



iCALL Corpus: Mandarin Chinese Spoken by Non-Native Speakers of European Descent

Nancy F. Chen, Rong Tong, Darren Wee, Peixuan Lee, Bin Ma, Haizhou Li

Institute for Infocomm Research, Singapore

{nfychen, tongrong, mabin, hli}@i2r.a-star.edu.sg

Abstract

We present *iCALL*, a speech corpus designed to evaluate Mandarin Chinese pronunciation patterns of non-native speakers of European descent, developed at the Institute for Infocomm Research (I²R) in Singapore. To the best of our knowledge, *iCALL* is larger than any reported non-native corpora to date in terms of utterance number, duration, and number of speakers: *iCALL* consists of 90,841 utterances from 305 speakers with a total duration of 142 hours. The speakers are gender-balanced, from a diverse native language background, and represent a realistic sampling of the adult age of Mandarin learners. The read utterances are phonetically balanced and are of varying lengths (words, phrases, and sentences). The spoken utterances are phonetically transcribed and perceptually rated with fluency scores by trained native speakers of Mandarin. In this work, we share our experience in corpus design, data collection, and human annotation and analyze phonetic and tonal error patterns, in particular their relationship with speaker demographics and utterance length. Potential applications of the *iCALL* corpus include computer-assisted pronunciation training (CAPT), lexical tone recognition, automatic fluency assessment, accent recognition, and accented Mandarin speech recognition.

Index Terms: computer-assisted language learning (CALL), computer-assisted pronunciation training (CAPT), first language (L1), second language (L2), database, language resource, lexical tone

1. Introduction

An increasing number of people are learning Mandarin Chinese as a second language (L2), yet there is few speech corpora reported to date targeting non-native Mandarin (see Table 1 and Table 2). Over two-thirds of the non-native speech corpora focus on English as the target language, while the target language for the remaining corpora include Arabic, French, German, Japanese, Russian, and Spanish [1, 2]. In addition, many of these speech corpora are small in size [3], making them challenging to be used directly in speech technology applications.

Compared to the acquisition of English as a second language, the peculiarities of Mandarin Chinese present some interesting challenges for second language learners. In fact, Mandarin Chinese is listed as one of the most difficult languages for native English speakers [4]. The most well-known challenge is the perception and production of lexical tones in Mandarin Chinese [5, 6]. In Mandarin Chinese, the pitch contour of a syllable defines the lexical tone and thus the meaning of a Chinese character is changed by merely changing the pitch contour without altering the phonetic characteristics. For example, the sound "ma" changes meaning from *mom* to *hemp* simply by inflecting the tone upwards (like an interrogative in English) instead of a

high, level pitch. In addition to pitch contours, it has also been reported that timing is an important perceptual cue in determining how native Mandarin Chinese productions sound [7].

Mandarin is also well-known for its variety of affricate and fricative consonants with subtle phonemic differences in place of articulation or aspiration, which have been reported to be challenging for non-native speakers [5, 6, 8].

Motivated by the reasons mentioned above, we designed and collected a large-scale speech corpus of non-native Mandarin production with detailed annotations. We targeted speakers of European descent since virtually no European language is characterized with lexical tones, in contrast to languages in Africa, Asia, and in the Americas. Table 2 lists non-native speech corpora that are comparable in size with the *iCALL* corpus presented in this work in at least one of the following three attributes: speaker number, utterance number, and duration. We see that *iCALL* is largest in size in all three attributes.

In particular, *iCALL* possess the following qualities: (1) large in speaker size (305), utterance number (90,841), duration (142 hr, of which 59 hours are estimated to be speech from force alignments), and per-speaker data (~300 utterances/speaker, including words, phrases, and sentences); (2) complete and balanced in phonetic coverage; (3) diverse in speaker demographic background: non-tonal first languages of European origin, gender-balanced, and realistic distribution of the adult ages of Mandarin learners; (4) detailed in human annotations, which consists of phonetic and tonal transcriptions along with fluency ratings.

To the best of our knowledge, few, *if any*, non-native speech corpora of *any* target language reported to date satisfies all the aforementioned desirable qualities.

