An Automatic Classification of Intonation Using Temporal Structure in Utterance-level Pitch Patterns for British English Speech

Chiranjeevi Yarra
Electrical Engineering
Indian Institute of Science
Bangalore 560012, India
chiranjeeviy@iisc.ac.in

Prasanta Kumar Ghosh
Electrical Engineering
Indian Institute of Science
Bangalore 560012, India
prasantg@iisc.ac.in

***Abstract*—In spoken communication, intonation often conveys meaning of an utterance. Thus, incorrect intonation, typically made by second language (L2) learners, could result in miscommunication. We, in this work, consider the problem of automatically detecting the intonation of British English (BE) utterances which could be useful for providing feedback to the L2 learners. Typically, in BE, the meaning is conveyed through four intonation classes – Glide-up, Glide-down, Dive and Take-off. We hypothesize that these classes could be discriminated using temporal structure in utterance-level pitch patterns. These patterns could be represented by either stylized pitch or tones from automatic tone and break indices (AuToBI) tool. We model these temporal structures for the intonation classification using three techniques, namely, n-gram, deep neural network and long short term memory recurrent networks. Experiments are conducted on the speech data collected from a spoken English training material for teaching intonation of BE. We obtain better unweighted average recall (UAR) with the proposed schemes compared to the baseline scheme, that does not exploit temporal structure in the utterance-level pitch patterns. Among different proposed schemes, the highest absolute improvement in the UAR is found to be 9.33% over the baseline scheme.**

***Index Terms*—intonation classification, pitch stylization, tones, computer assisted language learning.**

I. Introduction

Intonation often adds meaning to words and word groups [1], [2]. It also adds the speaker's feelings, particularly in British English (BE) [1]. Hence, incorrect intonation would result in miscommunication. Thus, in the second language (L2) training, for learning BE, L2 learners require to learn BE intonation for obtaining a good proficiency in spoken communication. Further, the L2 learners would be benefited with an automated system that detects the learners' proficiency in the intonation [3][4]. Such systems would also be useful in the applications like computer assisted language learning (CALL) [3]. These systems could be designed by using reliable native speakers' intonation models. For example, in detecting L2 learners' phoneme proficiency, previous studies have used phoneme models built from the native speakers' speech data [5]. In a similar manner, in this work, we propose models to identify/classify BE intonation which could be useful for detecting L2 learners' proficiency automatically.

In BE intonation training, L2 learners are required to learn four different ways of producing intonation to match the competence of a native BE speaker [1]. These are Glide-up, Glide-down, Dive and Take-off [1] and we refer them as intonation classes. These classes can be identified by modeling variations in the pitch, since pitch is the acoustic manifestation of the intonation [6]. Figure 1 shows the native BE speakers' pitch contours from these four intonation classes, referred to as raw pitch. From the figure, it is observed that in Glide-down the pitch falls from a high to low value at the end of the utterance (shown in red rectangular box). In contrast to this, in the remaining three classes, pitch rises from a low to high value. Typically, with the pitch fall, the Glide-down could be discriminated from the remaining classes [1]. Within the three remaining classes, as observed from the figure, the Take-off could be discriminated from Glide-up and Dive with consistently low pitch values before the pitch rise. Finally, Dive and Glide-up could be discriminated with a level difference in the pitch at the end of the utterance as well as with the temporal structures in the pitch before the pitch rise [1]. From the figure, it is observed that in the case of Glide-up, the low pitch value before the pitch rise is reached with a gradual change from a high pitch value. However, in the Dive, this value is reached with a fast change from a high pitch and maintained consistently around the low value until the pitch rise. Hence, the intonation classes could be discriminated based on the properties of pitch rise, fall and level (RFAL) as well as from their associated temporal structures.

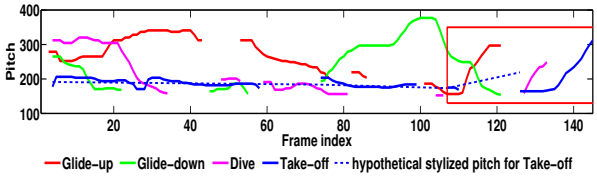

Fig. 1. Pitch contours of four utterances belonging to four intonation classes – Glide-up, Glide-down, Dive and Take-off along with a hypothetical stylized pitch contour. The hypothetical stylized pitch is obtained by joining the average pitch values within each voiced segments.

