INTERSPEECH 2019
September 15–19, 2019, Graz, Austria

Low resource automatic intonation classification using gated recurrent unit (GRU) networks pre-trained with synthesized pitch patterns

Atreyee Saha[1], *Chiranjeevi Yarra*[2], *Prasanta Kumar Ghosh*[2]
[1]Electrical Engineering, Jadavpur University, Kolkata 700032, India
[2]Electrical Engineering, Indian Institute of Science (IISc), Bangalore 560012, India
as4297492@gmail.com,{chiranjeeviy,prasantg}@iisc.ac.in

Abstract

Second language learners of British English (BE) are typically trained to learn four intonation classes – Glide-up, Glide-down, Dive and Take-off. We predict the intonation class in a learner's utterance by modeling the temporal dependencies in the pitch patterns with gated recurrent unit (GRU) networks. For these, we pre-train the GRU network using a set of synthesized pitch patterns representing each intonation class. For the synthesis, we propose to obtain pitch patterns from the tone sequences representing each intonation class obtained from domain knowledge. Experiments are conducted on speech data collected from experts in a spoken English training material for teaching BE intonation. The absolute improvements in the unweighted average recall (UAR) using the proposed scheme with pre-training are found to be 4.14% and 6.01% respectively over the proposed approach without pre-training and the baseline scheme that uses hidden Markov models (HMMs).

Index Terms: intonation classification, computer assisted language learning, LSTM with pre-training, synthetic pitch for intonation

1. Introduction

In spoken communication, intonation refers to the modulation of pitch that gives meaning to an utterance [1] and it acts as an emotional indicative of the speaker [2]. In addition, for second language (L2) learners, intonation is an important prosodic aspect in learning, because an incorrect intonation can result in miscommunication. Hence, in spoken L2 training, for example in learning British English (BE), L2 learners require to learn BE intonation for a better spoken communication. Though the intonation of BE varies across different geographic regions [3], L2 learners are initially trained to learn four different patterns of BE intonation in the received pronunciation [4–6] – Glide-up, Glide-down, Dive and Take-off [4, 7], referred to as intonation classes. Later, they are trained to add finer changes to those patterns [4] to produce more detailed intonation variations. In this work, we propose models to classify those four classes in the BE expert's intonation. Thus, these models could be useful in the L2 training similar to the work proposed by Witt [8, 9], where the quality of phonemes in L2 learner's utterance has been assessed using a model built from expert's data.

In general, intonation is defined by a sequence of discrete patterns called tones [3–7, 10–12]. Although the last tone in the sequence, called nuclear tone [3, 4], plays a critical role in the intonation class, all tones in the sequence together convey the meaning [4, 7, 12]. For example, the nuclear tone in Glide-up and Take-off are the same, but these classes are discriminated based on the entire tone pattern. Most of the existing works have studied the variations of intonation among different nativities [13–18] and variations of BE intonation across the nativities [3, 10, 11]. However, a few works have addressed the problem of intonation assessment of L2 learners [19, 20] and the problem of intonation classification [21–23]. Most of the works on intonation assessment and classification have considered temporal structures in either pitch or the tone sequence in an utterance [22, 23]. This is because intonation of an utterance depends on the entire sequence of tones and each tone in the sequence depends on pitch variations within the tone [4, 7, 12].

Li et al. [21] have used two tones at the end of the utterance and performed intonation classification using deep neural network (DNN) models. Instead of only last two tones, Ke et al. [19] have considered tone duration based features from all the tones for assessing the L2 learners' intonation. Yarra et al. [23], have modelled the temporal dependencies in the tone sequence of an utterance for the intonation classification task. However, in these tone based modeling a small error in estimating tone sequence could cause large degradation in the classification [22, 23]. On the other hand, Arias et al. [20] have assessed the L2 learners using pitch contours from learners and experts. Further, the temporal dependencies in the pitch patterns have been used in the intonation classification task [22]. Among these works, the works considering the temporal dependencies in pitch patterns have been shown to be effective in the intonation classification task. Yet, most of these approaches do not consider any deep learning based sequential modelling techniques such as recurrent neural networks (RNNs). In this work, for the intonation classification task, we use RNNs to capture the temporal dependencies in both the pitch pattern and the tone sequences. However, in order to avoid the degradation due to the errors in estimating tones from the pitch, which typically happens in typical tone sequence based modelling, we construct the tone sequence from the domain knowledge belonging to the intonation classes.