In the following sections, in addition to delineating the corpus design of *iCALL*, we also present tonal and phonetic error analysis in relation to speaker background and utterance length, and discuss potential applications of the *iCALL* corpus¹.

2. iCALL Corpus Design

2.1. Background: Mandarin Chinese Phonology

Each Chinese character corresponds to a syllable, which takes on the structure of [C]V[N]T, where C is a consonant, V is a vowel, N is a nasal consonant, T is a lexical tone, and [] denotes the sub-syllable enclosed to be optional. Note that for each syllable, it is mandatory to have a vowel and a tone.

In this paper, we provide both Pinyin (with numerical tone markers) and their IPA representations, which are in italics and enclosed in square brackets.

¹This work would not be complete without the help of M. Dong and X. Wang's data collection and coordination efforts and Y. Li, X. Hu, C. Zhang and E. Hsieh's detailed annotations and analysis.

Table 1: Abbreviations used in Table 2.

Cantonese	C	Danish	D	English	E	German	G	Italian	I	Mandarin	M	Russian	R
Czech	Cz	Dutch	Du	French	F	Indonesian	In	Japanese	J	Portugese	P	Spanish	S

Table 2: Comparison of large-scale non-native speech corpora. Blanks indicate unavailable information. Phon. Trans.: Phonetic Transcription; Prof. Rating: Proficiency Rating; P: partially transcribed.

Corpus	Author	Source	Target Language	Native Language	#Spkrs	#Utt	Dur	Phon. Trans.	Prof. Rating
AMI [9]			E	G et al.			100h	N	N
ATR-Gruhn [10]	Gruhn	ATR	E	C G F J In	96	15,000		N	Y
C-AuDIT [11]	Honig		E	F G I S	56	18,424			
CU-CHLOE [12]	Meng	CUHK	E	C M	211			P	N
Cross Towns [13]	Schaden	U. Bochum	E F G I Cz Du	E F G I S	161	72,000	133h	Y	N
ERJ [14]	Minematsu	U. Tokyo	E	J	200	68,000		N	Y
ISLE [15]	Menzel	U. Hamburg	E	G I	46	11,484	18h	Y	Y
LeaP [16]	Gut	U. Bielefeld	G E	32 in total	131	359	12h	Y	N
Sunstar [17]		EU	E	G S I P D	100	40,000		N	N
Tokyo-Kikuko [18]	Kikuko	U. Tokyo	J	10 countries	140	35,000		N	Y
NTU [19]	Wang	Nat. Taiwan U.	M	36 countries	278	8340		Y	N
iCALL	Chen	I2R	M	E F G I P R S et al. (24 in total)	305	90,841	142h	Y	Y

Tone is the use of pitch in speech. Tones primarily express paralinguistic information such as emotions in languages like English. In contrast, Mandarin Chinese uses *lexical tones* to encode semantics; i.e., a change in tone changes the meaning of a word. For example, *ma1* [ma¹] and *ma2* [ma⁴] are phonetically the same but differ in tone, resulting in different meanings: mom vs. hemp.

In Table 3, we introduce the lexical tones in Mandarin. Note that there is no equivalent in English for Tone 3, and Tone 5 is a neutral tone or *lack of tone*, similar to an unstressed syllable in English.

Table 3: Lexical Tones in Mandarin.

Tone	Pitch Contour	English Equivalent
1	High-level	Singing
2	High-rising	Question-final intonation; e.g., What?!
3	Dipping	N/A
4	Falling	Curt commands; e.g., Stop!
5	Undefined	Unstressed syllable

2.2. Speakers: Diverse Native Language Background

A total of 305² speakers of European descent whose first language is non-tonal were recruited. The speakers' native languages were of European origin: 1/2 are Germanic (e.g. English, German); 1/3 are Romance (e.g. French, Spanish, Italian); and 1/6 Slavic (e.g. Russian)³. The gender ratio is balanced. The speaker age group was sampled such that it represents the distribution of adult second language learners of Mandarin in Beijing, ranging from 18 to 52.

²The corpus size (e.g. number of speakers, number of utterances) and speaker demographic statistics are slightly different from that reported in our previous work [5, 20, 21] due to data cleanup (e.g. noisy audio) and additional speaker recruitment.

