Better Nonnative Intonation Scores through Prosodic Theory

Joseph Tepperman and Shrikanth Narayanan

Signal Analysis and Interpretation Laboratory
University of Southern California, Los Angeles, USA
http://sail.usc.edu/
tepperma@usc.edu, shri@sipi.usc.edu

Abstract

Pronunciation scoring is one important task for software designed to give feedback to students practicing a second language. English intonation can convey information about a speaker's nativeness, so previous studies have proposed using intonation-based models to score nonnative pronunciation. One past approach trained models for a set of pronunciation scores using ad hoc features derived from the frequency contour. We use prosodic theory to train models for categorical intonation units, inspired by work in modeling tone languages. These HMM-based models offer 0.398 correlation between automatic and listener scores on the ISLE nonnative speech corpus, compared to the 0.156 baseline correlation.

Index Terms: prosody, nonnative speech, pronunciation

1. Introduction

Why might we expect intonation - the patterns of pitch in speech - to inform automatic assessment of an English learner's pronunciation quality? Though English is not a tone language like Mandarin (i.e. one in which intonation can determine the meaning of an isolated word), intonational variation in English conveys a wide variety of information. The placement and choice of pitch accents and boundary tones - manifested through the shape and range of the fundamental frequency (f0) contour - is well-correlated with speaker intentions and listener perception on the phrase level. Boundary tones within or at the ends of phrases delineate syntactic units of various sizes, offering the listener hints for appropriate processing and interpretation [15]. Pitch accents within a phrase can intimate if the information offered is new, contrastive, accessible, or uncertain [16]. Even a speaker's emotional state can be inferred from phrase-level suprasegmental features [1], and listeners can discern a speaker's regional accent from hearing the intonation alone in filtered speech [5]. It follows that the extent to which an English learner sounds native must be due partly to their intonation, and that intonation modeling is potentially useful for automatic score generation in a second-language practice environment.

Perhaps the most complete study in using text-independent intonation-based features to generate pronunciation scores was reported by Teixeira et al in [12]. The proposed solution was to derive many text-free features from the phrase-level f0 contour, and then train a decision tree to assign each feature vector an integer score on a 1 to 5 scale. Relying only on these pronunciation scores as training class labels, this method used no prosodic annotation, which requires expert linguistic knowledge and suffers from low inter-annotator agreement. It also allowed a certain versatility in the selection of features and the size of the feature set, though the linguistic relevance of many of the features is not clear since they are chosen ad hoc. Ultimately, features derived from alignment of the text and from other sources resulted in automatic scores better correlated with listener scores than the text-free intonation-based features did.

The modest performance in scoring pronunciation based on intonation in [12] is perhaps due both to the ad hoc choice of features and the fact that the models trained represent uncertain perceptual scores rather than relevant linguistic units of intonation (i.e. pitch accents and boundary tones). Our previous work in [13] introduced improved intonation-based scoring by training Hidden Markov Models for categorical intonation units on continuous f0 and energy contours from native speech. These units could then be decoded from nonnative speech in the same way that words commonly are, allowing us to estimate scores for how well the nonnative features fit the native models. In this paper we expand on that work by investigating several new methods for improving HMM-based intonation models for the pronunciation evaluation task, and by reproducing the baseline decision tree scoring procedure from [12] for comparison on our data set. Our hypothesis (and main finding) is that significant improvements in automatic score correlation with listener perception can come from the use of linguistic theory about intonation and prosodic structure, in the form of proper f0 processing, meaningful feature sets, theoretical grammars for recognition of intonation events, and accounting for contextual effects. Intonation is only one factor contributing to a listener's qualitative assessment of phrase-level pronunciation, so we expect improvements in this domain to result in correlation still well below that of agreement among listeners for scoring pronunciation in general.