Typically, RNN based modelling requires large amount of data [24]. However, annotating the intonation class labels requires highly skilled experts thereby limiting the data size. In order to obtain better model under this low resource scenario, it requires memory units in RNN with less number of parameters and a good initialization point for training. These can be achieved respectively by considering gated recurrent unit (GRU) as the RNN unit [25] and by pre-training the model using artificial data that closely matches the distribution of the training data.

In general, each intonation class is characterized by a set of tone sequences, which can be obtained by domain knowledge. For example, the intonation class Take-off always ends in a rising tone that may span multiple segments, and it is preceded by a sequence of low level tones. Considering this, a set of representative samples of Take-off class can be designed with a rise tone at the end preceded by a sequence of low tones of different lengths. Similarly, such sets can be designed for the other intonation classes as well. Considering these knowledge based tone

We thank the Department of Science & Technology, Government of India and the Pratiksha Trust for their support.

http://dx.doi.org/10.21437/Interspeech.2019-2351

sequences, we show that the pitch patterns can be synthesized and used for pre-training. Hence, this pre-training setup allows us to model the temporal dependencies in the tone sequences without actually estimating them from test utterances as done in a typical tone based modelling pipeline.

Pitch may have unwanted variations which may not be present in the intonation class specific pattern. In this work, we hypothesized that a time-distributed neural network layer (TDL) that performs transforms on the pitch pattern would suppress such unwanted variations when jointly trained with GRU. Experiments are performed on the speech data collected from a spoken English training material for teaching BE intonation [7]. We consider the work proposed by Yarra et al. [22] as the baseline scheme. The absolute improvement in unweighted average recall (UAR) [26] is found to be 6.01% with the proposed method compared to that with the baseline scheme. The highest average UAR is also found to be 4.14% more than the UAR obtained without pre-training indicating the benefit of the proposed pre-training.

2. Database

In this work, the speech data is considered from a spoken English training material [7] used for teaching BE. The speech recordings selected for our experiments contain all the utterances of intonation phrases belonging to intonation lessons. The entire speech recording is manually segmented into individual speech files belonging to every utterance. Further, the annotated text transcriptions are obtained along with the respective intonation class label and the tone sequence for each utterance. In the speech data, the total number of utterances is 233 out of which 50, 68, 82 and 33 belong to Glide-up, Glide-down, Dive and Take-off intonation classes respectively. The entire speech data considered in this work has been spoken by one male and one female native BE speaker. To the best of our knowledge, there is no larger speech data that has these four intonation class labels annotated by experts. This could be because recording and labeling of such corpora require highly trained specialists, which, in turn, limits the size and the availability of such corpora.

3. Proposed approach

Figure 1 shows the three major stages involved in the proposed approach. In the first stage, a 3-dimensional (3D) feature sequence ($f(t),\ 1 \leq t \leq T$) is computed from the speech signal, where T is the total number of frames in the signal. In the second stage, we perform pre-training to obtain parameters for initializing the GRU network in the classifier in three steps. The first step derives the tone sequence ($\tau(n),\ 1 \leq n \leq N$) containing discrete symbols of an arbitrary length N for each intonation class using the class specific knowledge. The second step synthesizes a 2D artificial feature sequence ($\tilde{f}(t),\ 1 \leq t \leq \tilde{T}$) from the tone sequence of length N, where $\tilde{T}$ is the number frames obtained based on the range of typical syllable duration and N. The third step trains the GRU network with $\tilde{f}(t)$ and obtains parameters for the initialization. In the third stage, we classify the feature sequence ($f(t)$) into one of the four intonation classes with a classifier containing a TDL and a GRU network. The first step estimates the posterior probability of each class given $f(t)$ by jointly training TDL and GRU network. The TDL is used to obtain a 2D sequence from 3D $f(t)$ and the GRU network is initialized with parameters from the pre-training. With the joint model, we believe that the TDL could suppress the unwanted variations in $f(t)$ that are not present in 2D $\tilde{f}(t)$. The second step estimates the class with highest probability as the predicted class.