³These numbers exclude speakers whom chose not to reveal more information beyond their L1 being non-tonal and of European origin.

All speakers are beginner learners of Mandarin and rely heavily on the Pinyin phonetic representations (instead of Chinese characters) to read the prompts. The non-native speech recordings were recorded in quiet office rooms, sampled at 16 kHz, and encoded in 16 bit pulse-code modulation (PCM).

2.3. Lexical Content: Reading Prompts

Each speaker was given a distinct set of Pinyin prompts to read, where some of the prompts are overlapping.

2.3.1. Complete Phonetic Coverage

The iCALL corpus is phonetically balanced such that its phonetic frequency matches that of the natural phonetic distribution in Mandarin [22]. This complete and balanced phonetic coverage is also achieved on a per-speaker basis.

2.3.2. Diverse Utterance Length

Since there is a trade-off tendency between pronunciation accuracy and syntax complexity, the corpus consists of utterances of different lengths to provide a richer context for modeling pronunciation errors. Each speaker reads 300 utterances, where the first 100 consists of 2-syllable words (the most common length of words in Mandarin Chinese), the second 100 consists of phrases that are 3 or 4 syllables long (idioms usually are 4 syllables long in Chinese), and the last 100 are sentences of at least 5 syllables, with an average of 10.8 syllables per sentence. (See Table 5.)

2.4. Human Annotations

2.4.1. Phonetic and Tonal Transcription

Phonetic transcriptions were done in Pinyin, a phonetic system used to transcribe Chinese characters into Latin script, while tones were transcribed in numerical form.

Two rounds of phonetic and tonal transcription were carried out. In the first round, 64 native speakers of Mandarin Chinese shared the load of transcribing the non-native utterances phonetically using Pinyin. Results reported in [5] were done using transcriptions from this round.

During fluency scoring (Section 2.4.2), the phonetic and tonal transcriptions of 77,895 utterances were checked (and corrected if necessary) and further refined. For tones and phones that were ambiguous, the 2nd-round human annotators were instructed to make a forced decision by appending a * sign next to the tone/phone to indicate the decision was perceptually ambiguous. For example, the pitch contour of a syllable might go up and then go down, which is undefined among lexical tones in Mandarin. The annotators would be asked to make a forced decision among the 5 lexical tones, but mark it as ambiguous at the same time.

2.4.2. Fluency Scoring Protocol

The fluency scoring protocol was developed by two native Mandarin speakers using several runs of pilot data to iteratively establish guidelines (including mispronunciation rate) for fluency score ratings into four levels: fluent (4), good (3), average (2), poor (1). Zero mispronunciations at the phonetic and tonal level are necessary but insufficient for obtaining a score of 4. In other words, to be scored as 4, in addition to no mispronunciations, the utterance also had to be read in a fluent manner. On the other hand, if every syllable was mispronounced, the score is 1.

The two raters practiced scoring with the established guidelines to ensure inter-rater and intra-rater consistency. Before the two raters could officially start scoring the fluency levels, their scores needed to reach at least 0.8 for Pearson correlation coefficients and at least 0.6 for Cohens Kappa coefficients on a subset of utterances designated for training purposes. After the training phase, inter-rater consistency tests were conducted regularly to ensure the consistent scoring quality.

3. Pronunciation Error Pattern Analysis

3.1. Lexical Tone Errors

Overall, there are 32.0% tonal errors across the five tones. Table 4 shows the confusion matrix breakdown of the tonal errors. Tone 3 is the most challenging, resulting in the lowest production accuracy (58.8%), while Tone 1 and Tone 5 are the easiest for the non-native speakers.

This pattern is expected as Tone 3 is the most exotic to non-tonal languages, requiring the pitch contour to fall and rise within a syllable, whereas Tone 1 is the high-level tone with a flat pitch contour and Tone 5 is the equivalent to unstressed syllables, both of which are common in non-tonal languages. Tone 3 is the most commonly mispronounced as Tone 2, which corresponds with prior work on perceptual and production studies of American learners of Mandarin [23].

