EXPLORING SECOND LANGUAGE ATTITUDINAL PROSODY WITH A MACHINE LEARNING APPROACH

Xiaolin Xu, Izabelle Grenon

University of Tokyo
shelynxu@gmail.com

ABSTRACT

This research explored the language universal and specific effects in the second language (L2) prosody. Our multivariate analysis covers over 20 prosodic features from phonetics (e.g. F0) and phonology(e.g. pitch pattern) aspects. The importance of features was defined by a linear Support Vector Machine (SVM) model performing classification on both speakers' groups and attitudes. Speech data were recorded from Japanese speakers (n=34) and English speakers (n=16) having semi-spontaneous question-reply conversations with neutral, strict, and gentle attitudes. Our model successfully classified the gentle attitude from neutral as well as the two speaker groups. The feature weights indicate a significant role of phonetic features with attitudinal prosody. Japanese speakers with higher English proficiency tend to change sentence-level F0 and intensity with a gentle attitude similar to the English natives. While the lower English proficiency group prefers syllable-level features which is the same as the case in L1 Japanese.

Keywords: second language acquisition, attitudinal prosody, SVM,L1 transfer, pitch

1. INTRODUCTION

Human beings have developed the ability to produce various prosody and intonation patterns throughout evolution, with which we can directly express pragmatic information such as attitudes (e.g. strict, friendly) and emotions (e.g. comfort, distress)[1]. This crucial ability enriched our social interactions and facilitates close relationships between individuals[2]. With the promotion of globalization and cultural exchange, cross-language pragmatics receives more attention. Research on prosody cross-language not only helps us to avoid misunderstandings but also deepens our knowledge of language acquisition, and supports automatic speech modeling technologies development.

1.1. language-universal aspect of prosody

To a large extent, prosodic features are considered universal across different languages. For instance, people can distinguish a question from a statement (sentence mood) when listening to languages they have no proficiency in [3]. The theory of the Biological Codes proposed by Gussenhoven in 2004 with the previous work of [4] indicates that different languages, even of separate typologies, share unified functions of reflecting attitudes. Within this framework, the low-level fundamental frequency (F0) and the falling-shaped pitch contour are associated with an unfriendly attitude while the high pitch level and rising contour tend to be produced when conveying a friendly attitude. It is supported by previous research on many languages including English and Japanese [5].

1.2. language-specific differences of prosody

On the other hand, the specific differences between the two languages are supposed to cause difficulty in L2 prosody acquisition. Previous studies reported the first language (L1) transfer due to language-specific differences. In the Intonation learning theory (LILt) [6], the differences can be categorized as phonological, and phonetic dimensions. Phonologically, when a particular pitch pattern of L2 doesn't exist in L1, the learners tend not to learn that L2 pitch pattern [7][8][9]. Phonetically, F0 peaks produced by learners are either higher or lower than those produced by native speakers; Learner's steepness of pitch contour is either greater or smaller than native speaker's [10][11].

1.3. machine learning (ML) models

Models such as attention-long Long Short-Term Memory (LSTM), Convolutional Neural Network (CNN), Hidden Markov model(HMM) and Support Vector Machines (SVM) were widely adopted on robotic speech emotion recognization[12][13], automatic L2 English proficiency evaluation [14]

with Test of English as a Foreign Language (TOEFL) database and Clinical detection[15] of Depression and Autism [16]with utterance level acoustic features. Most of the models reached a high accuracy with a full aspect of features covering intensity, F0, speech rate reflected by duration, voice quality, and zero crossing rate features.

1.4. The significance of this study

Most of the linguistic studies on prosodic features focus on the syllable level in order to interpret each specific feature's effect in a clear domain. Multiple phonetic and phonological features have been proven significant in L2 prosody acquisition in separate research but difficult to rank which feature matters the most in each case. In contrast, studies with machine learning models are able to make accurate predictions with a massive amount of features at once. But they focus more on accuracy increase in dealing with big-size utterance corpus. As a drawback, background factors like scenarios could not be controlled without an experimental design. In addition, it is difficult to interpret each feature's effect with black box models. ML models on speech emotion detection are mostly trained on monolingual data, and models involving L2 barely cover pragmatic functions of conveying emotions or attitudes. Particularly noteworthy is the study of cross-language prosody of Autism[16] pointing out the language-specific prosody differences in Autism classification. It found that classifications using F0 values were significant for English, but not for Cantonese speakers, which is potentially due to long exposure to their native tone language. Regarding the comparison of previous studies above, it's significant for our research to bridge the gap between L2 prosody research and acoustic ML model development territories. To achieve this goal, we classify attitudes and speaker groups, utilizing an ML approach on sentence and syllable-level prosodic features from a controlled experimental design. We focus on two languages of separate typologies, English and Japanese, and particularly choose a linear SVM model. The SVM models perform well and were frequently adopted by classification tasks with acoustic features [16][14], and the linear SVM is rather transparent (c.f. black box) since it outputs a weight interpreting the importance of each feature[17].

