Automatic intonation classification using temporal patterns in utterance-level pitch contour and perceptually motivated pitch transformation

Chiranjeevi Yarra, and Prasanta Kumar Ghosh

Citation: The Journal of the Acoustical Society of America **144**, EL471 (2018); doi: 10.1121/1.5080466
View online: https://doi.org/10.1121/1.5080466
View Table of Contents: http://asa.scitation.org/toc/jas/144/5
Published by the Acoustical Society of America

Automatic intonation classification using temporal patterns in utterance-level pitch contour and perceptually motivated pitch transformation

Chiranjeevi Yarra[a] **and Prasanta Kumar Ghosh**
Department of Electrical Engineering, Indian Institute of Science, Karnataka 560012, India
chiranjeeviy@iisc.ac.in, prasantg@iisc.ac.in

Abstract: Second language learners of British English (BE) are typically trained for four intonation classes: Glide-up, Glide-down, Dive, and Take-off. Automatic four-way intonation classification could be useful to evaluate a learner's pronunciation. However, such automatic classification is challenging without having manually annotated tones, typically considered in intonation analysis and classification tasks. In this, a three-dimensional feature sequence is proposed representing temporal patterns in the utterance-level *f*0 contour using a perceptually motivated pitch transformation. Hidden Markov model based classification experiments conducted using a training material for teaching BE intonation demonstrate the benefit of the proposed approach over the baseline scheme considered.

[BHS]
Date Received: September 19, 2018 **Date Accepted:** November 7, 2018

1. Introduction

Intonation often adds meaning to words and word groups.[1,2] In general, intonation contains a sequence of discrete patterns called tones.[1–3] Although the last tone in the sequence, called nuclear tone,[1,3] plays a critical role in the intonation, all tones in the sequence together convey the meaning.[1,2] Hence, an incorrect tone sequence would result in wrong intonation, thus, miscommunication. Therefore, in the second language (L2) training, for example, in learning British English (BE), L2 learners are required to learn BE intonation for a better spoken communication.

In general, the intonation of BE varies across different geographic regions.[3] However, instead of teaching the L2 learners with many varieties of BE intonation, they are typically trained to learn intonation of the received pronunciation of BE (Ref. 1) containing four different patterns—Glide-up, Glide-down, Dive, and Take-off,[1,2] referred to as intonation classes. Later, they are trained to add finer changes to those patterns.[1] In this work, models are proposed to classify those four classes automatically in the expert's BE intonation which could be useful for detecting L2 learner's proficiency similar to the work proposed by Witt,[4] where the quality of phonemes in L2 learner's utterance has been assessed using the models built from expert's data. For this, the temporal structures in the pitch over the entire sentence are considered which, in turn, could represent intonation class dependent variabilities in the tone sequence.[1,2] It has been shown that the intonation also depends on the linguistic patterns like syllable stress;[1,2] however, in this work only the pitch patterns are exploited for the intonation classification task. Most of the existing works have studied the variations of intonation across different nativities.[3] However, a few works have addressed the problem of intonation assessment of L2 learners.[5–7]

Li *et al.*[7] have classified the intonation using two manually annotated tones, namely, a tone at the final pitch accent and an edge tone instead of considering an entire tone sequence. Ke and Xu[6] have assessed the L2 learners' intonation by comparing the tone duration based features from their pitch contour with those from experts' pitch contour. However, the manual annotation is not possible to do on a learner's utterance in an automatic assessment task. Thus, the tone labels need to be estimated. However, the tools available for estimating the tones are very limited and also errors in the tone estimation could cause degradation in the classification accuracy. In contrast to using tones, Arias *et al.*[5] have assessed the L2 learners using original pitch contours from learners and experts. However, it cannot be extended to the case of

[a]Author to whom correspondence should be addressed.

spontaneous speech where the learners' utterances are different from those of experts. Further, all the existing studies on the intonation assessment do not exploit the intonation dependent temporal structures in the pitch contour or tone sequence of the entire utterance.