3.1.1. Tonal Error Rate vs. Utterance Length

Table 5 shows that lexical tone error rate correlates with utterance length, whether it is computed at the utterance level or syllable level. For two-syllable words, half of them have at least one tonal mistake. For sentences of at least 5 syllables, only 13% are completely correct in terms of tone production.

3.1.2. Mispronounced Tone Pairs

Table 6 lists the tone pairs that are most mispronounced. Tone pair (3,2) is only pronounced correctly 39.9% of the time, and

Table 4: Confusion matrix of lexical tones (%).

		Non-native Production				
		Tone 1	Tone 2	Tone 3	Tone 4	Tone 5
Canonical	Tone 1	72.6	8.3	7.1	11.7	0.2
	Tone 2	13.3	67.7	8.0	10.7	0.3
	Tone 3	12.0	19.3	58.8	9.7	0.2
	Tone 4	13.5	9.9	7.9	68.4	0.3
	Tone 5	6.4	6.3	3.6	8.1	75.5

Table 5: Error rate correlates with utterance length (measured in the number of Chinese characters) of the subset with refined transcriptions in Section 2.4.1.

	Word	Phrase	Sentence	Mean
Utterance length	2.0	3.5	10.8	5.4
Number of utterances	25,340	26,790	25,765	N/A
At least 1 tonal error	49 %	63 %	87 %	66 %
Tonal errors/syllable	0.34	0.39	0.45	0.39
At least 1 phonetic error	18 %	23 %	49 %	30 %
Phonetic errors/syllable	0.13	0.11	0.13	0.12

Table 6: Top mispronounced tonal patterns. **Due to tone sandhi, the accurate implementation of (3,3) should be (2,3).

Tone Pattern	Percent Correct (%)	Top Mispronunciation(s)
(4,4)	55.2	(1,4): 5.3%
(2,2)	53.2	(1,2): 8.6%
(4,3)	46.4	(4,2): 11.9%
(3,2)	39.9	(2,2): 13.1%
(3,3)**	9.6	(3,3): 37.5%; (3,1): 7.4%

most often mispronounced as (2,2). Tone pair (3,2) is often reported as challenging on blogs of Mandarin learners. For example, America, *mei3 guo2* [mei³ guo²], is often mispronounced as *mei2 guo2* [mei² guo²]. Other tone pairs that are challenging to L2 learners include (4,3), (4,4), and (2,2). We suspect these aforementioned tone pairs to be challenging because the difference between the pitch level of the latter portion of the former tone to the initial portion of the latter tone is larger, making the pitch transition more challenging.

*Tone sandhi*⁴ is a phonological change where the tones assigned to individual words or morphemes change based on the pronunciation of adjacent words or morphemes [24]. The most well-known example of tone sandhi in Mandarin is when two consecutive Tone 3's occur, native speakers will produce a Tone 2 followed by a Tone 3 instead: (3,3) → (2,3). This tone sandhi pattern only achieves 9.6% accuracy in non-native speech.

3.1.3. Tonal Preferences vs. Speaker Demographics

In terms of tone production accuracy, we find that male speakers are better than female speakers, younger speakers are better than older speakers, and Germanic speakers are better than both Romance and Slavic speakers ($p < 0.05$ for all). Further analysis reveals that female speakers, Romance speakers, and Slavic speakers have a bias to produce Tone 1's, whereas older speak-

⁴Sandhi means *joining* in Sanskrit

ers have a bias to produce Tone 4's.

French speakers are more likely than other Romance speakers to produce Tone 1's when the reference is not Tone 1. This preference for producing Tone 1's might be partially because French lacks lexical stress, unlike English and other Romance languages such as Italian [25].

3.2. Phonetic Errors

Overall, there are 5.0% phonetic errors. Table 7 shows the top phonetic errors including affricates, aspirated stops, the alveolar fricative *s* [s], and the vowel *iu* [iəu]. These phonetic error patterns mirror those mentioned in [6] about difficult consonants and vowels German speakers encounter when learning Mandarin Chinese.

For *t* /t^h/ and *p* /p^h/, they are predominantly substituted by their unaspirated counterparts (*d* [t] and *b* [p]) 73.4% and 95.0% of the time, respectively. For the other phones, the distribution of the substituted phones are more spread out. Note that 6 out of the top 10 phonetic errors are related to aspiration.

