INTERSPEECH 2021
30 August – 3 September, 2021, Brno, Czechia

A Preliminary Study on Discourse Prosody Encoding in L1 and L2 English Spontaneous Narratives

Yuqing Zhang[1], *Zhu Li*[1], *Binghuai Lin*[2], *Jinsong Zhang*[1]

[1]Beijing Language and Culture University, China
[2]Smart Platform Product Department, Tencent Technology Co., Ltd, China
{yuqingelsa, lzblcu19}@gmail.com, jinsong.zhang@blcu.edu.cn, binghuailin@tencent.com

Abstract

Relatively little attention has been devoted to the discourse-level prosodic encoding and speech planning in second language (L2) speech. This study reports a preliminary study on learners' discourse prosody encoding pattern and makes a comparison with that of native speakers. Using a corpus of spontaneously produced picture story narratives, we analyzed general characteristics of prosodic units (PUs) and explored relationships between pitch encoding (cross-boundary f0 heights and f0 slopes) of PUs and the semantic completeness of PUs in English spontaneous speech by native speakers, beginning learners, and advanced learners. The results indicated that beginning learners showed neither sensitivity to semantic units in discourse (DUs) in their f0 encoding nor distinct signs of pitch-related preplanning based on DUs, suggesting improper phrasing of the least proficient non-native speakers. Both native speakers and advanced learners were sensitive to the initiation and termination of DUs in their prosodic encoding; however, only native speakers showed clear signs of DU-based preplanning. We argue that the observed between-group differences in L1 and L2 speech might be attributed to differences in the scope of speech planning, i.e., compared with native speakers, who mostly produce complete semantic units, learners' speech is produced step by step with pauses between phrases.

Index Terms: spontaneous narratives, speech planning

1. Introduction

Prosodic aspects of non-native speech have received a large amount of attention in phonetic and psycho-linguistic research, and most researchers agree that L2 prosody deviates to some degree from the native norm, notably in pitch variation [1], pitch range [2] and rhythm [3]. In addition, it is generally recognized that learners' organization of discourse prosody differs from that of native speakers, and those differences are likely to have an impact on comprehensibility [4]. Studies focusing on discourse-level prosody have examined L2 speech from the perspective of speech rate, locations of discourse boundary breaks as well as size and scope of speech planning and chunking units. For instance, using a corpus of multilingual recordings of a standard text, [5] demonstrated that L2 English was characterized by slower speech rate, and therefore lower information density. [6] suggested that L1 and L2 speech differed in the distribution of prosodic break levels and break locations, and attributed this to between-group differences in the size and strategy of discourse-level speech planning (i.e., L2 speakers use smaller scope of speech chunking and fewer large-scale planning units). Prior work also demonstrated that in L2 speech, realization of prosodic cues to discourse structure (e.g., use of intonation to signal topic change) in continuous speech correlated with language proficiency [7].

Here, we extend this line of research and analyze whether the pitch encoding (initial f0 height, the final f0 height, and the f0 slope) of a PU varies as a function of its semantic completeness in L1 and L2 spontaneous speech, and attempt to answer the following questions: 1) whether the PUs produced by native speakers and L2 learners of different proficiency levels show different characteristics, 2) whether native speakers and L2 learners of different proficiency levels are sensitive to semantic units in discourse (DUs) in their prosodic encoding, and 3) whether they are capable of DU-based pitch-related preplanning in spontaneous speech production. Analyses of the relationship between prosodic encoding and the semantic structure of the speech segments may shed light on the mechanisms of speech preplanning and intonation acquisition.

1.1. Sensitivity to the DU

Following prior work [8, 9], the semantically-defined unit is referred to as "semantic unit in discourse", which is often structurally encoded by a clause and considered a basic semantic unit in our daily interactions; the perceptually-defined unit is referred to as "prosodic unit". The semantic completeness of the PU is defined based on its co-extensiveness (i.e., its match/mismatch of left/right alignment) with a DU.