In contrast, in this work an automatic BE intonation classification is performed which has three key aspects: (1) temporal structures in the utterance-level pitch contour are considered and modeled using hidden Markov models (HMMs), (2) a transformed pitch contour is used instead of the original pitch, for which frequency transformations (FTs) are considered that approximate the perceptual properties of the ear, and (3) a three-dimensional (3D) feature sequence is proposed using the transformed pitch. Experiments are performed on the speech data collected from a spoken English training material for teaching BE intonation.[2] Among all FTs in the proposed scheme, the highest absolute improvement in the unweighted average recall (UAR) over the baseline scheme is found to be 30.53%. The highest UAR is also found to be 13.56% more than the UARs obtained, respectively, without considering the FT indicating the benefit of the proposed FT.

2. Database

In this work, the speech data are considered from a spoken English training material[2] used for teaching BE. The entire speech recording is considered that contains all the utterances of intonation phrases belonging to intonation lessons for our experiments. 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 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, limit the size and the availability of such corpora.

3. Proposed approach

The block diagram in Fig. 1 describes the three major stages involved in the proposed approach. The first stage extracts pitch $p(t)$; $1 \leq t \leq T$ in the tth frame of the speech signal and a confidence score in estimating the pitch,[8] where T is the total number of frames. The second stage computes a feature sequence, $f(t)$, using $p(t)$ and the confidence scores in four steps. In the first step, the original pitch contour is transformed using one of the three FTs, namely, equivalent rectangular bandwidth (ERB), bark, and Mel to obtain transformed pitch $\phi(t)$.

In the second step, the range of the confidence score is normalized for each utterance separately such that it spans between 0 and 1, referred to as normalized score sequence $s(t)$. In the third step, utterance specific mean and range normalization of $\phi(t)$ is performed using $s(t)$ to obtain frequency transformed normalized pitch contour $\phi_n(t)$. In the fourth step, a 3D feature sequence $f(t)$ is obtained by concatenating the $\phi_n(t)$, $s(t)$, and first order differentiation of $\phi_n(t)$. However, in the unvoiced regions, $\phi_n(t)$ values are interpolated from $\phi_n(t)$ in the voiced regions. Since the interpolated values are not obtained directly from a pitch estimation technique, the $s(t)$ values are set to zero in those regions. The third stage estimates class conditional probabilities $[p(f|C)]$ using HMM. The HMMs are trained for each intonation class C and the parameters are optimized on the development data. The class conditional probabilities are compared from each model with the 3D feature from a test utterance and the class with the highest probability is considered as the estimated intonation class $\hat{C}$.

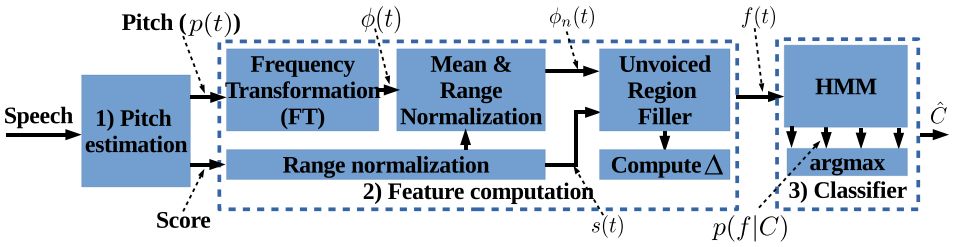

Fig. 1. (Color online) Block diagram summarizing the stages involved in the proposed approach.

3.1 Frequency transformed normalized pitch $\phi_n(t)$

Typically, the intonation classes are dependent on the dynamic information in the pitch contour and their associated temporal structures. Pitch frequency and its dynamics are, in general, speaker specific. Hence, the original pitch without the speaker specific normalization could have a large variation within every intonation class, which is not desirable for high classification accuracy. Thus, across all the classes, it is ensured that $\phi_n(t)$ has a similar range using the following utterance specific mean and range normalization:

$$\phi_n(t) = \frac{\tilde{\phi}(t)}{\max_t\left(\tilde{\phi}(t)\right) - \min_t\left(\tilde{\phi}(t)\right)}, \tag{1}$$