3.2.1. Phonetic Error Rate vs. Utterance Length

Table 5 shows that similar to tonal errors, phonetic error rate (at the utterance level) gets higher as the utterance length increases. However, phonetic error rate at the syllable level is insensitive to utterance length.

3.2.2. High Deaspiration Rate for Romance Speakers

There is no aspiration in Romance languages [26] such as French [27], Italian [28] and Spanish [29], which might explain their de-aspiration error patterns. The patterns for each subgroup are different, which we elaborate below.

Spanish speakers consistently de-aspirate; for de-aspiration error patterns *p* /p^h/ → *b* [p], *c* /t^h/ → *z* [ts], *ch* /t^h/ → *zh* [tʂ], *q* /t^h/ → *j* [tɕ], and *k* /k^h/ → *g* [k], Spanish speakers are at least 1.4 more times likely to de-aspirate than French speakers.

French speakers de-aspirate all aspirated phonemes; the only exception is the aspiration substitution *j* /t^h/ → *q* [tɕ^h] occurring twice as likely as the opposite *q* /t^h/ → *j* [tɕ].

Italian speakers' exception is the aspiration error *ch* /t^h/ → *zh* [tʂ], twice as likely as the opposite *zh* /t^h/ → *ch* [tʂ^h]. Whether these differences are due to L1 or other reasons like individual speaking style remains a topic for future research.

3.3. Fluency Score Analysis

The per speaker mean and standard deviation of the fluency scores are 2.60 and 0.88. We categorize the fluency scored utterances based on their first language family and find that the means of the three language family groups are statistically significant: Germanic > Slavic > Romance (*p* < 0.05).

We also conducted an oracle experiment using decision tree clustering to evaluate how well we can predict fluency levels if we assume we know both the canonical reference pronunciation and the non-native pronunciation (manual transcription). The data partition is the same as [20]. Other setup configurations are similar to [5] but with features like mispronunciation rate, whether there were consecutive mistakes, if the spoken utterance is of a different length from the reference utterance (implying deletions or insertions), utterance length, the syllable position of the mispronunciations, speaking rate. The classifier output fluency scores achieves a high correlation with human ground-truth scores: 0.87.

Table 7: Top mispronounced phones (in IPA and Pinyin), their corresponding most substituted phone (in IPA and Pinyin), and the error type.

Phoneme IPA	Pinyin	Error Rate	Top Substitution Pattern	Rate	Error Type
/t ^h /	c	16.4%	/t ^h / → [ts] c → z	6.5%	Deaspiration
/t ^h /	q	10.5%	/t ^h / → [tʂ ^h] q → ch	3.6%	Fronting
/t ^h /	ch	10.2%	/t ^h / → [tʂ] ch → zh	3.0%	Deaspiration
/ts/	z	9.8%	/ts/ → [tʂ] z → zh	2.8%	Backing
/tʂ/	zh	8.2%	/tʂ/ → [tʂ ^h] zh → ch	2.4%	Aspiration
/s/	s	7.7%	/s/ → [ʂ] s → sh	2.2%	Backing
/t ^h /	t	7.4%	/t ^h / → [t] t → d	5.5%	Deaspiration
/p ^h /	p	6.6%	/p ^h / → [p] p → b	6.3%	Deaspiration
/iəu/	-iu you	6.1%	/iəu/ → [u] iu → u	2.2%	Monophthongization
/t ^h /	j	6.0%	/t ^h / → [tɕ ^h] j → q	2.2%	Aspiration

4. Discussion

We presented a non-native Mandarin corpus for developing computer-assisted language learning applications. Our analysis reveals that lexical tones are difficult for L2 learners whose L1 is of European origin, utterance length inversely correlates with phonetic and tonal error rate, and the speaker's L1 affects the error patterns. We also showed that if the acoustic implementation of the non-native production is known, automatic fluency predictions correlate highly with human fluency labels.

We have used the iCALL corpus for lexical tone error detection [21], lexical tone recognition [30], automatic fluency assessment [20]. We plan to use iCALL for computer-assisted pronunciation training, accent recognition and analysis (similar to [31]), accented Mandarin speech recognition, and speech prosody research.