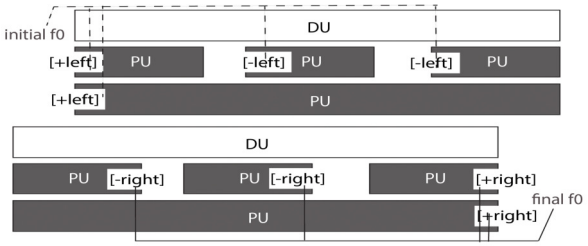

Figure 1: *A graphic illustration of research question 2.*

The analysis of learners' sensitivity to DUs is based on the assumption that if speakers are sensitive to the initiation of a DU, the initial f0 encoding of the PU that starts a DU is expected to differ strongly from the initial f0 of PUs starting in the latter part of the DU. Therefore, the initial f0 height of the PU is expected to positively correlate with the PU-DU left alignment (cf. the top part of Figure 1). Similarly, if speakers show sensitivity to the termination of a DU, the final f0 encoding of the PU that ends a DU is expected to to differ significantly from that of those PUs terminating in the middle of the DU. Thus the final prosodic encoding of the PU is expected to negatively correlate with the PU-DU right alignment (cf. the bottom part of Figure 1). Consequently, the overall f0 slope is expected to be steeper if more sides of the PU are aligned with the DU. The for-

http://dx.doi.org/10.21437/Interspeech.2021-1082

mer relationship indicates that speakers are aware of the initiation of a semantic unit at the onset of the PU articulation, which is reflected in the higher initial f0 height in speech production. The latter relationship indicates that speakers are aware of the termination of a semantic unit when arranging the prosodic contour, which is prosodically marked by the lower final f0 height in speech production. The match/mismatch of PU-DU boundaries is consistently marked by varying prosodic encoding, indicating speakers' sensitivity to semantic units.

1.2. DU-based pitch-related preplanning

Studies on pitch-related preplanning have established the existence of a significant relationship between phrasal f0 parameters (i.e., initial f0 heights, final f0 heights, and f0 declination slopes) and the length of intonation phrases (IPs) in read speech [10,11], scripted speech [12], and spontaneous speech [8,13]. In addition, [8] reported that semantic completeness of the speech segment was strongly associated with its prosodic parameters, which was taken as evidence for speakers' capability of DU-based pitch-related preplanning.

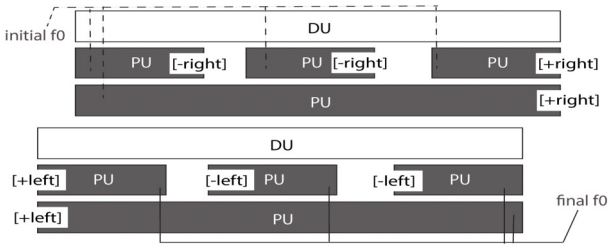

Figure 2: *A graphic illustration of research question 3.*

Based on [8]'s study, this study hypothesizes that if speakers show signs of DU-based pitch-related preplanning, the initial f0 height of PUs is expected to negatively correlate with the PU-DU right alignment (cf. the top part of Figure 2) and the final f0 height is expected to positively correlate with the left alignment (cf. the bottom part of Figure 2). The former relation suggests that speakers provide anticipatory prosodic cues for the semantic completeness of the subsequent unit at the speech onset. The latter relation suggests that speakers are aware of the beginning of a semantic unit leaving a carryover prosodic cue at the end of articulation. Namely, for PUs which terminate a DU, there will be an overall f0-lowering resulting from the PU-termination anticipatory effect. And for PUs which initiate a DU, there will be an overall f0-raising due to the PU-initiation carryover effect if speakers are capable of DU-based pitch-related preplanning.

2. Materials and method

2.1. Corpus of spontaneous narratives

Speech recordings were taken from the ALLSSTAR Corpus [14], which consists of scripted and spontaneous speech produced by native English speakers and L2 learners. The subset of corpus selected in the present study comprised L1 and L2 English spontaneous narratives elicited by two cartoon prompts "Bird's New Hat" [15] and "Bubble Bubble" [16]. Word and phoneme boundaries have been provided in the time aligned textgrids [17]. Since there were limited data available for each language, we gathered speech produced by learners from different language backgrounds. L2 learners were then divided into two groups by their average pronunciation and fluency scores. Information of talkers selected here is provided in table 1.

