Computational perception of information foci produced by Chinese English learners and American English speakers

Juqiang Chen* and Xuliang He†
*MARCS Institute, Western Sydney University, Sydney, Australia
E-mail: j.chen2@westernsydney.edu.au Tel: +61 420286368
†Nantong University, Nantong, China
E-mail: xlnick@ntu.edu.cn Tel: +86-13515201442

***Abstract*— This study used computational perception, via SVM and Random Forest models, to examine phonetic features used by American English speakers (AE) and Chinese second language learners of English (CE1 with low proficiency and CE2 with high proficiency) in realizing different information foci. For all participant groups, the machine learning models achieved above chance level accuracy. Coda duration and the duration of the rising contour were two phonetic features that ranked top across three participant groups in terms of their importance to the models. The SVM models trained with the AE data classified different foci by CE1 and CE2 with above chance level accuracy, but English proficiency had little effect on the classification results.**

I. INTRODUCTION

A focus is the informative part of an utterance, contrast to the background information of the sentence. Foci vary in terms of how informative they are. For example, a focus can be broad as in (1a, Broad Focus, BF) where the whole sentence is the focus constituent in response to the question, or narrow as in (1b, c) [1]. A further distinction between narrow focus (NF, 1b) and corrective focus (CF, 1c) has been made in accordance with the different information status [2]. "New York" in (1b) provides the requested information "where", while "New York" in (1c) corrects the old information "Chicago" in the preceding question.

(a) A: What are your plans for tomorrow?
B: I would like to go to **New York**.
(b) A: Where would Karel like to take you? (1)
B: He would like to take me to **New York**.
(c) A: Did your mother want to send you to Chicago?
B: No, she wanted to send me to **New York**.

II. LITERATURE REVIEW

A. Phonetic correlates of different information foci

Native speakers of different languages were reported to employ a range of phonetic means, such as pitch accent distribution, phrase boundary, pitch range and duration to express the meanings in different focus conditions.

In [3], speakers of American English (AE) were asked to read sentences with target words in different focus conditions (in response to a question) and in sentence-initial, sentence-medial and sentence-final positions. The f0 peak of a word was found to be consistently higher in NF compared with the neutral-focus sentence and the general locations of the f0 peaks were largely the same with or without a narrow focus.

Similarly, in [4], Dutch speakers read short dialogues with the answer containing three different focus conditions (BF, NF, CF). The target words were trisyllabic pseudo place names with the first syllable as the stressed syllable. It was found that the onsets of the stressed syllable in both NF and CF were longer than those in BF. And the coda durations of the stressed syllable were significantly longer in BF and NF than in CF. Moreover, the f0 peak of the word was higher and earlier in BF than in CF and NF. The falls of the rising-falling accentual contour in CF and NF descended more steeply than BF.

However, British English speakers, when asked to read sentences with target words in different sentence positions and under three focus conditions in [5], did not show significant effects of focus conditions on the height of f0 nor peak alignment. Only some participants produced accented words with longer duration in NF and CF than in BF. The authors interpreted the results in support of the claim that there were no robust phonetic cues in the accented word itself for distinguishing BF, NF and CF.

In summary, the phonetic details of the target words vary in different focus conditions and sentence positions. There seems to be a lack of reliable phonetic cues that distinguish different foci across different languages. Moreover, the above studies approached the issue from a production perspective. Very few studies have examined whether phonetic differences in the target word are sufficient to distinguish different focus conditions either by human listeners or machine learning models.

B. Phonetic realization of different information foci by non-native English speakers

A group of studies have found L1 prosodic structures influence how participants realize focus in their second

language (L2). In [6], Japanese and Korean English speakers transferred the intonation patterns in their mother tongues when realizing focus in English. However, the authors did not test the effect of different types of focus. Similarly, native Zulu English second language (L2) speakers in South Africa marked focus in different ways from the native speakers [7]. Proficient speakers of English realized the focus more like native English speakers, whereas less proficient ones transferred features from L1 Zulu when signaling the focus. When native English speakers perceived the production data of Zulu English speakers, they judged that Zulu speakers' production of focus had a different perceived prominence of the test words. However, this study did not provide detailed phonetic analysis of the target words, making the comparison of the results with previous native production studies difficult.