2. Corpora and Annotation

The nonnative speech in this study comes from the ISLE corpus of English learners [8]. It consists of read sentences by native speakers of Italian and German, at various levels of British English proficiency. This corpus was divided into training and test sets, each with an equal number of German and Italian speakers per set, and no speakers in common between them. Information about their relative sizes is given in Table 1. The sentences in these sets were scored by one native listener for overall pronunciation (taking into account intonation and all other cues) on a 1 to 5 scale, as in [12]. To measure inter-annotator agreement, 138 sentences from this corpus were scored by five other native listeners. Average inter-labeler correlation was 0.657, which can be considered an upper bound on all automatic scores' performance. The one native listener who scored all sentences had a correlation of 0.732 with the medians of the other five listeners' scores; since this exceeds the inter-labeler agreement, we

[1] This work was supported in part by the NSF and the NIH.

	native train set	*nonnative train set*	*nonnative test set*
corpora	BURNC, IViE	ISLE	ISLE
total speakers	7	8	8
total sentences	3657	1238	307
total minutes	80.2	82.8	25.5

Table 1: Relative sizes of the training and test sets.

can consider that listener's scores to be a reliable reference.

For training native models of intonation events, we used the Boston University Radio News Corpus [10], and the IViE corpus [2]. The former consists of read news by American English speakers, transcribed for intonation using the ToBI system; the latter uses a similar transcription convention for its read southern British English. Previous work in [13] has shown that, due to broad similarities in tone realization (if not placement), both dialects can be used together for training categorical prosodic models for nonnative pronunciation evaluation. Because of low transcriber agreement for some of the finer tone categories, and to reconcile minor differences in the two transcription conventions, all sub-categories within both accents and boundaries were collapsed into only two: high and low (i.e. high and low pitch accents, high and low phrase boundaries, etc.). Intonational “silence” was inserted into the transcripts at the start and end of each phrase, and at every ToBI break of 3 or higher.

3. Baseline: Decision Tree Score Models

For comparison with our proposed improvements in Section 4, we reproduced the method in [12] of pronunciation score modeling with decision trees trained on text-independent features derived from the f0 contour. The f0 contour of each sentence in the nonnative training and test sets was estimated using the standard autocorrelation method with a frame size of 10 msec. The curves were then smoothed as described in [11] using a 5-point median filter (i.e. each frame's value was replaced by the median of itself and that of the four frames immediately surrounding it) to minimize any phonetic segment effects, and then log-scaled. Then a piecewise linear stylization was fitted to each voiced segment. Using this processed representation of the f0 contour, for each sentence we derived the same 23 “pitch signal” features used in [12], mainly related to f0 range and slope. Note that, though all these features are text-independent (i.e. they can be derived from any spoken phrase without accounting for differences in lexical items or sentence structure), they are not entirely text-free. The rate of speaking (ROS) used for normalizing some of the features is derived from alignment of the text, though unsupervised methods of estimating ROS exist and could potentially be used here.

A decision tree performs classification by asking of a feature vector a sequence of questions which can be thought of as “branches” of the tree. The results of the final branches in the question sequence (called the “leaf” nodes of the tree) represent subsets of the training data, and we can use these subsets to obtain estimates of the probability of each class for every leaf. As it relates to pronunciation scoring, the posterior probability of the human score h (integers from 1 to 5) given the feature vector $\underline{f} = \{f_1, f_2, \dots, f_{22}\}$ can be estimated as $P(h|\underline{f}) \cong P(h|l_{\underline{f}})$ where $l_{\underline{f}}$ is the leaf node that the questions asked of $\underline{f}$ lead to. In [12], these posteriors are used to estimate pronunciation scores as continuous values based on the minimum error criterion,

$$E[h|l_{\underline{f}}] = \sum_{i=1}^{5} h_i \cdot P(h_i|l_{\underline{f}}) \qquad (1)$$

The study in [12] evaluated these automatic scores in terms of correlation with listener scores and the mean absolute error between the automatic and listener scores (as a percentage of the maximum possible error, which is 4). The optimal tree is grown as in [9] by specifying the minimum size of a leaf subset that maximizes correlation on the test set, and then for that minimum subset size choosing a pruning confidence factor that again maximizes correlation. Baseline results for this optimal tree are given in Table 2 (row 2), alongside results for random scores generated using the same proportions of 1 to 5 scores found in the nonnative train set and averaged over ten random realizations. These best decision tree results in Table 2, though better than random, are lower than those obtained on the corpus used in [12] (they reported 0.247 correlation, 25.4% error), but their listener agreement (0.8 score correlation) was higher than for the test set used here.