Table 1: *Information for talkers in the ST1&ST2 sub-corpus.*

Group	Speakers	Average Age	Proficiency scores
L1_ENG	12 English	20.2 ± 2.0	–
L2_expert	5 Turkish, 5 Mandarin, 1 Russian, 1 Portuguese	24.1 ± 3.2	71.5 ± 6.4
L2_novice	2 Turkish, 5 Mandarin, 1 Vietnamese, 2 Korean, 1 Portuguese, 1 Japanese	24.6 ± 2.5	53.8 ± 5.6

2.2. PU and DU annotation

Prosodic unit boundaries were annotated based on the pitch variation and timing patterns perceived from the audio signals, i.e., a clear perceivable pitch reset, lengthening of the final syllables in the previous segment and the shortening of the initial syllables in the subsequent segment, and noticeable pauses.

DUs have been studied as basic analytic units in research on cross-linguistic comparative analysis of spontaneous speech segmentation [18]. Technically, a DU was characterized as a part of an utterance including a predicate and the predicate's key arguments, and was structurally encoded as a clause. The main predicate was the semantic core of the DU and annotators used it as a clue to the boundaries of a DU.

Figure 3 provides a graphic illustration of possible PUs and PU-DU alignment conditions. The first PU is left-aligned with a DU but not right-aligned, i.e., in the match of left alignment, and mismatch of right alignment condition. PU and DU annotation was performed by two students majoring in linguistics, and annotation points at which two annotators expressed disagreement were removed from final analysis. We also excluded those PUs which were abandoned by the speakers in the ensuing speech or too short to extract reliable f0 values.

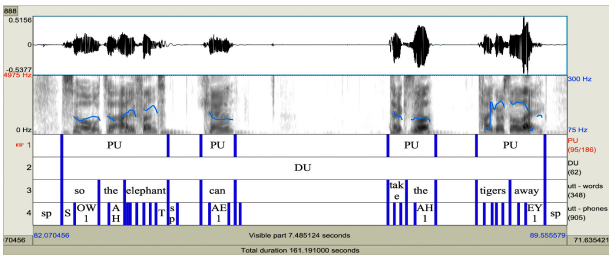

Figure 3: *An example of PU and DU annotation.*

2.3. Prosodic parameter extraction

F0 extraction was performed using the auto-correlation method with recommended pitch ranges (75-300 Hz for male) through the Parselmouth API of Praat [19, 20]. The raw F0 values were then normalized to semitones with the formula (1), where the f0 base was computed as the 5th percentile of all f0 values of a speaker.

$$\text{Semitone} = 12 * \log_2 \left(\frac{F_0}{F_0 \text{ base}} \right) \quad (1)$$

Raw f0 values in Hertz were transformed into the logarithmic semitone scale to minimize individual differences. To measure f0 declination, the raw f0 points were linearly interpolated and then smoothed by passing through a Sabitzky-Golay filter with a third-order polynomial and a 5-sample window [21]. After that, a linear regression line was fitted to the continuous f0

points of each PU using the least-squares method. The slope term of the fitted lines was extracted to describe the slope of f0 declination [12]. The highest f0 value of the first syllable and the lowest value of the last syllable of each PU were extracted to represent the PU initial and the PU final f0 heights respectively.

2.4. Statistical analysis

Linear mixed effect models [22, 23] were fitted on the PU-initial, PU-final f0 heights and the slope of the f0 declination. As previous studies have demonstrated the strong correlation between utterance length and f0 slope, we included PU length as the control variable. Speakers were treated as random factors.

3. Results and discussion

3.1. General characteristics of PUs

Table 2 summarizes the distribution of PUs for the three groups according to the four PU-DU alignment conditions. Figure 4 shows speech rate and planning unit size comparisons of the three groups. Here speaking rate was computed at the PU level, i.e., average syllable number per second in a PU. Planning unit size was calculated as the number of orthographic syllables in a PU. The average pitch range is shown in Figure 5.