In [8], phonetic details of the target words were analyzed. Mandarin L1 Dutch L2 speakers were asked to read short dialogues with three different information foci (BF, NF, CF). The target words were tri-syllabic words placed at the end of sentences analogous to [4]. Mandarin L1 Dutch L2 speakers differed from Dutch native speakers in onset duration, coda duration, rime duration and the duration of the falling contour, as well as the relative peak delay in the rime and the f0 excursions of the nuclear syllable.

However, we did not know whether the above-mentioned phonetic differences in focus realization by L2 speakers would affect perception/classification of focus by native speakers, and whether their realization is phonetically informative to disambiguate different focus types.

C. *Computational perception*

Traditional statistical modeling (e.g. the General Linear Model), as used in most previous studies that investigated the phonetic realization of information foci, examines the whole dataset and detects the effect of focus types in terms of discrete measures. Usually several models are built based on different acoustic measures to test the effect of different foci and language groups, but a comprehensive picture is hard to get. Human listeners, on the other hand, hear the utterance one by one and make their judgement accordingly. This process can be simulated by machine learning algorithms that take in a set of features and output classification results based on these features.

In [9], Support Vector Machine (SVM) and Random Forest models have been used to classify nasal versus oral vowels based on different sets of acoustic measures. We understand that computational perception is not identical to human perception but the advantage of computational perception in phonetic research is that it is based purely on acoustic features and thus tears apart syntactical or lexical effects. In addition, unlike identifying vowel or consonants, it is difficult to only give listeners some target words and test how listeners identify different foci. Machine learning algorithms can identify different foci without context and thus approach this issue from a perception point of view.

D. *The present study*

The present study employed two types of machine learning models, SVM and Random Forest, to investigate to what extent different types of information focus can be perceived/classified based on phonetic features. We also compared among native American English (AE) and Chinese English learner of low proficiency (CE1) and Chinese English learner of high proficiency (CE2) groups in terms of the accuracy of classification and the ranking of different phonetic features. In addition, we trained two models with AE production data and tested them with CE1 and CE2 to simulate how AE listeners perceive non-native production of different information foci.

III. METHOD

A. *Data collection*

16 native AE speakers (female: 8) from New York University participated in the study. 20 first-year Chinese English major student (all female, CE1) and 20 (all female) third-year Chinese English major students (CE2) students from Nantong University were chosen randomly based upon their registered student numbers. No participants reported they were dyslexic or had hearing problems.

All the participated were asked to read twelve dialogues (in total) blocked into three information foci condition (four for each BF, NF, CF) in English. In each dialogue, "A" was context and "B" was the test sentence, the response to "A". Each sentence B had one of four tri-syllabic words as targets (pseudo place names: Manderen, Momberen, Memberen, Munderen) at the end of the sentence. The nuclear pitch accent in all sentences was expected to occur on the target words.

Before the recording, participants were asked to read the response sentence in the dialogues with a falling tone and to accentuate the target word. During the recording, participants were asked to first read the practice dialogue and their performance was supervised and checked to ensure the correct accent placement and intonation contour. Participants were allowed to repeat sentences to the best of their abilities.

Participants were recorded with a Zoom H4 Handy Digital Recorder in quiet study rooms of Nantong University library and New York university library. A total number of 672 utterances were collected but 4 utterances were removed due to mistakes in production, resulting in 668 utterances for the analysis.

B. *Annotation*

The segmental boundaries of the target word were labeled manually based on auditory information and visual inspection of the spectrogram and the waveform. Since the boundaries between segments are rarely clear-cut, the point where phonemes least affected each other was chosen as the location of the boundary. All segmental boundaries on this tier were placed at negative-to-positive zero-crossings (Figure 1).

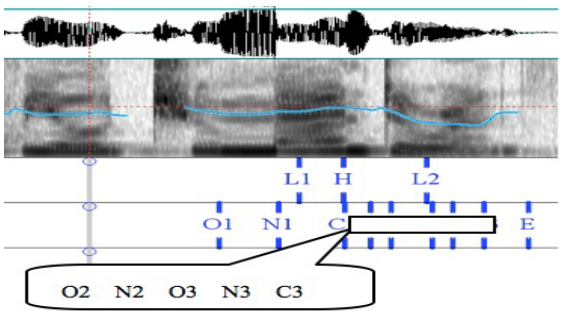

Fig. 1 Target word annotation

The intonation contours were annotated in form of ToDI (Gussenhoven, et al., 2002) to make the data comparable to the previous studies[4], [8] with similar target word structures (Table 1). The pitch accent H was placed automatically with a Praat function that gives the location of the highest pitch in a selected interval. L1 and L2 were placed at the point where a sudden change in the slope was visible.