Typically, in teaching these four intonation categories, L2 learners' are trained by representing the entire raw pitch contour into small discrete symbols within syllable segments called tones [1][6], instead of teaching them by showing a

native speakers' raw pitch. Generally, tones are proposed to indicate RFAL changes in the raw pitch. For BE, these tones are obtained automatically from the speech signal using the automatic tone and break indices (AuToBI) tool [7]. Li et al. have shown the effectiveness of these tones for intonation classification [8]. However, they have considered manually annotated tones. Also, they have not considered the temporal structures in utterance-level tone patterns. In contrast to the tones reflecting RFAL changes, pitch stylization techniques convert the raw pitch into a contour composed of line segments called stylized pitch [9]. Typically, these techniques retain perceptually relevant information in the raw pitch [10], which could be indicative of the RFAL changes. Figure 1 shows a hypothetical stylization of a raw pitch belonging to Take-off. From the figure, it is observed that the stylized pitch has less variations than the raw pitch. Hence, we assume that the stylized pitch could represent the RFAL changes better compared to the raw pitch. Thus, considering temporal structures in the stylized pitch could provide better discrimination among all the intonation classes. Wypych has introduced models based on thr stylized pitch for Polish language intonation recognition [11]. However, no experiments on the intonation classification accuracy is reported. To the best of our knowledge, there have been no studies with the stylized pitch for the BE intonation classification task.

Arias et al. [3] have assessed the L2 learners intonation by comparing their raw pitch with the experts' raw pitch contour. Ke et al. have used the same strategy to assess Chinese learners, considering tone and duration based features from the pitch contours [4]. However, in both the studies, temporal structures in the raw pitch have not been exploited. Unlike the previous studies, we use temporal structures in the utterance-level pitch patterns (tones or stylized pitch) for the BE intonation classification task. We hypothesize that these temporal structures of the pitch patterns would provide better class discrimination than using the temporal structures of the raw pitch, which is evaluated using two baselines.

In this work, we represent the raw pitch of an intonation class using two pitch patterns – tones and stylized pitch. For each class, we model the temporal structures in these patterns separately using three modeling techniques – n-gram, deep neural networks (DNN) and long short term memory (LSTM) recurrent networks, for which, we compute features specific to each model using each of these patterns. We use posterior probabilities or likelihoods obtained from these models to classify the BE intonation. Experiments are performed on the speech data collected from a spoken English training material for teaching BE intonation [1]. We consider three baseline schemes to validate our experiments – work proposed by Li et al. [8] for overall comparison and hidden Markov model (HMM) with raw pitch & DNN with raw pitch to know the effectiveness of the pitch patterns compared to the raw pitch. Among the different proposed schemes, the highest improvement in the unweighted average racall (UAR) [12] is found to be 9.33%, 8.90% and 7.80% (absolute) respectively on three baseline schemes.

II. Database and experimental setup

In this work, we use the speech data from a spoken English training material [1] used for teaching BE. We consider entire speech recording containing all the utterances belonging to intonation lessons for our experiments. We manually segment the entire speech recording into individual speech files belonging to every utterance. We obtain the annotated text transcription as well as the intonation class labels belonging to each utterance from the same training material [1]. In the speech data, the total number of utterances is 232 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 speakers. To the best of our knowledge, there is no speech data larger than this that has these four intonation class labels annotated from the experts.

III. Proposed approach

The block diagram in Figure 2 describes the two major stages involved in the proposed approach. The first stage extracts two different sets of pitch patterns from the speech signal, namely, stylized pitch and tone labels. The second stage computes features ($\boldsymbol{f}$) from each of these sets of pitch patterns and estimates the class conditional probabilities, $\boldsymbol{p(f|C)}$, or class posterior probabilities, $\boldsymbol{p(C|f)}$, using three modeling techniques separately – n-gram, DNN and LSTM, where $\boldsymbol{C}$ denotes the class label. We train these models using the parameters optimized on the development data. We compare the class conditional or posterior probabilities from each model with each feature set belonging to a test utterance and consider the class with the highest probability as the estimated intonation class ($\hat{C}$).