Table 2: *PUs' distributions according to alignment conditions.*

PU-DU Alignment		N			%		
Left	Right	L1_ENG	L2_expert	L2_novice	L1_ENG	L2_expert	L2_novice
-	-	34	152	252	0.08	0.25	0.39
-	+	108	199	171	0.25	0.33	0.27
+	-	112	201	202	0.26	0.33	0.32
+	+	185	57	15	0.42	0.09	0.02

It is noticeable that the three groups differ in terms of the patterns of distribution. Around 42% of the PUs align with DUs on both sides in L1 English speech, whereas only 9% of the PUs span a whole DU in L2 speech by advanced learners, and less than 5% in L2 speech by beginning learners. About 92% of the PUs in L1 speech are aligned with DUs on at least one side, while the number of PU-DU alignment on at least one side is comparatively smaller in L2 spontaneous speech (75% in speech by L2_expert and 61% in L2_novice).

3.1.1. Speaking rate

As shown in Figure 4, learners' average speaking rate is slower than native speakers. This observation is consistent with previous studies on articulation rate (i.e., syllables/second excluding silent pauses) in L2 speech [5,6]. In addition, as language proficiency level increases, learners' speaking rate tends to be faster and more native-like.

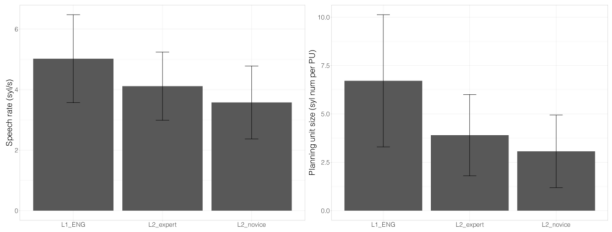

Figure 4: *The left subplot shows speaking rate (y-axis) plotted against speaker groups (x-axis). The right subplot shows planning unit size (y-axis) plotted against speaker groups (x-axis).*

3.1.2. Planning unit size

As suggested in Figure 4, the size of each PU produced by native speakers is larger and almost twice as large as the average PU size produced by language learners. In addition, advanced learners seem to be able to plan speech on a larger scale compared to beginning learners.

3.1.3. Pitch range

Figure 5 shows that the PUs produced by learners have narrower pitch range, suggesting less amount of pitch variation in L2 speech. This observation echos prior work on pitch characteristics of L2 speech [24]. Moreover, PUs of advanced learners have slightly larger pitch range than those produced by beginning learners.

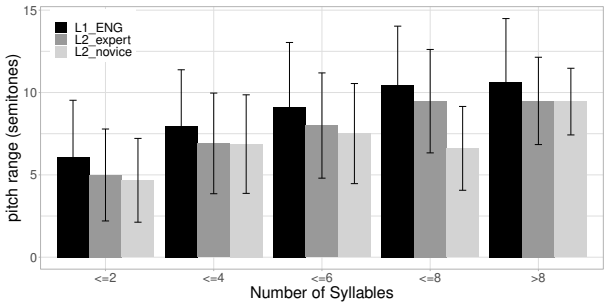

Figure 5: *The average pitch range in L1 and L2 for different sizes of PUs (e.g., $<= 2$ in the x-axis means the number of syllables in the PU is less than or equal to 2).*

3.2. Sensitivity to the DU

Table 3 summarizes the model parameters for the three groups. It can be seen that after controlling the effect of PUL_{ength}, a significant relationship between $PUI_{nitial}F0$ and L_{eft} alignment and a strong correlation between $PUF_{inal}F0$ and R_{ight} alignment exist in English spontaneous narratives by native speakers, and in L2 speech by the proficient learners as well. There is a significant initial f0 up-shifting in left-aligned PUs (Figure 6), compared to those not left-aligned ($\beta = 1.67, t = 5.74, p < 0.001$ in L1_ENG; $\beta = 1.03, t = 4.57, p < 0.001$ in L2_expert), and a significant final f0 down-shifting in right-aligned PUs (Figure 7), in comparison with not right-aligned ones ($\beta = -1.31, t = -5.22, p < 0.001$ in L1_ENG; $\beta = -1.43, t = -7.06, p < 0.001$ in L2_expert). With regard to f0 declination slopes (Figure 6, Figure 7), both PU-DU left alignment and right alignment will introduce a steeper negative f0 declination slope of the PU in native English spontaneous speech (L_{eft}: $\beta = -1.46, t = -3.67, p < 0.001$, R_{ight}: $\beta = -1.17, t = -2.84, p < 0.01$ in L1_ENG). However, no evidence can be found in the PU slope shifting in the negative direction in relation to PU-DU alignment in L2 speech.