Table 1. Labels on the target words

Name	Position
L1	Elbow before the nuclear peak
H	Maximum f0 of the pitch accent
L2	Elbow after the nuclear peak
O1	Beginning of nuclear onset
N1	Beginning of nuclear vowel
C1	Beginning of nuclear coda
O2	Beginning of the onset in the second syllable
N2	Beginning of the vowel in the second syllable
O3	Beginning of the onset in the third syllable
N3	Beginning of the vowel in the third syllable
C3	Beginning of the onset in the postnuclear stressed syllable
E	End of the postnuclear word

C. *Phonetic features*

A set of phonetic features were selected for this study (Table 2) to be comparable to previous studies [4], [8]. It should be noted that theoretically there are an infinite number of features that could be considered and the set we used here was not exhaustive. In order to reduce talker variability, all the features were z-score transformed before they were fed into the machine learning algorithms.

Table 2. Phonetic features of the target words

Variable	Variable meaning	Labels
Odur	Onset duration	O1 to N1
Vdur	Nuclear vowel duration	N1 to C1
Cdur	Coda duration in nucleus	C1 to O2
Rdur	Rime duration in nucleus	N1 to O2
Sylldur	Duration of accented syllable	O1 to O2
Wdur	Accented word duration	O1 to E
O1Hexc	Excursion from onset to peak	f0 excursion: O1 to H
Fallexc	Excursion of fall	f0 excursion: H to L2
Nucleusfall exc	Excursion of nucleus fall	f0 excursion: H to O2
H-N2 exc	Excursion from peak to the second vowel	f0 excursion: H to N2
pd	Peak relative to nuclear vowel(peak delay)	Timing from H to N1
Pd/rime	Percentage pd in the rime	Pd/duration of rime
Falldur	Duration of Fall	Duration from H to L2
Risedur	Duration of Rise	Duration from L1to H
Hf0-Ef0	Peak relative to end	f0 excursion from H to E
L1-Hexc	Excursion from the first elbow to peak	f0 excursion from L1 to H
Risesp	Speed of Rise	(H– L1)/(H(sec) – L1 (sec))
Fallsp	Speed of Fall	(H – L2)/(L2 (sec) −H(sec))
O1f0-Ef0	Onset f0 relative to the end of the post word	O1f0−Ef0
L1f0-Ef0	f0 of the first elbow relative to the end	L1f0−Ef0
L2f0-Ef0	f0 of the second elbow relative to the end	L2f0−Ef0

D. *Machine learning classifiers*

We used two types of machine learning classifiers, RandomForests and SVM in the present study. RandomForests is a type of machine learning model that consists of a group of decision trees trained and tested on randomized subsets of the data. In this study, we used 500 trees in all the RandomForest models as in [9]. The RandomForest classfier examines the outputs from each decision tree and produces an accuracy for the classification. In addition, the RandomForest classfier provides direct measures of feature importance for each model. The *randomForest* package [10] in R was used to build the classifiers in the study.

However, RandomForests sacrifice some accuracy in classification for their interpretability. To make up for this, we used SVM, which has an overall good performance in the field of machine learning.

SVM works by finding a line that best separates the different classes in the data. When the data is not linearly separable, a kernel can be used to handle the non-linear relations. In this study, kernel "radical" was used. However, we lost feature weightings when using a kernel. Thus we relied mostly on RandomForest to indicate the importance of each phonetic feature.

In addition, due to a relative small dataset (from a machine learning perspective), we used "10-fold cross-validation", in which the analysis was run 10 times on the data, and each time a different 9/10ths of the data as "training" and the remaining 1/10th as "test". The *e1701* package [11] in R was used for building the SVM classifiers in this study.

IV. COMPUTATIONAL PERCEPTION

A. *Classifying information foci produced by AE and CE speakers.*

First, we built three RandomForest classfiers with all acoustics features of the target word by AE, CE1 and CE2. AE model was specified in (2) as an example.

```
RF_AE = randomForest(focus_type ~ .,
                     ntree = 500,                    (2)
                     data = datasetAE,
                     importance = TRUE)
```

We obtained the confusion matrices for each model and accuracy for each focus type and calculated overall accuracy for each language group in Table 3. The overall classification accuracy (the mean of three focus types) for each group is round 50%, higher than the chance level 33.3% (given that there were three choices). However, this accuracy was not very high, suggesting that classifying focus types based solely on the phonetic features of the target word was difficult for the RandomForest models. Even for AE speaker data, the models did not classify with a very high accuracy, suggesting phonetic details of the target word alone were not sufficient.