1.5. Research question

Our research questions can be narrowed down to:

- What are the key important prosodic features when conveying different attitudes?
- Do these important features vary when produced with a second language?
- If the features varies, could it be explained by a language-specific difference (L1 transfer)?

For the record, the feature's importance is defined at the ML model level, which should be treated differently from the human perception level.

2. HYPOTHESIS AND METHOD

2.1. Hypothesis

English is a stressed-time language while Japanese is a pitch-accent language where the specific pitch patterns are decided by a particular lexical word. The two languages differ both phonologically and phonetically: Phonologically, English interrogative utterance typically has a rising pitch pattern over the focused word with the Yes/No question, and a falling pitch pattern with the Wh question. Japanese doesn't have such variation. When expressing an attitude, English speakers use the pitch pattern over the focused word's syllable. In Japanese, pitch pattern over the final syllable (Boundary pitch movement, BPM) of utterance functions to convey attitudes[18]. Phonetically the Japanese overall pitch level is higher than in English[19]. According to [11], the English interrogative rising pitch contour produced by Japanese learners has a steeper slope than native speakers. Additionally, Japanese has a higher speech rate generally[20]. 1. Based on the language-universal Biological code framework, we first expect that pitch features (measured by F0 values of focused syllable and utterance domains) are important when conveying positive and negative attitudes in both L1 English and L1Japanese. 2. L1 Japanese will have a higher pitch level and a steeper slope than L1 English with a higher speech rate (measured by focused syllable duration, sentence duration, and duration-per-syllable over utterance) 3. When classifying Japanese L2 English from native speakers' L1 English, the language-specific difference of hypothesis 2 should be more important than the language-universal features of hypothesis 1.

2.2. Participants and procedure

34 participants speak in Tokyo Japanese accent. Their English proficiency ranges from Elementary (n=12), Intermediate(n=15) and Advanced (n=6).16 English speakers are native to North American English. They were traveling in Tokyo at the time of data collection. All participants produced 10 Yes/No and 10 Wh questions, respectively in

Neutral, Strict, and Gentle attitudes from slides. Pictures of faces of corresponding attitudes were also shown to speakers to better trigger attitudes. The focus word is underlined on slides (half in the middle of the sentence and half at the end). To simulate natural conversations, utterances were produced by finishing semi-spontaneous production tasks, in which the participants asked each question as having a real conversation with the researcher after memorizing a short question from the slide and moving to the next question after having a short answer. Japanese speakers asked the same questions translated into Japanese in the same procedure.

2.3. Features extraction

The phonological pitch pattern is annotated as rising (R), falling (F), level (L), and combinations of them (e.g.RF stands for rising-falling pattern) within focus syllable using the original python script. Then we visually re-check the annotation on Praat.Phonetic features of F0 and intensity's max, min, range, and mean values were extracted from Praat on syllable and utterance levels. The range, slope, and deviation values were also calculated as subfeatures. We also recorded the duration, intensity, jitter, and zero crossing rate values and input them into the following model together with focus word position information (middle or final in a sentence.)

2.4. SVM classification

As the figure of the modeling process shows, three types of classifications were conducted to explore how language-universal and specific prosodic features work together on L2 production. We first classify utterances under Gentle and Strict from Neutral ones with L1 Japanese and L1 English together and separately, to see the common prosodic cues used in English and Japanese with and without an attitude.Secondly, we seek the specific features that distinguish the two languages with and without an attitude.Finally, speaker groups of English/Japanese, as well as Japanese speakers of three English proficiencies were classified with L1 + L2 English speech data together. The important features extracted based on SVM weights were compared with other important features from the first two classification types. Outliers were filtered by rescale feature data within the 1st quartile and the 3rd quartile (75th quantile) with RobustScalar of the Sklearn library, then standardized with Sklearn's StandardScaler function before input into mode. Since feature data consisted of continuous data with different units and scales (e.g. F0, intensity, duration), as well as binary data of focus position sentence mood, and nominal data of pitch patterns. The important feature was extracted from classification models that were evaluated as effective based on 10-fold cross-validation.

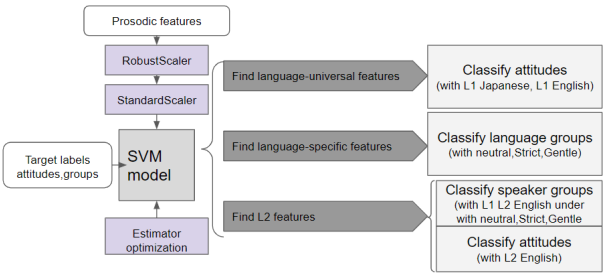

Figure 1: The SVM classification process.