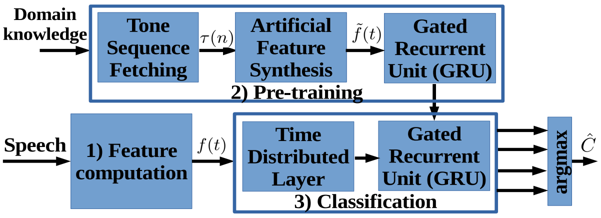

Figure 1: *Block diagram summarizing the stages involved in the proposed approach.*

3.1. Feature computation

Following the work by Yarra et al. [22], we consider a 3D feature sequence $f(t) = [f_1(t), f_2(t), f_3(t)]^T$. $f(t)$ is an estimated pitch contour in Mel scale with mean and range normalization over every utterance. $f_2(t)$ is the derivative of $f_1(t)$. $f_3(t)$ is a sequence of confidence score in estimating $f_1(t)$. It has been observed that the score values are lower when there are pitch estimation errors and vice versa. Thus, by having the score sequence in $f(t)$, it has been hypothesized that the classifier trained with $f(t)$ could model the dependencies between the score sequence and the estimated pitch contour and minimizes the inaccuracies caused due to the estimation errors. In this work, using $f(t)$, we propose a classifier that would learn transformations representing those dependencies more explicitly using TDL with the help of pre-training.

3.2. Pre-training

RNNs have large number of parameters compared to a standard MLP network when both the networks are considered with the same number of units. This is due to the multiple internal gating functions present in each RNN unit. Hence, ideally, a large set of training data samples is required to learn the parameters for better generalizability of these networks [24]. However, a good initialization could reduce a need for large data size. It has been shown that pre-training with artificial data set allows a better initialization of the classifier [27]. In this work, we pre-train the classifier by synthesizing a 2D artificial feature sequence and explore its benefit in minimizing the inaccuracies due to the pitch estimation errors.

3.2.1. Tone sequence fetching

A tone sequence ($\tau(n)$) comprises of discrete symbols called tones. Typically, the tone is associated with a syllable and there are four tones – rise (R), fall (F), low (L) and high (H). In some cases, a fifth tone is also considered, which is mid (M) tone. In R and F tones, pitch changes from a low to a high value and from a high to a low value respectively. In L, M and H tones, the pitch is at a low, an average and a high value of the normalized pitch contour respectively. It is observed that the temporal dependencies in the tone sequence are the representative of each intonation class. However, the utterances with the same number of syllables (N) could have different tone sequences. For example, two exemplary utterances of Take-off class with $N = 4$ from the corpora considered have the following two different tone sequences – $\{L, L, L, R\}$ and $\{L, L, R, R\}$. However, the temporal pattern of a low (L) tone followed by a rise (R) tone is sufficient to identify the Take-off class. Similarly, this holds for other class as well.

In the corpora, the total percentage of such different tone sequences are found to be 64.00%, 66.18%, 45.12% and 75.76% across all N in Glide-up, Glide-down, Dive and Take-off classes respectively. The percentages below 100% in all four classes indicate that different spoken texts within a class have the same tone sequence. In addition, we also observe that the same spoken text is uttered in the tone sequences belonging to different intonation classes. From these three observations, we assume that the tone sequence is a supra-segmental information that is embedded onto a spoken text. Thus, we hypothesize that the tone sequences of each intonation class can be collected irrespective of the spoken text but based on the knowledge of relation between the temporal dependencies in tone sequences and the intonation classes. Considering this hypothesis, we propose to synthesize feature sequences $\tilde{f}(t)$ that approximately resemble $f(t)$ derived from the tone sequences independent of the spoken text. Further, using $\tilde{f}(t)$, we pre-train the GRU network and propose to initialize the classifier which takes $f(t)$ as the input.