Table 3. Classification accuracy of the RandomForest classifiers for each group

Group	Focus types	Error rate	Accuracy	Overall
AE	BF	58.7%	41.3%	
	CF	53.1%	46.9%	45.5%
	NF	51.7%	48.3%	
CE1	BF	45.0%	55.0%	
	CF	66.3%	33.8%	47.1%
	NF	47.5%	52.5%	
CE2	BF	46.3%	53.8%	
	CF	61.3%	38.8%	49.6%
	NF	43.8%	56.3%	

In addition, we extracted the importance indices for each feature in three models and plotted the top five of them in Figure 2, Figure 3 and Figure 4. *MeanDecreaseAccuracy* indicates how much the model accuracy decreases if we drop that variable and *MeanDecreaseGini* calculates feature importance based on the Gini impurity index. For both indices, the higher the value the more important that feature is in the model.

For the three groups in this study, only the coda duration and the duration of the rising contour remained the top five most important features in term of both importance indices, suggesting that these two features were most stable in classifying different information foci across AE and CE. Other features varied between two importance indices and across different groups.

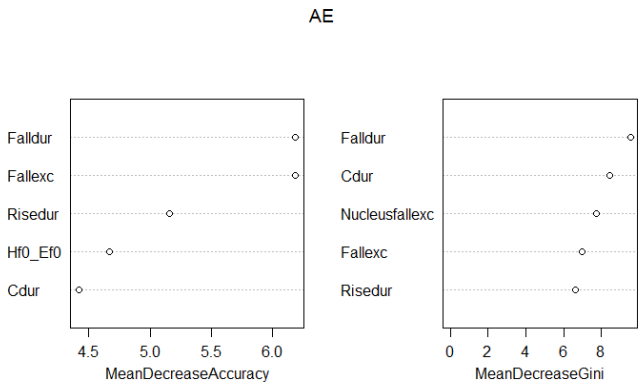

Fig. 2 Importance ranking of phonetic features in the RandomForest trained and tested with AE data.

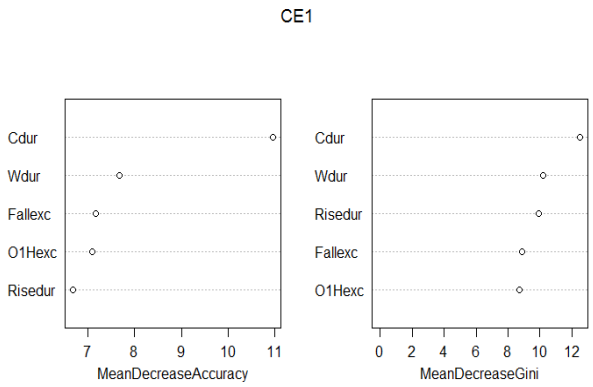

Fig. 3 Importance ranking of phonetic features in the RandomForest trained and tested with CE1 data.

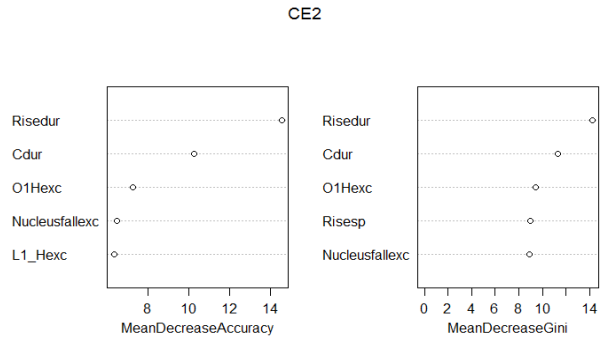

Fig. 4 Importance ranking of phonetic features in the RandomForest trained and tested with CE2 data.

In addition to the RandomForest models, we also built three SVM models for each group. For example, an SVM model for AE listeners was specified as in (3).

```
svm_AE = svm (focus_type ~ .,
              data = datasetAE,
              type = 'C-classification',
              kernel = 'radial',
              cost = "2",
              cross = 10)
```
(3)

For the SVM models, the overall accuracy for AE, CE1 and CE2 were 42.78%, 48.33 and 49.16% respectively. The results from SVM models resembled those produced by the RandomForest classifiers.