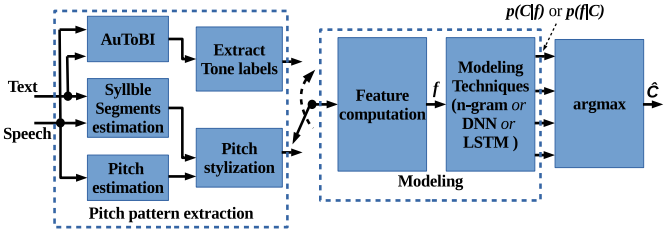

Fig. 2. Block diagram summarizing the steps involved in the proposed approach.

A. Pitch pattern extraction

We obtain tone labels and stylized pitch from speech signal separately. The tone labels are extracted using AuToBI, which takes the speech signal and its text as the inputs. A total of 8 different tone labels is obtained from AuToBI on the entire speech data considered in this work. These tone labels are !H*, H*, H+!H*, L*, L*+H, L+!H*, L+H* and NOACCENT. The stylized pitch is estimated using the pitch stylization technique proposed by Ghosh et al. [13]. This technique provides stylization of the raw pitch using a set of line segments for each voiced region separately, taking the number of line segments per voiced region as the input. We consider the number of

line segments to be equal to the number of syllables in a voiced region, since the teaching of intonation is done using the representations within the syllable segments [1][6][9]. We estimate the number of syllable segments from speech signal and its text in two steps. Firstly, phone segments along with their transcriptions are obtained using force alignment process and then syllable transcriptions are obtained by automatically syllabifying the phone transcriptions. We obtain the raw pitch using the sub-harmonic to harmonic ratio (SHR) algorithm proposed by Sun et al. [14].

B. Modeling

We propose different sets of features depending on the modeling technique as well as the pitch pattern to represent the temporal structures in that pitch pattern. Hence, we discuss the feature computation for each pitch pattern along with the following modeling techniques.

3.2.1. n-gram: In n-gram modeling, we create a finite set of symbols to represent the pitch patterns. In the case of tone labels, we consider tone symbols obtained from AuToBI directly for modeling. However, for stylized pitch, we propose a method to obtain symbols for each line segment. In general, the slope and intercept values of the line segments are real. Hence, it results in an infinite set of symbols when the absolute values of these line parameters are considered directly as symbols. We empirically find a finite set of symbols from these line segments in three steps.

1) We obtain a normalized stylize pitch, $x_n(p) = \frac{x(p)-min(x(p))}{max(x(p))-min(x(p)))}$, where $x(p)$ is stylized pitch with the frame index p.
   We use $m(k)$ and $c(k)$, $1 \leq k \leq K$, to indicate the slope and the intercept values of the lines belonging to $x_n(p)$ in each segment, where K is the total number of segments in an utterance.
   With this normalization, it is easy to see that the values of $|m(k)|$ and $c(k)$ are in between 0 and 1.
2) We obtain quantized slopes, $m_q(k) = round(m(k)*Q)$, for each line segment k, and indicate those segments with the symbol $S_{m_q(k)}$, where Q is the number of quantization levels.
3) We re-symbolize the $S_{m_q(k)}; \forall k$ such that $m_q(k) = 0$ as $L_{c_q(k)}$, where $c_q(k) = round(Q * c(k))$.

After these steps, at each k-th line segment, we obtain the symbol either $S_{m_q(k)}$ or $L_{c_q(k)}$. We empirically set the Q to be 4, for which the symbols are S_1, S_2, S_3, S_4, S_{-1},S_{-2}, S_{-3}, S_{-4}, L_0, L_1, L_2, L_3 and L_4. However, in DNN and LSTM modeling, we use $m(k)$ and $c(k)$ directly as a feature vector instead of their quantized symbols ($S_{m_q(k)}$ or $L_{c_q(k)}$) representing as a binary feature vector. With this, we allow the models to learn these representations from the training data.