3.2.2. Feature synthesis

In order to obtain $\tilde{f}(t)$, first, we synthesize its 1-st dimension sequence ($\tilde{f}_1(t)$) which is of similar nature as that of $f_1(t)$. Later, we compute derivative on $\tilde{f}_1(t)$ and consider it as the 2-nd dimension sequence ($\tilde{f}_2(t)$) of $\tilde{f}(t)$. With this, we hypothesize that the two dimensions of $\tilde{f}(t)$ would have characteristics similar to that of $f_1(t)$ and $f_2(t)$. Let a tone sequence be $\{\tau_i, 1 \leq i \leq N\}$, where $\tau_i \in \{L, M, H, R, F\}$ is the tone of i-th syllable with duration d_i. The d_i is chosen randomly between the minimum and maximum duration (d_{min} and d_{max}), where d_{min} and d_{max} are computed using the range of syllable rate in the corpora considered and it is found to be 1.26 to 6.47 syllables per second. Thus, 155ms and 793ms are assigned to d_{min} and d_{max} respectively. Further, in the synthesis of $\tilde{f}_1(t)$, we propose to use the following parameters – p_L, p_H and $p_M = \frac{1}{2}(p_L + p_H)$, where p_L and p_H are computed by averaging the least and highest $f_1(t)$ values across all utterances considered in the corpora. The values of p_L, p_H and p_M are found to be -0.57, 0.44 and -0.06 respectively. Considering these values, we synthesize $\tilde{f}_1(t)$ for a tone sequence as follows:

1. Divide N syllable segments into K, where $K \leq N$, sub segments where each sub-segment consists of consecutive syllable segments that belong to the same tone.
2. $\tilde{f}_1(t)$ is assigned with pL, pM and pH in a sub-segment, when the tone in that sub-segment is equal to L, M, and H respectively. If the current sub-segment has either R or F tone, the $\tilde{f}_1(t)$ in this sub-segment is obtained by linearly interpolating the $\tilde{f}_1(t)$ values at the previous and next sub-segments. If either the previous or the next sub-segment is absent, for interpolation, the $\tilde{f}_1(t)$ at the begin and end of the current sub-segment are considered as $pL(pH)$ and $pH(pL)$ respectively when the tone in the current sub-segment is $R(F)$.
3. Add a white Gaussian noise at 20dB SNR to the synthesized $\tilde{f}_1(t)$ obtained from steps 1 and 2.

It is to be noted that the step 3 is used for better generalizability of the classifier, since real data mostly contains small variations around the ground-truth pitch values.

Figure 2 shows the $\tilde{f}_1(t)$, $f_1(t)$ and $f_3(t)$ for an exemplary tone sequence, $\{L, L, L, R\}$, belonging to Take-off class from the corpora. From the figure, it is observed that the pattern in $f_1(t)$ and $\tilde{f}_1(t)$ are having similar trend except in the black rectangular box, where a sudden variation is observed in $f_1(t)$ from the typical trend in the tone sequence. It is also observed that the confidence score values ($f_3(t)$) are lower in the box compared to those in other locations. This suggests that the variation is due to an estimation error. However, it is to be noted that, we propose to minimize such unwanted variations using TDL before applying to GRU network.

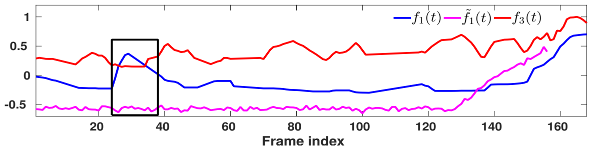

Figure 2: *Original and synthesized normalized pitch for an utterance belonging to Take-off*

3.3. Classifier

3.3.1. Time-distributed layer (TDL)

We transform the 3D input feature sequence of length T to 2D output sequence of the same length using many to many TDL. We consider TDL with one input layer and one hidden layer with two units. Each unit is considered without any activation function.

3.3.2. GRU network

In general, GRU network is similar to the typical RNN, where the hidden layers containing memory cells are considered with GRU [28]. Compared to LSTM unit cell in an LSTM network, a GRU employ a few number of parameters. Thus, with increase in the number of hidden units the computational burden in the GRU network is reduced compared to that in the LSTM network. Further, we observe that the accuracy obtained with GRU network is better than that with LSTM network for the intonation classification task. In this work, the GRU network takes 2D sequence from the TDL and produces a 4D sequence at its output. The GRU network is composed of one layer with four GRU units. Further, the output at the last time step from the GRU network is fed to a softmax layer containing four units to obtain class posterior probabilities for four intonation classes.