3. RESULTS

The following chart shows the top-weighted features of each effective (Accuracy> 0.6, accuracy - null accuracy >0.1, cross-validation AUC >0.6) classification case. These values are the average from 1000 times ML sampling iterations.

Classification types	Data scope	Top weighted features
Attitude Gentle/ Neutral	L1 Japanese	Nmax_F0 Nmean_F0 Nduration
	L1 English	Smean_F0 Smean_intensity Sdevi_F0
	L2 English Advanced	Sdevi_F0 Smax_intensity Smean_F0
	L2 English Elementary	Nmean_F0 Nmin_F0 Sdevi_pitch
Languages	Neutral	Smean_intensity Nmean_intensity Nunvoiced
Speaker groups L1/L2 English	Neutral	Smean_F0 Sdevi_F0 Nmax_F0

Figure 2: SVM models chart of means values of 1000 times itterations

The classification of languages (L1 Japanese from L1 English) is with the highest accuracy of over 0.9. The results indicate the intensity feature on both

nucleus syllable (Nmean intensity) and sentence levels (Smean intensity) is the most important prosodic feature to distinguish the two languages with the SVM model. Attitudes classification was only on Gentle and Neutral utterances only. English speakers' sentence-level F0 and intensity features (Smean F0, Smean intensity, Smax intensity) seem to be more dominant. In L1 Japanese, on the other hand, nucleus syllable level F0 and speech rate features (Nmax F0, Nmean F0, Nduration) are most important, for instance when conveying a Gentle attitude. They tend to produce longer duration and lower speech rate. The phonological features also contribute to the model, Gentle utterances were produced with more rising nucleus patterns (PitchPattern R) in English as well as more rising BPM (Tone R) in Japanese after ruling out the unvoiced data. But still, the pitch pattern features are not as highly weighted as phonetic features. This set of results also presents the primary features of Japanese speakers' L2 English compared with native English speakers and L1 Japanese. L2 English of all three proficiency groups doesn't have an effective classification though the Advanced and Elementary English level groups do. The two proficiency groups have opposite tendencies. Japanese speakers with higher English proficiency tend to change sentence-level features (Sdevi F0, Smax intensity, Smean F0) with a gentle attitude similar to the English natives. While the lower English proficiency group prefers syllable-level features which is the same as the case in L1 Japanese. Finally, the SVM can classify L1 and L2 English groups with both sentence and syllable level F0 features (Smean F0, Sdevi F0, Nmax F0), which is consistent with the previous study that Japanese speakers' overall pitch level is higher than English. Our results show clear language-universal use of higher pitch levels associated with a positive attitude and language-specific differences in syllable and sentence domains.

4. DISCUSSION

Our research re-examined the language-universal and specific differences effects on attitudinal prosody. Overall, partially as the Biological code has summarized, both Japanese and English adopt higher F0 with positive attitudes. It can be utilized by ML models. But we also find that F0 ques in the nucleus syllable domain work better with Japanese while the sentence domain features are more primary in English. And Japanese speakers are able to learn much differently when proceeding

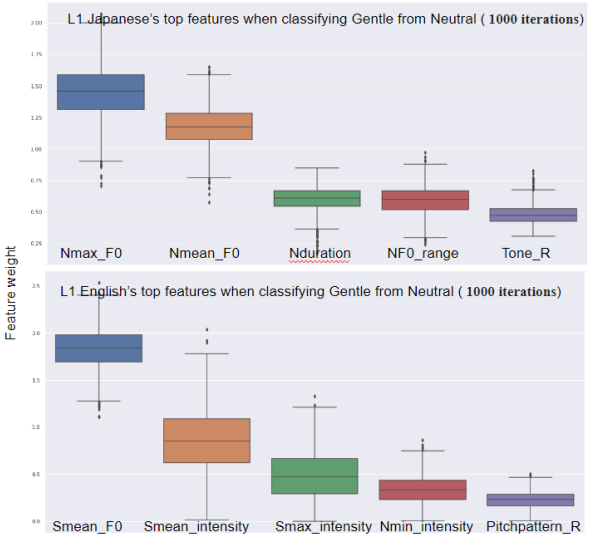

Figure 3: Different primary prosodic features of English and Japanese.

to a higher level. These findings not only help ML models recognize human speech better but also provide an interpretation from an L1 transfer perspective. One question is why the negative attitude was not effectively classified.[21] probably shows us the answer that negative emotions like anger can be divided into hot anger and cold anger. When someone is suppressing their anger, the pitch and intensity will go into low-level, so-called cold anger. The explosive type of anger should be the opposite. Therefore it was difficult to classify the two subcategories mixed together. Another problem with our research is gender unbalanced data. We find that when restricted to female, the model accuracy tends to be higher. Gender should be considered an important factor in future studies. More for the futural direction, it would be meaningful to investigate the difference between the ML model and human speech perception, which features that matter might be quite different. And mental disorder diagnosis using prosody may be conducted with a different standard with blingos regarding the cross-language differences found in our research.