3.2.2. DNN: In order to model DNN, it is required to obtain feature vectors with uniform length across all the utterances. However, we observe that the sequence length of tone labels and the number of stylized pitch line segments are not uniform across all utterances. Among all the utterances considered in this work, the maximum sequence length and the maximum number of segments are found to be 7 and 15 respectively. By considering these maximum values (7 and 15) separately for tone labels and stylized pitch, we obtain a 56-dim and a 30-dim feature vector across all utterances, respectively, in three steps. In the case of tone labels, we follow a procedure for each utterance as below.

1) We represent a tone label sequence of an utterance as an 8-dim binary vector sequence by replacing each label with an 8-dim binary vector, since the number of distinct tone labels is 8 in the considered speech data.
2) We append 8-dim zero vectors before the 8-dim binary vector sequence to make that sequence length as 7.
3) We convert these seven length 8-dim vectors to a 56-dim feature vector by stacking 8-dim vectors one below another.

Similarly, for the stylized pitch, we follow the above three steps to obtain a 30-dim feature vector. However, the 8-dim binary vector is replaced with a 2-dim vector containing $m(k)$ and $c(k)$ values of the line segments and the sequence length is made to 15 in the place of 7.

We experiment with the 56-dim and 30-dim vectors separately. For the training, we represent the class label as a 4-dim binary vector and train DNN with the number of output units equal to 4 and input units equal to 56 or 30. We use the softmax activation function for all 4 output units. We consider DNN with 3-hidden layers and we learn the number of hidden units per layer ($\in \{128, 256, 512, 1024, 2048\}$) in the training stage. For each hidden layer, we consider the relu activation function for all units. We use categorical cross-entropy as the objective function for DNN training.

3.2.3. LSTM: LSTM has been shown to be effective in capturing pattern in temporal sequences [15]. Hence, we use LSTM to model the temporal patterns using the tone labels and the stylized pitch for the intonation classification task. For LSTM modeling, we use the sequences of 8-dim binary vectors and 2-dim vectors respectively for tone labels and stylized pitch. These sequences are obtained following the steps outlined in Section 3.2.2. We use these sequence of vectors directly to train and test LSTM. Similar to DNN, we represent the class label as a 4-dim binary vector and train LSTM with four number of output layer units with the softmax activation functions. The number of units in the input layer is 8 and 2 respectively for tone labels and stylized pitch. In LSTM, we use 2-hidden layers with 128 hidden units using the relu activation function for each hidden unit. Similar to DNN, categorical cross-entropy is used as the objective function for training LSTM. We consider the number of loops as 50 to train LSTM.

IV. Experimental results

A. Experimental setup

We consider UAR as the performance measure to evaluate the classification performance [16]. We conduct the experiments in a 10-fold cross validation setup where eight folds are used for training, one fold for development and one fold

for testing in a round robin fashion. n-gram model is built using the SRILM language model toolkit [17]. DNN and LSTM are implemented using Theano [18] and Keras [19]. We obtain force-aligned phone boundaries using Kaldi speech recognition tool kit [20] with Fisher-English [21] acoustic models. Phone transcriptions are converted into syllables using P2TK automated syllabifier [22]. For comparison, the work by Li et al. [8] using DNN is considered as a baseline, referred to as Li-baseline. Li et al. have used tone labels belonging to the final pitch accent and the following edge tone as the features for the DNN. However, in our implementation, we consider only final pitch accent tone label, since we do not obtain any edge tone labels with AuToBI.

Further, we analyze the usefulness of the pitch patterns in comparison with raw pitch for the intonation classification task. For this, we build classifiers using HMM with HTK toolkit [23] as well as using DNN, referred to as HMM-baseline and DNN-baseline respectively. In HMM-baseline, we train HMM for each class using a low pass filtered raw pitch as a feature. On the development data, the filter cut-off frequency and number of states for HMM are learnt. With these parameters, likelihoods are calculated using HMMs for all classes for a test utterance and the class with the highest likelihood is considered as the estimated label for that utterance. In DNN-baseline, we use a setup similar to that described in Section 3.2.2. However in this setup, we consider the low-pass filtered raw pitch with a fixed number of resampled values (N) as a feature. We learn the filter cut-off frequency and N on the development data. With these parameters the class labels are estimated for the test utterance.