4. HMM Intonation Models

This section explains our intonation modeling and score generation paradigm, as a contrast to the purely score-based models of the baseline method in Section 3. We then describe some potential improvements inspired by prosodic theory and intonation modeling in Mandarin, and present experimental results.

4.1. Score Generation with HMMs

We trained native HMMs for eight different intonation units: high and low pitch accents ($\mathbf{H}^*$ and $\mathbf{L}^*$), high and low intermediate phrase boundaries ($\mathbf{H}-$ and $\mathbf{L}-$), high and low phrase-final boundaries ($\mathbf{H}\%$ and $\mathbf{L}\%$), a high initial boundary tone ($\%\mathbf{H}$), and a silence model (**SIL**). Without needing to know the text spoken, we can decode these units from a nonnative speaker's suprasegmental features and estimate scores that represent how native the speaker's intonation sounds. Our approach is simply to estimate posterior probabilities of each decoded tonal unit and then take their product over the phrase to estimate a phrase-level posterior score. This is similar to the method of scoring with phoneme models used in [9]. See [13] for our models' accuracy in tone label recognition.

Assuming a bigram model of intonation, each decoded unit's posterior is calculated as

$$P(M_t|O_t, M_{t-1}) = \frac{P(O_t|M_t)P(M_t|M_{t-1})}{\sum_n P(O_t|M_n)P(M_n|M_{n-1})} \qquad (2)$$

where O_t is the speech observation in suprasegmental features at time t, M_t is the recognized unit, $P(M_t|M_{t-1})$ is its bigram probability given the previous segment, and n takes values over all HMMs. In the case of an unweighted tone network (as we will see in certain experiments), $P(M_t|O_t, M_{t-1})$ reduces to $P(M_t|O_t)$, and $P(M_t|M_{t-1})$ reduces to its prior, $P(M_t)$. Then we approximate an overall utterance score ρ as the product of the posteriors of the T decoded intonation segments:

$$\rho = P(M_1, M_2, \dots, M_T|O_t) \approx \prod_{t=1}^{T} P(M_t|O_t, M_{t-1}) \qquad (3)$$

To make our scores comparable with those of the baseline method, we normalized all log-posteriors for the test set so that their mean and variance matched that of the 1 to 5 scores in the nonnative train set. Then all scores below 1 were set equal to 1, and all scores above 5 were set equal to 5. It is possible that other score calibration heuristics might give better results, but this one was simple and required no further training.

All context-independent (CI) HMMs were trained with a flat-start initialization and five iterations of embedded re-estimation. We arbitrarily chose three hidden states and 16 Gaussian mixtures per state, though these parameters could be optimized on a development set. A bigram tone recognition model and three features - f0, plus its first and second derivatives (estimated with the standard regression formulae) - were used for all initial experiments (rows 3-6 in Table 2).

4.2. Proposed Improvements

4.2.1. Mandarin-style f0 Preprocessing

Studies in automatic recognition of Mandarin speech have made good use of f0-related features, because of the importance of tone in lexical disambiguation. There is currently no standard method of preprocessing or normalizing estimates of the f0 contour (or at least nothing akin to the MFCCs of spectral features), but studies of Mandarin have developed a number of techniques aimed to compensate for variations in raw f0 [4]. Some of them are used in the baseline approach's preprocessing step (e.g. median smoothing to remove discontinuities caused by segment-level production). When working with HMM models trained on continuous f0 contours, it helps to interpolate regions of unvoiced speech - areas which would in theory have had voiced intonation if the phonemes of the utterance had been different. Here we use simple linear interpolation. Next, instead of only log-scaling the frequency contour, we convert it to the ERB scale, to better match the human ear's perception of frequency. Gradual declination in f0 over the course of a phrase is common in many languages, including both English and Mandarin. We normalized for declination by subtracting from each f0 frame the mean within a 1.5 second window. Finally, we performed 7-point median filtering to smooth out discontinuities.