It is noticeable that for the least proficient L2 learners in our study, no significant relationship has been found between $PUI_{nitial}F0$ and L_{eft} alignment, $PUF_{inal}F0$ and R_{ight} alignment, and f0 slope with respect to the two alignment conditions. A possible explanation is that for less proficient learners, their speech planning strategies differ from those adopted by native speakers, in that they may use smaller planning or chunking units, which cover incomplete semantic units in spontaneous speech production. As a result, less proficient learners produce

Table 3: *Summary of the linear mixed effect models*

		L1_ENG				L2_expert				L2_novice			
	PU-DU Alignment	β	SE	t value	Pr(>|t|)	β	SE	t value	Pr(>|t|)	β	SE	t value	Pr(>|t|)
PU initial f0	L_{eft}	1.67	0.29	5.74	1.80e-08 ***	1.03	0.23	4.57	5.96e-06 ***	0.43	0.27	1.56	0.12
	R_{ight}	-0.66	0.30	-2.18	0.0295 *	-0.36	0.23	-1.58	0.11	-0.51	0.29	-1.77	0.0768
PU final f0	L_{eft}	0.57	0.24	2.33	0.0205 *	0.19	0.20	0.94	0.35	0.54	0.25	2.15	0.0322 *
	R_{ight}	-1.31	0.25	-5.22	2.76e-07 ***	-1.43	0.20	-7.06	4.66e-12 ***	-0.32	0.26	-1.22	0.22
F0 slope	L_{eft}	-1.46	0.40	-3.67	0.00027 ***	-0.06	0.56	-0.11	0.91	0.86	0.61	1.40	0.16
	R_{ight}	-1.17	0.41	-2.84	0.004775 **	-0.55	0.57	-0.96	0.34	1.05	0.64	1.63	0.10

Notes: *** p < 0.001; ** p < 0.01; * p < 0.05.

these small units step by step, without paying much attention to using pitch shifting to signal the start and the end of a DU. This observation is in line with previous findings [25] and provides supporting evidence that beginning speakers show limited capacity to plan on a large scale compared with L1 speakers.

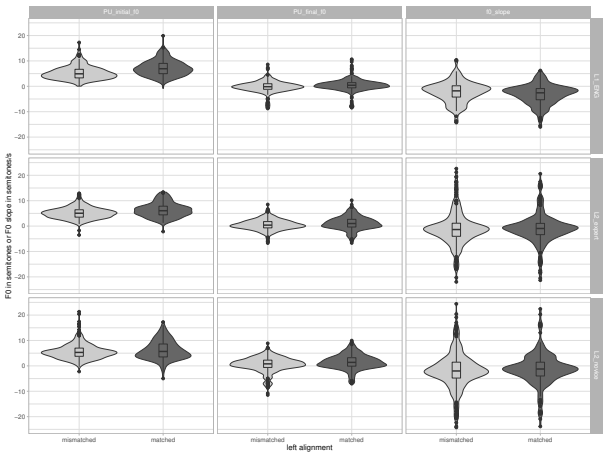

Figure 6: *Mean values of $PUI_{nitial}F0$, $PUF_{inal}F0$ and $PUS_{lope}F0$ in relation to L_{eft}.*