4. Experimental results

4.1. Experimental setup

We consider unweighted average recall (UAR) as the performance measure to evaluate the classification accuracy. We conduct the experiments in a 10-fold cross validation setup where eight folds are used for training, one fold for development (dev) and one fold for testing in a round robin fashion. We use SWIPE algorithm to estimate pitch and to obtain confidence scores [29]. We implement TDL and GRU networks using Theano [30] and Keras [31]. For the comparison, we consider the work proposed by Yarra et al. [22] as the baseline. While a domain expert can provide illustrative tone sequence for an intonation class irrespective of the utterances in the training data, we did not use such for pre-training due to unavailability of such tone sequences provided by experts. Rather, we use tone sequences from the training data as annotated by the experts.

4.2. Results and discussion

Table 1 shows the the average (standard deviation (SD)) of UARs on the test and dev sets with the baseline and the proposed approach. From the table, it is observed that the average UAR obtained using the baseline is 6.01% and 5.41% lower than those using the proposed approach with TDL and pre-training on test and dev sets respectively. This indicates that the proposed approach is better than the baseline for intonation classification task. In the table, we also show the average UAR

Table 1: *Average (SD) of UARs obtained with the baseline and the proposed approach with the combination of with & without (w/o) TDL and with & without pre-training*

	Baseline	Proposed approach			
		with pre-training		w/o pre-training	
		with TDL	w/o TDL	with TDL	w/o TDL
test	61.77 (8.6)	67.78 (9.8)	63.64 (6.9)	63.54 (5.4)	60.45 (7.3)
dev	62.32 (7.2)	67.73 (8.5)	62.67 (5.5)	63.59 (6.6)	60 (6.0)

obtained using proposed approach for each combination of with and without TDL as well as with and without pre-training. In the case when TDL is not used, there is no transformation from 3D to 2D feature sequence, thus we modify the GRU network so that it accepts directly 3D $f(t)$. In order to match model parameters with this modified GRU network and those from pre-trained model, we modify the pre-training setup so that it is trained with a 3D artificial feature sequence. We deduce the 3D feature sequence from 2D $\tilde{f}(t)$ by adding the 3-rd dimensional feature sequence as confidence score and it is chosen as one throughout the sequence length. This is because we assume that there is no error in the tone sequence thus in the synthesized feature sequence.

From the table, it is observed that the average UARs obtained using the proposed approach are higher than that with the baseline for all combinations except the that without TDL and without pre-training on both the test and dev sets. This indicates the benefit of the deep network based models for intonation classification task compared to the traditional HMM models. From the table, it is also observed that all the average UARs obtained with the proposed approach are higher when the classifier is pre-trained compared to those when the classifier is not pre-trained on both the test and dev sets. This indicates the benefit of the proposed pre-training approach using synthetically generated feature sequence from the tone sequences. Similarly, it is observed that all the average UARs obtained with proposed approach are higher than when the TDL is considered. This indicates the benefit of the TDL that performs the transformation on the feature sequence. It is interesting to observe that the highest and least UARs are obtained using proposed approach with pre-training & with TDL and without pre-training & without TDL respectively. This suggests the benefit of the proposed pre-training and TDL.