While it is widely acknowledged that prosody is essential to L2 fluency, few studies using large speech corpora have taken prosody into account. An example of a prosody annotated corpus for language learning is LeaP [16], which examines the acquisition of prosody by non-native speakers of German and English using variants of ToBI [32] to label intonations. We believe the development of iCALL and further investigations in non-native Mandarin productions could potentially shed light in how prosody carries over from L1 in L2 learning. In addition, a known challenge in prosody research is that ground-truth is relatively more subjective compared to phonetic annotation. Since the tonal structure is regulated and lexical in Mandarin, it is easier to establish objective ground truth using lexical tones from the canonical pronunciation.

For readers interested in the iCALL corpus, please email nfychen@i2r.a-star.edu.sg.

5. References

- [1] Martin Raab, Rainer Gruhn, and Elmar Noeth, “Non-native speech databases,” in *IEEE Automatic Speech Recognition and Understanding (ASRU) Workshop*, 2007, pp. 413–418.
- [2] “Non-Native Speech Databases Wikipedia Page: http://en.wikipedia.org/wiki/Non-native_speech_database#cite_note-37,” last accessed, March 17, 2015.
- [3] Center for English Corpus Linguistics, Universite Catholique de Louvain, “Learner corpora around the world: <http://www.uclouvain.be/en-cecl-lcworld.html>,” last accessed, July 1, 2015.
- [4] Effective Language Learning, “Language Difficulty Ranking: <http://www.effectivelanguagelearning.com/language-guide/language-diLanguage>,” last accessed, July 1, 2015.
- [5] Nancy F. Chen, Vivaek Shivakumar, Mahesh Harikumar, Bin Ma, and Haizhou Li, “Large-scale characterization of Mandarin pronunciation errors made by native speakers of European languages,” in *INTERSPEECH*, 2013, pp. 803–806.
- [6] Chia-Yu Chiu, Yuan-Fu Lia, Daniel Külls, Hansjorg Mixdorff, and Shing-Lung Chen, “A preliminary study on corpus design for computer-assisted German and Mandarin language learning,” in *Speech Database and Assessments, Oriental COCODA International Conference*, 2009, pp. 154–159.
- [7] Chiharu Tsurutani and Dean Luo, “Naturalness Judgement of L2 Mandarin Chinese-Does timing matter?,” in *INTERSPEECH*, 2013.
- [8] Yi-Hsiu Lai, “Asymmetry in Mandarin affricate perception by learners of Mandarin Chinese,” *Language and Cognitive Processes*, vol. 24, no. 7-8, pp. 1265–1285, sep 2009.
- [9] Jean Carletta, Simone Ashby, Sebastien Bourban, Mike Flynn, Mael Guillemot, Thomas Hain, Jaroslav Kadlec, Vasilis Karaiskos, Wessel Kraaij, Melissa Kronenthal, et al., “The AMI meeting corpus: A pre-announcement,” in *Machine Learning for Multimodal Interaction*, pp. 28–39. Springer, 2006.
- [10] R. Gruhn, T. Cincarek, and S. Nakamura, “A multi-accent non-native English database,” *Proceedings of Acoustical Society of Japan*, pp. 195–196, 2004.
- [11] Florian Honig, Anton Batliner, Karl Weilhammer, and Elmar Noth, “Islands of Failure: Employing word accent information for pronunciation quality assessment of English L2 learners,” in *ISCA Workshop on Speech and Language Technology for Education (SLaTE)*, 2009.
- [12] Helen Meng, Yuen Yee Lo, Lan Wang, and Wing Yiu Lau, “Deriving salient learners’ mispronunciations from cross-language phonological comparisons,” in *IEEE Automatic Speech Recognition and Understanding (ASRU) Workshop*, 2007, pp. 437–442.
- [13] S. Schaden, *Regelbasierte Modellierung fremdsprachlich akzent-behafteter Aussprachevarianten*, Ph.D. thesis, University Duisburg-Essen, 2006.