3.3. DU-based pitch-related preplanning

Consistent with prior research on prosodic encoding in native Mandarin [8], the results give supporting evidence for DU-based pitch-related preplanning within L1 speech. However, there is little positive evidence for DU-based prosodic preplanning in L2 spontaneous speech. There is a significant initial f0 down-shifting in right-aligned PUs (Figure 7), compared to non-right-aligned ones in L1 speech ($\beta = -0.66, t = -2.18, p < 0.05$ in L1_ENG), but no significant relationship has been found between the initial f0 height and PU-DU right alignment in L2 speech produced by learners of different proficiency levels. Speakers show a significant final f0 up-shifting in left-aligned PUs (Figure 6), in comparison with non-left-aligned PUs in L1 speech ($\beta = 0.57, t = 2.33, p < 0.05$ in L1_ENG). However, no significant relationship has been found between the final f0 height and PU-DU left alignment in L2 speech produced by learners of high proficiency. Notably, $PUF_{inal}F0$ is significantly related to the L_{eft} alignment condition in beginning learners' speech. The reason why this is the case is not explainable based on our prediction. The speech planning pattern of novice learners seems to deviate a lot from native speech and even from speech produced by advanced learners, in that f0 slope varies in the opposite direction with respect to different alignment conditions compared with the other two groups. Overall, it can be concluded that compared with native speakers, learners showed less sign of preplanning a semantic units in spontaneous narratives.

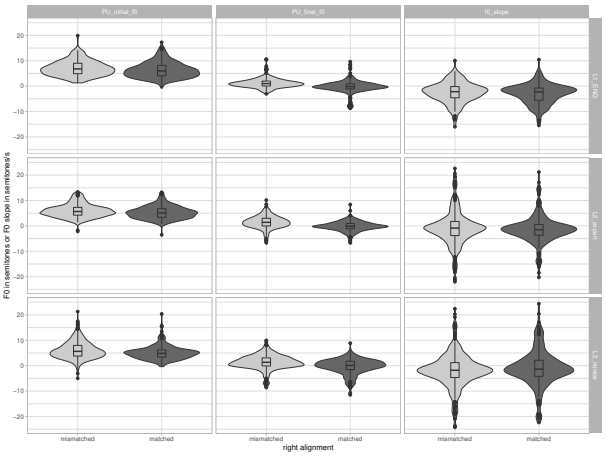

Figure 7: *Mean values of $PUI_{nitial}F0$, $PUF_{inal}F0$ and $PUS_{lope}F0$ in relation to R_{ight}.*

4. Conclusions

The present study conducted a preliminary analysis of discourse-level prosodic encoding and speech planning pattern in L1 and L2 speech. Analyses of the general characteristics of PUs indicated that L1 and L2 spontaneous narratives differed in speaking rate, planning unit size and pitch range. Furthermore, native speakers and L2 learners of different proficiency levels showed varying sensitivity to DUs in their prosodic encoding and differing capability of DU-based pitch-related preplanning.

The results presented in this paper are subject to several important qualifications, since we only used a relatively small dataset to draw inferences about speakers' prosodic encoding patterns. Future direction includes performing more thorough analysis to address these issues.

5. Acknowledgements

This study was supported by Advanced Innovation Center for Language Resource and Intelligence (KYR17005), National Social Science Foundation of China (18BYY124), Wutong Innovation Platform of Beijing Language and Culture University (19PT04), the Science Foundation and Special Program for Key Basic Research fund of Beijing Language and Culture University (the Fundamental Research Funds for the Central Universities) (21YJ040004, 21YCX180). Jinsong Zhang is the corresponding author.

6. References

[1] F. Zimmerer, J. Jügler, B. Andreeva, B. Möbius, and J. Trouvain, "Too cautious to vary more? A comparison of pitch variation in native and non-native productions of French and German speakers," 2014.

[2] I. Mennen, F. Schaeffler, and C. Dickie, "Second language acquisition of pitch range in German learners of English," *Studies in Second Language Acquisition*, vol. 36, no. 2, pp. 303–329, 2014.

[3] M. Ordin and L. Polyanskaya, "Acquisition of speech rhythm in a second language by learners with rhythmically different native languages," *The Journal of the Acoustical Society of America*, vol. 138, no. 2, pp. 533–544, 2015.

[4] P. Warren, I. Elgort, and D. Crabbe, "Comprehensibility and prosody ratings for pronunciation software development," *Language Learning & Technology*, vol. 13, no. 3, pp. 87–102, 2009.

