. Välimaa, "Narrative comprehension in simultaneously bilingual finnish-swedish and monolingual finnish children," *Developing Narrative Comprehension: Multilingual Assessment Instrument for Narratives*, vol. 61, p. 149, 2020.

[10] J. Lindgren and U. Bohnacker, "How do age, language, narrative task, language proficiency and exposure affect narrative macrostructure in german-swedish bilingual children aged 4 to 6?" *Linguistic Approaches to Bilingualism*, vol. 12, no. 4, pp. 479–508, 2022.

[11] N. V. Gagarina, D. Klop, S. Kunnari, K. Tantele, T. Välimaa, I. Balčiūnienė, U. Bohnacker, and J. Walters, "Main: Multilingual assessment instrument for narratives," *ZAS papers in linguistics*, vol. 56, pp. 155–155, 2012.

[12] B. H. Huang and R. Ramírez, "Research methods for evaluating second language speech production," in *Research Methods for Understanding Child Second Language Development*. Routledge, 2022, pp. 84–101.

[13] A. A. Lopez, S. Turkan, and D. Guzman-Orth, "Conceptualizing the use of translanguaging in initial content assessments for newly arrived emergent bilingual students," *ETS Research Report Series*, vol. 2017, no. 1, pp. 1–12, 2017.

[14] H. Hung, A. Maier, and T. Piske, "Building a non-native speech corpus featuring chinese-english bilingual children: Compilation and rationale," 2023.

[15] J. S. Garofolo, L. Lamel, W. M. Fisher *et al.*, "Timit acoustic phonetic continuous speech corpus," *Linguistic Data Consortium*, 1993.

[16] A. Taylor, M. Marcus, and B. Santorini, "The penn treebank: an overview," *Treebanks: Building and using parsed corpora*, pp. 5–22, 2003.

[17] M.-C. De Marneffe, C. D. Manning, J. Nivre, and D. Zeman, "Universal dependencies," *Computational linguistics*, vol. 47, no. 2, pp. 255–308, 2021.

[18] K. Petrova, "Interjections and l2 learning and teaching," *Linguística: Revista de Estudos Linguísticos da Universidade do Porto*, pp. 323–336, 2020.

[19] J. Bland, "Teaching english to young learners: More teacher education and more children's literature!." *Online Submission*, vol. 7, no. 2, pp. 79–103, 2019.