where $\tilde{\phi}(t) = \phi(t) - \mu_{\phi(t)}$, $\phi(t)$ is the frequency transformed pitch, and $\mu_{\phi(t)} = [\sum_{t=1}^{T} s(t)\phi(t)/\sum_{t=1}^{T} s(t)]$. One of the obvious choices for $\phi(t)$ can be the original pitch $p(t)$ without any FT, which is equivalent to having transformation of the pitch uniformly across all the pitch frequencies. However, in general, the perception of the pitch is not uniform across the pitch frequencies due to non-linearities in the auditory system. Considering this, in the literature several FTs have been proposed to capture the non-linearities, for example, ERB, bark, and Mel. The respective transformations are ERB: $\phi(t) = 21.4 \times \log_{10}[1 + 0.00437p(t)]$, bark: $\phi(t) = 26.81p(t)/1960 + p(t)$, Mel: $\phi(t) = 2595 \times \log_{10}[1 + 0.0014p(t)]$. It is observed that the features obtained from the frequency transformed pitch are less dependent on the factors that are not discriminative across intonation classes.

3.2 Feature computation

The dynamics in the intonation patterns are represented using a 3D feature vector sequence $f(t)$ and their temporal dependencies are modeled using HMM in each utterance. In order to obtain the feature sequence, the first dimension is considered as the values of $\phi_n(t)$, referred to as the static feature. The second dimension represents the values obtained from the first order difference of the static features [at t and $(t-1)$th frames] for each tth frame, referred to as the Δ feature. It has been hypothesized that augmenting Δ features to the static features would provide better temporal dependent cues for time series modeling, for example, with HMMs. It is observed that adding a higher order difference does not improve the performance and, hence, they are not considered. The third dimension represents the score indicating the confidence in estimating the pitch contour. Due to normalization, the score sequence $s(t)$ lies between 0 and 1.

The benefit of $s(t)$ as a feature. Typically, the pitch estimation techniques are not robust in estimating the pitch in all parts of the sentence. In general, they are prone to halving and doubling errors, which, in turn, could cause errors in the classification task. However, these techniques provide a score, referred to as the confidence score, whose values are proportional to the accuracy of the estimated pitch. Thus, using the confidence score as a feature could provide a benefit in the classification task. This is illustrated with the help of Fig. 2 using two exemplary sentences belonging to the intonation class "Glide-down." Figures 2(a) and 2(b) show two pitch contours and their associated $s(t)$ profiles. In the examples there is no unvoiced segments except at the beginning and at the end of the utterances, where both the values of pitch and $s(t)$ are zero. From the inset in Fig. 2(a), it is observed that the pitch rises after the fall (a high to low value) at the 45th frame index, which is against the Glide-down property. On the other hand, in Fig. 2(b), the pitch contour matches well with the Glide-down

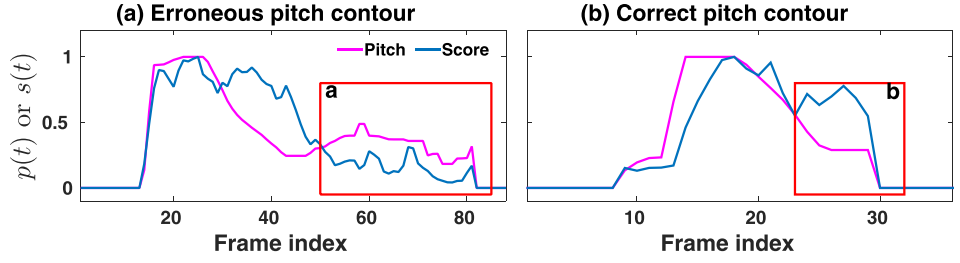

Fig. 2. (Color online) Illustrative examples of Glide-down class describing the benefit of $s(t)$ for intonation classification task. $p(t)$ shows the original pitch contour with the range normalization.

property. From both the figures, it is interesting to observe that the confidence scores are lower in the inset of Fig. 2(a) (where erroneous pitch rise is present) compared to those in the inset of Fig. 2(b). It is also observed that the confidence scores in the inset of Fig. 2(b) are above 0.5. The observations from both the figures indicate that the confidence scores could add complementary properties to the pitch based features for intonation classification task.