5. CONCLUSION

Our research shows that the phonetic features of F0 are significant during the ML classification, but still, they need to work together with intensity speech rate and phonological pitch pattern to achieve the best predictions. When classifying second language prosody, the effect from L1 can not be ignored. The syllable level features are more salient in the case of Japanese speakers producing L2 English.

6. REFERENCES

[1] E. Altenmuller, S. Schmidt, and E. Zimmermann, *The evolution of emotional communication: From sounds in nonhuman mammals to speech and music in man*. OUP Oxford, 2013.

[2] J. K. Hall, J. Hellermann, and S. P. Doehler, *L2 interactional competence and development*. Multilingual Matters, 2011, vol. 56.

[3] C. Gussenhoven and A. Chen, "Universal and language-specific effects in the perception of question intonation," in *6th International Conference on Spoken Language Processing (ICSLP 2000)*, 2000, pp. 91–94.

[4] J. J. Ohala, "Cross-language use of pitch: an ethological view," *Phonetica*, vol. 40, no. 1, pp. 1–18, 1983.

[5] C. Gussenhoven, *The phonology of tone and intonation*. Cambridge University Press, 2004.

[6] I. Mennen, "Beyond segments: Towards a l2 intonation learning theory," in *Prosody and language in contact*. Springer, 2015, pp. 171–188.

[7] E. Grabe and M. Karpinski, "Universal and language-specific aspects of intonation in english and polish," in *Proceedings of the 15th International Congress of Phonetic Sciences*, vol. 39, 2003.

[8] A. Chen, *Universal and language-specific perception of paralinguistic intonational meaning*. Utrecht: LOT, 2005.

[9] M. Zięba, *The acquisition of English intonation by Polish adult learners*. Wydawnictwo Naukowe Państwowej Wyższej Szkoły Zawodowej w Nowym Sączu, 2013.

[10] M. Jilka, "The contribution of intonation to the perception of foreign accent," Ph.D. dissertation, Universitat Stuttgart, 2000.

[11] M. Ueyama and S.-A. Jun, "Focus realization of japanese english and korean english intonation," *UCLA Working Papers in Phonetics*, pp. 110–125, 1996.

[12] H. Meng, T. Yan, F. Yuan, and H. Wei, "Speech emotion recognition from 3d log-mel spectrograms with deep learning network," *IEEE access*, vol. 7, pp. 125 868–125 881, 2019.

[13] T. Seehapoch and S. Wongthanavasu, "Speech emotion recognition using support vector machines," in *2013 5th international conference on Knowledge and smart technology (KST)*. IEEE, 2013, pp. 86–91.

[14] O. Kang, D. O. Johnson, and A. Kermad, *Second Language Prosody and Computer Modeling*. Routledge, 2021.

[15] M. Higuchi, M. Nakamura, S. Shinohara, Y. Omiya, T. Takano, D. Mizuguchi, N. Sonota, H. Toda, T. Saito, M. So *et al.*, "Detection of major depressive disorder based on a combination of voice features: An exploratory approach," *International journal of environmental research and public health*, vol. 19, no. 18, p. 11397, 2022.

[16] J. C. Lau, S. Patel, X. Kang, K. Nayar, G. E. Martin, J. Choy, P. C. Wong, and M. Losh, "Cross-linguistic patterns of speech prosodic differences in autism: A machine learning study," *PloS one*, vol. 17, no. 6, p. e0269637, 2022.

[17] D. Mladeni, J. Brank, a. Grobelnik, and N. Milic-Frayling, "Feature selection using linear classifier weights: interaction with classification models," in *Proceedings of the 27th annual international ACM SIGIR conference on Research and development in information retrieval*, 2004, pp. 234–241.

[18] Igarashi, "Handbooks of japanese language and linguistics.chapter 13: Intonation," pp. 525–568, 2015.

[19] Y. Ohara, "Gender-dependent pitch levels: A comparative study in japanese and english," in *Locating Power. Proceedings of the Second Berkeley Women and Language Conference, 1992*, 1992.

[20] C. Coupé, Y. M. Oh, D. Dediu, and F. Pellegrino, "Different languages, similar encoding efficiency: Comparable information rates across the human communicative niche," *Science advances*, vol. 5, no. 9, p. eaaw2594, 2019.

[21] F. Biassoni, S. Balzarotti, M. Giamporcaro, and R. Ciceri, "Hot or cold anger? verbal and vocal expression of anger while driving in a simulated anger-provoking scenario," *Sage Open*, vol. 6, no. 3, p. 2158244016658084, 2016.