TABLE I
AVERAGE UARS AND STANDARD DEVIATIONS, IN BRACKETS, OBTAINED WITH THE PROPOSED SCHEMES WITH STYLIZED PITCH AS WELL AS TONE LABELS AND HMM & DNN WITH RAW PITCH ON THE DEVELOPMENT (IN BLUE COLOR) AND THE TEST SETS.

	Raw pitch		Stylized pitch	Tone labels
HMM	41.1 (8.5) 26.0 (7.9)	n-gram	37.6 (7.2) 27.0 (11.0)	35.1 (9.7) **33.7** (9.4)
DNN	31.4 (4.9) 27.1 (5.2)	DNN	37.5 (13.7) 29.0 (10.4)	34.3 (9.1) 29.0 (11.7)
		LSTM	29.7 (8.1) **34.9** (9.0)	15.6 (7.6) 19.7 (10.4)

B. Results and discussion

Table I shows mean and standard deviations (SDs) of UARs across all the ten folds on the development and test sets for HMM-baseline & DNN-baseline as well as for all the proposed schemes with stylized pitch and tone labels. The average (SD) UAR using Li-baseline is 25.72% (5.36) and 25.58% (7.00) on the development and the test sets respectively. From the table, it is observed that the UARs obtained with HMM-baseline and DNN-baseline are found to be higher than that using Li-baseline by an UAR of 15% and 6% & 0.5% and 1.5% on development & test sets respectively. This large drop in UAR from the development to the test set indicates that although the HMM captures temporal structures in the raw pitch to maximize UAR on the development set, it does not generalize well in the test set. Similarly, the improvements in DNN and HMM performance over Li-baseline suggest that for the intonation classification task, it would be beneficial to consider the pitch pattern in the entire utterance compared to that within the tone labels at the end of the utterance.

From the table, it is observed that, when the stylized pitch is used, the highest (indicated by bold in Table I) and least UARs are found with LSTM and n-gram models respectively among all the proposed schemes. This indicates that n-gram model with quantized stylized pitch parameters fails to capture the temporal structure in the pitch patterns. Instead, original stylized pitch patterns would be more beneficial for the classification task as done in LSTM model. In contrast, when the tone labels are used, the highest (indicated by bold in Table I) and least UARs are obtained with n-gram and LSTM respectively. This could be because of poor training of LSTM model with few data points and 8-dim tone label feature unlike that with 2-dim stylized pitch feature. On the other hand, n-gram model works with a symbol-set of size eight when the tone labels are used compared to 13 when stylized pitch is used. With limited training data, this smaller symbol-set could lead to a higher UAR with tone labels compared to that with the stylized pitch. Lowering Q may result in smaller symbol-set in the case of stylized pitch. But that may not represent the temporal patterns well. Thus, a better quantization strategy with smaller sized symbol-set could also improve the UAR with stylized pitch.

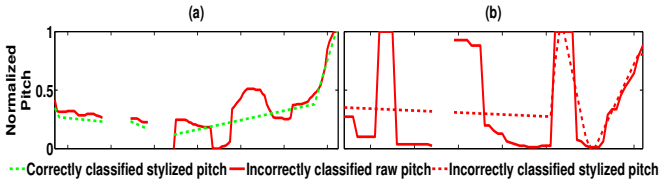

Fig. 3. Illustrations of Take-off stylized pitch and raw pitch that are classified (a) correctly by LSTM but incorrectly by HMM, (b) incorrectly by both LSTM and HMM.

From the table, it is also observed that all the proposed schemes perform better than HMM-baseline as well as DNN-baseline on the test data. This indicates that the intonation would be classified better with discrete pitch patterns like tone labels and stylized pitch instead of the raw pitch. This could be because the detail variations in the raw pitch may result in a noisy representation leading to a poor training of the classifier. As an example, in Figure 3a, we show a normalized raw pitch and a normalized stylized pitch of an utterance belonging to Take-off class. Where the utterance is misclassified by HMM with raw pitch while it is correctly classified by LSTM with stylized pitch. Similarly, it is observed that all the proposed schemes perform better than Li-baseline except using LSTM with tone labels. The highest improvement is found to be 9.33% over Li-baseline, when LSTM network is used with stylized pitch. It is also interesting to observe that the n-gram with tone labels results in an improvement of 8.1% over Li-baseline. These together support the hypothesis that the

temporal structures in the utterance-level pitch patterns would be better for the intonation classification task. Further, we study the effectiveness of these utterance-level pitch patterns on the classification of each intonation class using a confusion matrix computed on the test set.