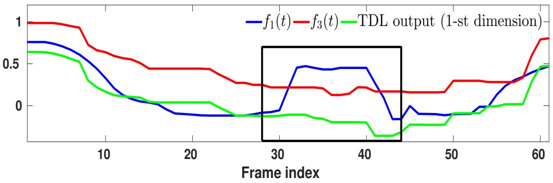

Figure 3: *Illustrative example showing the benefit of TDL*

Further, we analyse the benefit of the proposed pre-training and TDL with an illustrative example belonging to the Dive class taken from the corpora, which has the ground-truth tone sequence $\{F, R\}$. As per the tone sequence, it is expected that the pitch contour begins from a high value and decreases to a low value then again it increases to a high value. Figure 3 shows $f_1(t)$, 1-st dimension of the TDL output sequence and $f_3(t)$ for the utterance. It is to be noted that the considered utterance is correctly classified by the proposed approach with TDL and with pre-training combination and incorrectly classified by all other combinations of TDL and pre-training. In addition, we enclose the variations $f_1(t)$ that deviate from the expected ground-truth variations using the black rectangular box. From the figure, it is observed that $f_3(t)$ values are lower within the box compared to those at the outside the box. This indicates that the deviated $f_1(t)$ values could be due to the pitch estimation errors. However, it is interesting to observe that the TDL output sequence does not have such unwanted values. This suggests the benefit of the proposed TDL in suppressing the unwanted variations based on $f_3(t)$.

Table 2: *Confusion matrix obtained for the baseline and proposed approach with TDL and with pre-training. The rows and the columns represent the estimated and ground-truth classes respectively. Each cell entry is the average percentage across all the ten folds.*

	Baseline				Proposed			
	#1	#2	#3	#4	#1	#2	#3	#4
Glide-up (#1)	62.50	5.00	30.00	2.50	15.00	10.00	70.00	5.00
Glide-down (#2)	22.14	61.67	14.52	1.67	3.17	82.54	14.28	0.00
Dive (#3)	22.22	17.22	54.58	5.97	2.44	14.63	78.05	4.88
Take-off (#4)	0.00	6.67	25.00	68.33	5.71	0.00	17.14	77.14

Finally, we analyse the class specific performance of the baseline and the proposed approach with TDL & with pre-training combination using confusion matrix computed on the test set. From the confusion matrices shown in Table 2, it is observed that there is a significant improvement and decrement in the diagonal entries and off-diagonal entries respectively with the proposed approach compared to the baseline in all classes except Glide-up. This could be due to better ability of the proposed approach in handling the pitch estimation errors compared to the baseline. However, in the proposed method, Glide-up class mostly got confused with Dive, which is not the case with the baseline. This indicates that the baseline captures complementary information from the proposed approach that can discriminate well between these classes. These together suggest that a modelling technique that incorporates complementary characteristics of the baseline and the proposed approach could result in a better discrimination between the classes, thus the overall classification accuracy.

5. Conclusion

GRU network is used for the BE intonation classification task considering a pre-training with synthesized pitch contours and input from a time-distributed layer (TDL). For pre-training, we consider the tone sequences belonging to each intonation class obtained from domain knowledge. Experiments with the spoken English training material with four intonation classes reveal that the proposed scheme improves the UAR compared to the baseline scheme, which shows the benefit of the pre-training and TDL for GRU network in the intonation classification task. Further investigations are required to combine complementary properties of the proposed and the HMM based schemes for better discrimination between the Glide-up and Dive classes. Future works also include the use of linguistic features for better intonation classification.

6. References

[1] M. I. Collins, B, *The phonetics of English and Dutch*. BRILL, 2003.

[2] N. D. Cook, *Tone of Voice and Mind: The connections between intonation,emotion,cognition and consciousness*. John Benjamins Publishing Co., 2002.

[3] A. Cruttenden, "Intonational diglossia: a case study of Glasgow,"

Journal of the International Phonetic Association, vol. 37, no. 3, pp. 257–274, 2007.

[4] J. C. Wells, *English intonation: An introduction*. Cambridge University Press, 2006.

[5] B. Hlebec, "General attitudinal meanings in RP intonation," *Studia Anglica Posnaniensia*, vol. 44, pp. 275–295, 2008.

[6] A. Cruttenden, *Gimson's pronunciation of English*. Routledge, 2014.

[7] J. D. O'Connor, *Better English Pronunciation*. Cambridge University Press, 1980.

[8] S. M. Witt, "Automatic error detection in pronunciation training: Where we are and where we need to go," *Proc. ISADEPT*, vol. 6, 2012.

[9] ——, "Use of speech recognition in computer-assisted language learning," Ph.D. dissertation, University of Cambridge, 1999.

[10] E. Grabe, "Variation adds to prosodic typology," *Speech Prosody*, pp. 257–274, 2002.