3.3 Interpolation at unvoiced regions

In general, pitch confines to the voiced regions only; hence, the estimated pitch contour in an unvoiced region does not contain any meaningful information related to the intonation. However, Xu and Prom-On[9] have hypothesized that the dynamics involved in producing pitch in a voiced region are unchanged in the following unvoiced region. Based on the dynamics, they also suggested that the pitch contour of a voiced region can be extended into the following unvoiced region. However, the suggested computation is complex and involves many parameters, which are required to be learned from a reasonably large amount of data. In order to obtain values in the unvoiced regions, in this work, due to the limited data size, pitch values are interpolated in those regions considering the estimated pitch values in all voiced regions of an utterance using three interpolations, namely, linear, cubic, and spline.

3.4 HMM based classifier

HMM is one of the time series models which can model the temporal dependencies in feature sequences effectively. In the recent past, long short-term memory recurrent networks have been used for the time series modeling, however, these require a large amount of data for training. In this work, due to limitation on the data size, a HMM is considered to learn the temporal structures for each intonation class separately using the 3D feature vector sequences. In general, after the training, each class specific HMM results in a higher likelihood under matched conditions compared to the HMMs of mismatched intonation classes. With this hypothesis, for a test sentence the intonation class is estimated using two steps. First, class conditional probabilities are computed using each class specific HMM, following which the class with the highest class conditional probability is considered as the estimated intonation class. The number of states in each class specific HMM is learned during the training on development data by varying the number states from 3 to 5. Each HMM state is considered with four component Gaussian mixture model and a diagonal covariance for each mixture component.

3.5 Experimental setup

UAR, which is the average of each class recall,[10] is considered the performance measure to evaluate the classification accuracy. Experiments are conducted in a tenfold cross validation setup where eightfolds are used for training, onefold for development, and onefold for testing in a round robin fashion. The SWIPE algorithm is used to estimate pitch and obtain confidence scores.[8] Class specific HMMs are implemented using the HTK toolkit.[11] For comparison, the work proposed by Li *et al.*[7] is implemented and considered as a baseline. Li *et al.* have used manually annotated tone labels belonging to the final pitch accent and the following edge tones as the features. In this work, the pitch accents and tones are estimated using automatic tone and break indices (AuToBI) tool,[10] which requires syllable transcriptions. Syllable transcriptions are obtained by performing automated syllabification[12] on the phone transcriptions obtained from force alignment employing the Kaldi speech recognition tool kit.[13]

3.6 Results and discussion

The mean and standard deviations (SDs) of UARs are reported that are computed across all the tenfolds on the test set. First, the performance of the proposed method is discussed with $\phi_n(t)$ and $\phi(t)$ under the three interpolations. Next, considering the $\phi_n(t)$ with the interpolation that provides the best UAR, the performance of the proposed method is analyzed with the baseline scheme. The mean UARs averaged across all three FTs are found to be 60.06% and 55.29%, 55.34% and 54.90%, and 55.44% and 52.25% with $\phi_n(t)$ and $\phi(t)$, respectively, under three interpolations. From these values, it is observed that the average UAR values are higher when the mean and range normalization is considered under all three interpolations. This indicates that normalization of the transformed pitch improves the classification performance. It is also observed that the average UAR values with and without normalization are the highest under linear interpolation among all three interpolations. This indicates that the variabilities in the higher order interpolations cause unwanted variations, and hence result

Table 1. Average UARs and SDs, in brackets, obtained with the proposed method with four different combinations of features and all three FTs under linear interpolation on the test set.

Features	Original	ERB	bark	Mel
$\phi_n(t)$	42.91 (11.53)	46.44 (8.15)	43.58 (11.39)	44.69 (13.05)
$\{\phi_n(t), s(t)\}$	54.08 (11.53)	57.53 (11.74)	59.20 (12.82)	59.97 (12.11)
$\{\phi_n(t), \Delta\phi_n(t)\}$	43.99 (14.39)	44.92 (14.32)	49.95 (12.30)	45.73 (14.03)
$\{\phi_n(t), s(t), \Delta\phi_n(t)\}$	55.46 (8.39)	58.23 (7.39)	60.18 (5.75)	**61.77** (8.66)

in lower UARs. Further, the performance of the proposed method is analyzed using $\phi_n(t)$ with linear interpolation.