TABLE II
CONFUSION MATRIX FOR LI-BASELINE, N-GRAM WITH TONE LABELS AND LSTM WITH STYLIZED PITCH ON THE TEST SET. THE ROWS REPRESENT THE GROUND-TRUTH CLASSES – GLIDE-UP (#1), GLIDE-DOWN (#2). DIVE (#3) AND TAKE-OFF (#4) AND THE COLUMNS REPRESENT THE ESTIMATED CLASSES. EACH CELL ENTRY IS THE AVERAGE PERCENTAGE ACROSS ALL THE TEN FOLDS.

	Baseline				n-gram with tone labels				LSTM with stylized pitch			
	#1	#2	#3	#4	#1	#2	#3	#4	#1	#2	#3	#4
#1	17.0	23.3	59.7	0.0	25.8	31.4	32.9	10.0	26.4	29.4	22.9	21.3
#2	3.3	30.5	66.1	0.0	15.1	44.0	21.5	19.3	11.5	48.9	32.1	7.6
#3	13.0	32.2	54.8	0.0	25.2	29.6	22.7	22.5	13.3	25.0	56.0	5.8
#4	3.3	19.3	77.3	0.0	34.0	14.7	9.0	42.3	40.7	12.0	39.0	8.3

From the confusion matrices shown in Table II, it is observed that, with Li-baseline, the Take-off and Glide-up are classified as Dive in most of the cases. It is also interesting to notice that not a single utterance is classified as Take-off. This could be because Take-off has the same rising tone pattern as Dive and Glide-up at the end of utterances. Hence, considering the tone pattern only at the end of the utterance could not result in better discrimination among these classes. In contrast, the confusions among these classes are less in n-gram model. This could be because n-gram model uses the temporal structure over the entire sentence, hence, reduces the confusion among these classes. However from the table, it is observed that the accuracy (diagonal entry) of Dive class reduces in n-gram model compared to Li-baseline. But when LSTM is used, the accuracies are improved in all classes compared to Li-baseline; the overall UAR (from Table I) is also improved compared to Li-baseline and n-gram models. However, when the stylized pitch is used, the Take-off accuracy using LSTM is lower compared to n-gram model. It could be that the stylization process fails to produce the patterns that can distinguish Take-off from Dive and Glide-up. Hence, it suggests that an improved pitch stylization could reduce the confusion among the classes as well as improve the overall UAR in the classification task.

From Table I, it is observed that the proposed schemes result in an improvement in the UAR for the classification task. However, the highest UAR is only $\sim$35%, which has further scope to improve. The UAR from the proposed schemes are dependent on the reliability of the models as well as the pitch patterns – stylized pitch and tone labels. Further, these depend on the pitch estimation algorithms that are used to obtain stylized pitch or those used in AuToBI. Figure 3b shows a normalized stylized pitch and a normalized raw pitch of an utterance belonging to Take-off class, which is misclassified by both HMM with raw pitch and LSTM with stylized pitch. From the figure, it is observed that the raw pitch, hence, the normalized raw pitch has sudden changes in the pitch values from a low to high and a high to low probably due to pitch halving and doubling errors [24], [25]. This results in unwanted variations in the stylized pitch that lead to incorrect classification. On the other hand, intonation also depends on the linguistic measures like syllable stress [1][6], which can be considered in the classification task to improve the UAR further.