[5] A. R. Bradlow, "Speaking rate, information density, and information rate in first-language and second-language speech." in *INTERSPEECH*, 2019, pp. 3559–3563.

[6] C.-Y. Tseng, Z.-Y. Su, C.-F. Huang, and T. Visceglia, "An initial investigation of L1 and L2 discourse speech planning in English," in *2010 7th International Symposium on Chinese Spoken Language Processing*. IEEE, 2010, pp. 55–59.

[7] A. Wennerstrom, "Intonation as cohesion in academic discourse: A study of Chinese speakers of English," *Studies in Second Language Acquisition*, pp. 1–25, 1998.

[8] A. C.-H. Chen and S.-C. Tseng, "Prosodic encoding in Mandarin spontaneous speech: Evidence for clause-based advanced planning in language production," *Journal of Phonetics*, vol. 76, p. 100912, 2019.

[9] Y. Zhang, Z. Li, and J. Zhang, "A comparison study on the alignment of prosodic and semantic units and its effects on f0 shifting in L1 and L2 English spontaneous speech," in *2021 12th International Symposium on Chinese Spoken Language Processing (ISCSLP)*. IEEE, 2021, pp. 1–5.

[10] C. Shih, "A declination model of Mandarin Chinese," in *Intonation*. Springer, 2000, pp. 243–268.

[11] F. Scholz and Y. Chen, "Sentence planning and f0 scaling in Wenzhou Chinese," *Journal of Phonetics*, vol. 47, pp. 81–91, 2014.

[12] J. Yuan and M. Liberman, "F0 declination in English and Mandarin broadcast news speech," *Speech Communication*, vol. 65, pp. 67–74, 2014.

[13] E. L. Asu, P. Lippus, N. Salveste, and H. Sahkai, "F0 declination in spontaneous Estonian: implications for pitch-related preplanning in speech production," *Proceedings of Speech Prosody 2016*, pp. 1139–1142, 2016.

[14] A. R. Bradlow. (n.d.) ALLSSTAR: Archive of L1 and L2 scripted and spontaneous transcripts and recordings. [Online]. Available: https://oscaar3.ling.northwestern.edu/ALLSSTARcentral

[15] M. Mayer, "Two Moral Tales: Bird's New Hat and Bear's New Clothes," *Four Winds Press, New York*, pp. 1–48, 1974a.

[16] ——, "Bubble Bubble," *Parents' Magazine Press, New York*, pp. 1–20, 1973.

[17] A. R. Bradlow. (n.d.) Speechbox. [Online]. Available: https://speechbox.linguistics.northwestern.edu

[18] J. S.-Y. Park, "Cognitive and interactional motivations for the intonation unit," *Studies in Language. International Journal sponsored by the Foundation "Foundations of Language"*, vol. 26, no. 3, pp. 637–680, 2002.

[19] P. Boersma. (2020) Praat: doing phonetics by computer. [Online]. Available: http://www.praat.org/

[20] Y. Jadoul, B. Thompson, and B. De Boer, "Introducing parselmouth: A python interface to praat," *Journal of Phonetics*, vol. 71, pp. 1–15, 2018.

[21] U. D. Reichel and K. Mády, "Comparing parameterizations of pitch register and its discontinuities at prosodic boundaries for Hungarian," 2014.

[22] D. Bates, M. Mächler, B. Bolker, and S. Walker, "Fitting linear mixed-effects models using lme4," *arXiv preprint arXiv:1406.5823*, 2014.

[23] B. T. West, K. B. Welch, and A. T. Galecki, *Linear mixed models: a practical guide using statistical software*. CRC Press, 2014.

[24] J. Yuan, Q. Dong, F. Wu, H. Luan, X. Yang, H. Lin, and Y. Liu, "Pitch characteristics of L2 English speech by Chinese speakers: A large-scale study." in *INTERSPEECH*, 2018, pp. 2593–2597.

[25] T. Visceglia, C.-y. Tseng, Z.-y. Su, C.-F. Huang *et al.*, "Discourse prosody planning in L1 and L2 English," in *Proceedings of Oriental COCOSDA International Conference on Speech Database and Assessments*, 2010, pp. 24–25.