Table 1 shows the mean (SD) of UARs on the test sets for all three FTs and four feature combinations: (1) $\phi_n(t)$; (2) $\{\phi_n(t), s(t)\}$; (3) $\{\phi_n(t), \Delta\phi_n(t)\}$, and (4) proposed 3D feature set $\{\phi_n(t), \Delta\phi_n(t), s(t)\}$ using the linear interpolation. The mean (SD) of UARs using baseline is 31.24% (5.96%) across all tenfolds on the test sets. From Table 1 it is observed that the UARs obtained using the baseline are found to be lower than that using the proposed method under all combinations of the FTs and feature combinations. The improvements in the performance of the proposed method over the baseline suggest that for the intonation classification task, it would be beneficial to consider the temporal structures in the transformed pitch contour across the entire utterance compared to the tones at the end of the utterance. The lower UAR in the baseline scheme could also be due to the tone estimation errors; however, currently there is no readily available automatic tool which provides more accurate tone estimation than the AuToBI. It is also observed that the averaged mean UARs are the lowest in all four feature combinations when no FT is used. This indicates that the FT is helpful for the intonation classification.

From Table 1 it is interesting to observe that the classification accuracies in the second and fourth rows are higher than those in the first and third rows, respectively. This indicates that the performance of the proposed approach is improved by considering the confidence score in the feature sequence. Similarly, higher classification accuracies in the third and fourth rows compared to those in the respective first and second rows show the benefit of the Δ feature for the intonation classification task. In the proposed method the highest accuracy (indicated in bold) is found with the 3D feature sequence computed using the Mel FT. This could be because the Mel FT has been derived from the perceptual experiments on the pitch frequencies unlike bark and ERB FTs, which are based on the properties of critical bands in hearing.

4. Conclusion

Utterance-level temporal structures are modeled for the BE intonation classification task using HMM, for which a 3D feature sequence is proposed from the pitch contour. The pitch contour is obtained by transforming the original pitch using ERB, bark, and Mel scales. 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 utterance-level temporal patterns, FT, and the proposed 3D feature. Further investigations are required to develop a better feature sequence 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 3D feature sequence for improving the classification performance.

References and links

[1]J. C. Wells, *English Intonation: An Introduction* (Cambridge University Press, Cambridge, UK, 2006).
[2]J. D. O'Connor, *Better English Pronunciation* (Cambridge University Press, Cambridge, UK, 1980).
[3]A. Cruttenden, "Intonational diglossia: A case study of Glasgow," J. Int. Phonetic Assoc. **37**(3), 257–274 (2007).
[4]S. M. Witt, "Use of speech recognition in computer-assisted language learning," Ph.D. thesis, University of Cambridge, 1999.
[5]J. P. Arias, N. B. Yoma, and H. Vivanco, "Automatic intonation assessment for computer aided language learning," Speech Commun. **52**(3), 254–267 (2010).
[6]D. Ke and B. Xu, "Chinese intonation assessment using SEV features," in *IEEE International Conference on Acoustics, Speech and Signal Processing* (2009), pp. 4853–4856.
[7]K. Li, X. Wu, and H. Meng, "Intonation classification for L2 English speech using multi-distribution deep neural networks," Comp. Speech Lang. **43**, 18–33 (2016).
[8]A. Camacho and J. G. Harris, "A sawtooth waveform inspired pitch estimator for speech and music," J. Acoust. Soc. Am. **124**(3), 1638–1652 (2008).

[9]Y. Xu and S. Prom-On, "Toward invariant functional representations of variable surface fundamental frequency contours: Synthesizing speech melody via model-based stochastic learning," Speech Commun. **57**, 181–208 (2014).
[10]A. Rosenberg, "AuToBI-a tool for automatic ToBI annotation," in *Interspeech* (Makuhari, Japan, 2010), pp. 146–149.
[11]S. Young, G. Evermann, M. Gales, T. Hain, D. Kershaw, X. Liu, G. Moore, J. Odell, D. Ollason, D. Povey, V. Valtchev, and P. Woodland, *The HTK Book* (Cambridge University Press, Cambridge, UK, 2002).
[12]J. Tauberer, "P2TK automated syllabifier," Available at https://sourceforge.net/p/p2tk/code/HEAD/tree/python/syllabify/ (Last viewed March 14, 2018).
[13]D. Povey, A. Ghoshal, G. Boulianne, and L. Burget, "The Kaldi speech recognition toolkit," in *IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU)* (2011).