V. CONCLUSIONS

We model the temporal structures in the utterance-level pitch patterns for the BE intonation classification task using n-gram, DNN and LSTM. The pitch patterns are estimated from AuToBI tone labels and pitch stylization techniques. We also model the temporal structures in the raw pitch contour using HMM and DNN to know the effectiveness of the raw pitch. Experiments with the spoken English training material with four intonation classes reveal that the proposed schemes improve UAR compared to the baseline scheme as well as the models with raw pitch. Further investigations are required to develop better pitch patterns that could result in an improved UAR under typical halving and doubling errors in the pitch estimation. Future works also include the use of linguistic features in addition to the pitch patterns for improving the classification performance.

REFERENCES

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

[2] 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.

[3] 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.

[4] D. Ke and B. Xu, "Chinese intonation assessment using SEV features," *IEEE International Conference on Acoustics, Speech, and Signal Processing*, pp. 4853–4856, 2009.

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

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

[7] A. Rosenberg, "AuToBI-a tool for automatic ToBI annotation." *Proc. of INTERSPEECH, Makuhari, Japan*, pp. 146–149, 2010.

[8] 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.

[9] S. Ravuri and D. P. Ellis, "Stylization of pitch with syllable-based linear segments," *IEEE International Conference on Acoustics, Speech, and Signal Processing*, pp. 3985–3988, 2008.

[10] P. Mertens *et al.*, "Automatic pitch contour stylization using a model of tonal perception," *Computer Speech & Language*, vol. 9, no. 3, pp. 257–288, 1995.

[11] M. Wypych, "An automatic intonation recognizer for the polish language based on machine learning and expert knowledge." *Proc. of INTERSPEECH*, pp. 3305–3308, 2005.

[12] 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," *Proc. of INTERSPEECH, San Francisco, USA*, pp. 2001–2005, 2016.

[13] P. K. Ghosh and S. S. Narayanan, "Pitch contour stylization using an optimal piecewise polynomial approximation," *IEEE signal processing letters*, vol. 16, no. 9, pp. 810–813, 2009.

[14] X. Sun, "Pitch determination and voice quality analysis using subharmonic-to-harmonic ratio," *IEEE International Conference on Acoustics, Speech, and Signal Processing*, vol. 1, pp. 333–336, 2002.

[15] H. Sak, A. W. Senior, and F. Beaufays, "Long short-term memory recurrent neural network architectures for large scale acoustic modeling." *Proc. of INTERSPEECH*, pp. 338–342, 2014.
[16] A. Rosenberg, "Classifying skewed data: Importance weighting to optimize average recall," *Proc. of INTERSPEECH*, pp. 2242–2245, 2012.
[17] A. Stolcke *et al.*, "SRILM-an extensible language modeling toolkit." *Proc. of INTERSPEECH*, vol. 2002, p. 2002, 2002.
[18] 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.
[19] F. Chollet, "Keras: Deep learning library for Tensorflow and Theano," *Available at https://github.com/fchollet/keras*, last accessed on 14-03-2017.
[20] D. Povey, A. Ghoshal, G. Boulianne, L. Burget, O. Glembek, N. Goel, M. Hannemann, P. Motlicek, Y. Qian, P. Schwarz *et al.*, "The kaldi speech recognition toolkit," *IEEE workshop on automatic speech recognition and understanding (ASRU)*, 2011.
[21] C. Cieri, D. Miller, and K. Walker, "The Fisher Corpus: a resource for the next generations of speech-to-text." *4th international conference on Language Resources Evaluation*, vol. 4, pp. 69–71, 2004.
[22] J. Tauberer, "P2TK automated syllabifier," *Available at https://sourceforge.net/p/p2tk/code/HEAD/tree/python/syllabify/*, last accessed on 14-03-2017.
[23] S. Young, G. Evermann, M. Gales, T. Hain, D. Kershaw, X. Liu, G. Moore, J. Odell, D. Ollason, D. Povey *et al.*, "The HTK book," *Cambridge university engineering department*, vol. 3, p. 175, 2002.
[24] M. Asgari and I. Shafran, "Improving the accuracy and the robustness of harmonic model for pitch estimation." *Proc. of INTERSPEECH*, pp. 1936–1940, 2013.
[25] S. Gonzalez and M. Brookes, "PEFAC-a pitch estimation algorithm robust to high levels of noise," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 22, no. 2, pp. 518–530, 2014.