- [14] Nobuaki Minematsu, Yoshihiro Tomiyama, Kei Yoshimoto, Katsumasa Shimizu, Seiichi Nakagawa, Masatake Dantsuji, and Shozo Makino, “Development of English speech database read by Japanese to support CALL research,” in *International Congress on Acoustics (ICA)*, 2004, pp. 577–560.
- [15] Wolfgang Menzel, Eric Atwell, Patrizia Bonaventura, Daniel Heron, Peter Howarth, Rachel Morton, and Clive Souter, “The ISLE corpus of non-native spoken English,” in *International Conference on Language Resources & Evaluation (LREC)*, 2000.
- [16] Ulrike Gut and Englisches Seminar, “The Leap Corpus,” *online documentation: <http://www.phonetik.unifrieburg.de/leap/LeapCorpus.pdf>*, 2004.
- [17] Carlos Teixeira, Isabel Trancoso, and António Joaquim Serralheiro, “Recognition of non-native accents,” in *Eurospeech*, 1997.
- [18] Kikuko Nishina, Yumiko Yoshimura, Izumi Saita, Yoko Takai, Kikuo Maekawa, Nobuaki Minematsu, Seiichi Nakagawa, Shozo Makino, and Masatake Dantsuji, “Development of Japanese speech database read by non-native speakers for constructing CALL system,” in *International Congress on Acoustics (ICA)*, 2004, pp. 561–564.
- [19] Yow-Bang Wang and Lin-Shan Lee, “Improved approaches of modeling and detecting error patterns with empirical analysis for computer-aided pronunciation training,” in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2012, pp. 5049–5052.
- [20] Rong Tong, Boon Pang Lim, Nancy F Chen, Bin Ma, and Haizhou Li, “Subspace Gaussian mixture model for computer-assisted language learning,” in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2014, pp. 5347–5351.
- [21] Rong Tong, Nancy F. Chen, Boon Pang Lim, Bin Ma, and Haizhou Li, “Tokenizing Fundamental Frequency Variation for Mandarin Tone Error Detection,” in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2014.
- [22] “Chinese Text Computing: <http://lingua.mtsu.edu/chinese-computing/>,” last accessed, March 17, 2015.
- [23] Yue Wang, Allard Jongman, and Joan A Sereno, “Acoustic and perceptual evaluation of mandarin tone productions before and after perceptual training,” *The Journal of the Acoustical Society of America*, vol. 113, no. 2, pp. 1033–1043, 2003.
- [24] Moira Yip, *Tone*, Cambridge University Press, 2002.
- [25] Nawal Abboub, Ranka Bijeljac-Babic, Josette Serres, and Thierry Nazzi, “On the importance of being bilingual: Word stress processing in a context of segmental variability,” *Journal of Experimental Child Psychology*, vol. 132, pp. 111–120, 2015.
- [26] M. Celce-Murcia, D. M. Brinton, and J. M. Goodwin, *Teaching Pronunciation: A reference for teachers of English to speaker of other languages*, Cambridge University Press, 1996.
- [27] M. Campbell and M. Paquin, *French Fluency: Glossika Mass Sentences.*, Nolsen Bedon, Ltd., 2014.
- [28] Martin Kramer, *The Phonology of Italian*, Oxford University Press, 2009.
- [29] M. S. Whitley, *Spanish/English contrasts: A course in Spanish linguistics*, Georgetown University Press, 2002.
- [30] Rong Tong, Nancy F. Chen, Bin Ma, and Haizhou Li, “Goodness of Tone (GOT) for Non-native Mandarin Tone Recognition,” in *INTERSPEECH*, 2015.
- [31] Nancy F Chen, Sharon W Tam, Wade Shen, and Joseph P Campbell, “Characterizing phonetic transformations and acoustic differences across english dialects,” *IEEE Transactions on Audio, Speech, and Language Processing (TASLP)*, vol. 22, no. 1, pp. 110–124, 2014.
- [32] Kim E.A. Silverman, Mary E. Beckman, John F. Pitrelli, Mari Ostendorf, Colin W. Wightman, Patti Price, Janet B. Pierrehumbert, and Julia Hirschberg, “TOBI: a standard for labeling English prosody,” in *International Conference on Spoken Language Processing (ICSLP)*, 1992.