[11] E. Grabe, G. Kochanski, and J. Coleman, "Connecting intonation labels to mathematical descriptions of fundamental frequency," *Language and speech*, vol. 50, no. 3, pp. 281–310, 2007.

[12] J. D. O'Connor and G. F. Arnold, *Intonation of colloquial English*. Longman Ltd, 2004.

[13] P. Warren, "Issues in the study of intonation in language varieties," *Language and speech*, vol. 48, no. 4, pp. 345–358, 2005.

[14] M. E. Beckman and J. B. Pierrehumbert, "Intonational structure in Japanese and English," *Phonology*, vol. 3, pp. 255–309, 1986.

[15] M. d. M. Vanrell, I. Mascaró, F. Torres-Tamarit, and P. Prieto, "Intonation as an encoder of speaker certainty: Information and confirmation yes-no questions in Catalan," *Language and speech*, vol. 56, no. 2, pp. 163–190, 2013.

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

[17] H. J. Nibert, "Phonetic and phonological evidence for intermediate phrasing in Spanish intonation," Ph.D. dissertation, University of Illinois at Urbana-Champaign, 2000.

[18] L. E. De Ruiter, "Polynomial modeling of child and adult intonation in German spontaneous speech," *Language and speech*, vol. 54, no. 2, pp. 199–223, 2011.

[19] D. Ke and B. Xu, "Chinese intonation assessment using SEV features," *ICASSP*, pp. 4853–4856, 2009.

[20] J. P. Arias, N. B. Yoma, and H. Vivanco, "Automatic intonation assessment for computer aided language learning," *Speech communication*, vol. 52, no. 3, pp. 254–267, 2010.

[21] K. Li, X. Wu, and H. Meng, "Intonation classification for L2 English speech using multi-distribution deep neural networks," *Computer Speech & Language*, vol. 43, pp. 18–33, 2016.

[22] C. Yarra and P. K. Ghosh, "Automatic intonation classification using temporal patterns in utterance-level pitch contour and perceptually motivated pitch transformation," *The Journal of the Acoustical Society of America*, vol. 144, no. 5, pp. EL471–EL476, 2018.

[23] ——, "An automatic classification of intonation using temporal structure in utterance-level pitch patterns for british english speech," *Accepted in IEEE India Council International Conference (INDICON)*, 2018.

[24] S. Jain, "Nanonets : How to use deep learning when you have limited data," *Available at https://medium.com/nanonets/nanonets-how-to-use-deep-learning-when-you-have-limited-data-f68c0b512cab*, last accessed on 21-03-2017.

[25] F. M. S. Rahul Dey, "Gate-variants of gated recurrent unit (GRU) neural networks," *arxiv*, 2017.

[26] B. Schuller, S. Steidl, A. Batliner, J. Hirschberg, J. K. Burgoon, A. Baird, A. Elkins, Y. Zhang, E. Coutinho, and K. Evanini, "The interspeech 2016 computational paralinguistics challenge: Deception, Sincerity & Native language," *Interspeech, San Francisco, USA*, pp. 2001–2005, 2016.

[27] Q. V. L. Andrew M. Dai, "Semi-supervised sequence learning," *Advances in neural information processing systems*, 2015.

[28] R. Karim, "Counting number of parameters in deep learning models by hand," *Available at https://towardsdatascience.com/counting-no-of-parameters-in-deep-learning-models-by-hand-8f1716241889*.

[29] A. Camacho and J. G. Harris, "A sawtooth waveform inspired pitch estimator for speech and music," *The Journal of the Acoustical Society of America*, vol. 124, no. 3, pp. 1638–1652, 2008.

[30] J. Bergstra, F. Bastien, O. Breuleux, P. Lamblin, R. Pascanu, O. Delalleau, G. Desjardins, D. Warde-Farley, I. Goodfellow, A. Bergeron *et al.*, "Theano: Deep learning on GPUs with python," *NIPS 2011, BigLearning Workshop, Granada, Spain*, vol. 3, 2011.

[31] F. Chollet, "Keras: Deep learning library for Tensorflow and Theano," *Available at https://github.com/fchollet/keras*, last accessed on 14-03-2017.