B. Cross-language classification

In order to simulate how AE listeners would perceive the target words produced by CE1 and CE2 learners, an SVM model was trained with all features of the AE data and tested with data from two Chinese English learner groups separately. The overall accuracy was above chance level for both CE1 and CE2, and the difference between CE1 and CE2 was small (Table 4). Within each group, BF was relatively easy to classify while NF and CF were more difficult, reflected in lower accuracy.

Table 4. Cross-language classification accuracy

Group	Focus types	Accuracy	Overall
CE1	BF	53.5%	45.1%
	CF	41.7%	
	NF	40.3%	
CE2	BF	49.4%	44.4%
	CF	36.0%	
	NF	47.8%	

V. DISCUSSION

The general accuracy by machine learning algorithms of classifying different information foci based on phonetic features of the target word alone was not high. This, from a perception perspective, supports that there are no robust phonetic cues in the target word itself for disambiguating among broad, narrow and contrastive foci [12], [13]. Nonetheless, the high ranking in terms of importance indices of the duration related features was in line with previous literature [4], [5], [8].

However, the lack of robust phonetic features in the target word foci does not mean that phonetic information cannot be sufficient for modeling information focus. Rather, phonetic information beyond the target word may help to disambiguate different foci. In [3], the f0 peaks of all post-focus words were lower than those of the same words in the neutral-focus sentence. Including phonetic information in post-focus words may improve the classification accuracy.

Classifying second language learners' production by the SVM models trained with AE data generated accuracy comparable to the results of classifying AE production. This suggests that despite the phonetic differences between AE and CE, both CE groups produced target words that were comparable to AE at least in terms of how well these words can be distinguished by the SVM models. And the difference between two CE groups was very small, indicating English proficiency did not alter the pattern of focus realization by CE leaners.

VI. CONCLUSIONS

In conclusion, machine learning models, such as SVM and RandomForest, can be used to investigate focus realization by native speakers and second language learners. The phonetic features of the target word can disambiguate three information focus types above chance level, but to improve classification accuracy, phonetic information beyond the target word need to be included in the model. CE learners produced comparable target words to AE speakers in terms of how well it can be

distinguished by the AE-trained SVM model. But proficiency level did not affect the production of different foci as evaluated by the SVM models. This research has implications for theories of intonation modeling and second language teaching and testing.

REFERENCES

[1] D. R. Ladd, "Phonological features of intonational peaks," *Language*, vol. 59, p. 721, 1983.

[2] C. Gussenhoven, "Types of focus in English," in *Topic and focus: Cross-linguistic perspectives on meaning and intonation*, vol. 82, C. Lee and M. Gordon, Eds. Dordrecht: Springer Science & Business Media, 2007, pp. 83–100.

[3] Y. Xu and C. X. Xu, "Phonetic realization of focus in English declarative intonation," *J. Phonetics*, vol. 33, p. 159, 2005.

[4] J. Hanssen, J. Peters, and C. Gussenhoven, "Prosodic effects of focus in Dutch declaratives," 2008.

[5] D. Sityaev and J. House, "Phonetic and phonological correlates of broad, narrow and contrastive focus in English," presented at the 15th ICPhS, 2003, vol. 1822.

[6] M. Ueyama and S.-A. Jun, "Focus realization of Japanese English and Korean English intonation," *UCLA Working Papers in Phonetics*, pp. 110–125, 1996.

[7] M. Swerts and S. Zerbian, "Intonational differences between L1 and L2 English in South Africa," *Phonetica*, vol. 67, no. 3, pp. 127–146, 2010.

[8] X. He, *Mandarin Accented Dutch Prosody*. Utrecht: LOT, 2012.

[9] W. Styler, "On the Acoustical and Perceptual Features of Vowel Nasality," PhD Thesis, University of Colorado at Boulder, 2015.

[10] A. Liaw and M. Wiener, "Classification and regression by randomForest," *R news*, vol. 2, no. 3, pp. 18–22, 2002.

[11] E. Dimitriadou, K. Hornik, F. Leisch, D. Meyer, and A. Weingessel, "Misc functions of the Department of Statistics (e1071), TU Wien," *R package*, vol. 1, pp. 5–24, 2008.

[12] M. A. K. Halliday, *Intonation and grammar in British English*, vol. 48. Walter de Gruyter GmbH & Co KG, 2015.

[13] D. Bolinger and D. L. M. Bolinger, *Intonation and its parts: Melody in spoken English*. Stanford University Press, 1986.