Row 3 in Table 2 presents results for CI intonation HMMs using the baseline preprocessing explained in Section 3; the correlation results are comparable to those obtained using the decision tree scoring method. Correlation improves dramatically with the benefit of techniques first used for Mandarin (row 4). This indicates that the baseline method, row 2, is not ideal for HMM intonation models.

4.2.2. Additional Features

So far, the only features used have been the f0 contour and its first and second derivatives. This was simply for comparison with the baseline method, which used only features derived from f0. However, some studies (such as [6]) have emphasized the importance of energy in the perception of pitch accents. Experiments in automatic recognition of tones in Mandarin have found improvement through using MFCCs along with the f0-related features [14]. Could the addition of these features improve intonation modeling in English? Rows 5 and 6 in Table 2 report the results. Some modest improvement in correlation is seen with the use of RMS energy (and its first and second derivatives) but performance declines dramatically once MFCCs are introduced. This suggests that intonation units in English are not realized through contrasts in spectral characteristics so much as through suprasegmental features like f0 and energy, but feature weighting could offer some improvement.

	modeling method	*corr.*	*error*
(1)	random scores	-0.002	38.5
(2)	baseline decision tree scores	**0.156**	**29.1**
(3)	CI HMMs: preprocessing from (2)	0.153	34.1
(4)	(3) w/ Mandarin preprocessing	0.304	31.2
(5)	(4) + E and Δ E and $\Delta\Delta$ E	0.320	29.4
(6)	(5) + 12 MFCCs (+ Δ and $\Delta\Delta$)	0.054	35.7
(7)	(5) w/ tone loop grammar	0.223	32.2
(8)	(5) w/ theory-based FSG	0.318	29.9
(9)	(8) w/ CD HMMs	**0.398**	**27.4**

Table 2: Correlation and error of scores derived from different proposed modeling methods.

4.2.3. Intonation Grammars

In our previous work in [13], we proposed two types of intonation grammars for proper recognition of tone units before scores could be calculated. The better of these two was a bigram model: based on the transcripts, it estimated the probability $P(M_t|M_{t-1})$ where M_t is the current tone unit and M_{t-1} is the unit immediately preceding it. The other, poorer-performing finite-state grammar (FSG), was an unweighted network that decoded intonation models based on theories of prosodic structure in an English phrase [17]. At that time, the poor performance of this theory-based grammar seemed to indicate that FSGs for native intonation could not apply to nonnative speech. However, our acoustic models then did not include intermediate phrase boundaries ($\mathbf{H}-$ and $\mathbf{L}-$) as they do now. The new theory-based FSG dictated by the current choice of models requires **SIL** at the beginning and end of each utterance, allows an optional initial high boundary ($\%\mathbf{H}$) after the initial **SIL**, and then decodes zero or more intermediate phrase sequences:

$$< \mathbf{H}^*|\mathbf{L}^* > \ (\mathbf{H}-|\mathbf{L}-)\ [\mathbf{SIL}]$$

followed by a required phrase-final sequence:

$$< \mathbf{H}^*|\mathbf{L}^* > \ (\mathbf{H}-|\mathbf{L}-)\ (\mathbf{H}\%|\mathbf{L}\%)$$

where square brackets denote optional elements, angle brackets denote one or more repetitions, and vertical lines mean "or". For the sake of comparison, we also tried a simple "tone loop" grammar that allowed for any sequence of tone events, making use of no statistical or theoretical information. The two non-bigram models assumed unweighted arcs for all decoding paths.

Results for these grammars are compared in Table 2. Both the bigram model (row 5) and the new theory-based FSG (row 8) performed better than the simple "tone loop" grammar (row 7), showing that decoding can benefit from knowledge of linguistic theory or corpus statistics. Since there was no statistically significant difference in correlation performance between the bigram and the theory-based FSG, all subsequent experiments were conducted with the latter since it requires no training and is more in accord with prosodic theory than the bigram model - i.e. it conceives of intonation within a phrase as a whole sequence of tones consistent with the structure of prosodic information, rather than each tone dependent only on the one immediately preceding it.

4.2.4. Context-Dependent Tone Models

Again making use of ideas first proposed for Mandarin, we trained new HMMs that were dependent on their context. Often for Mandarin, tone models are trained based on the syllable over which they occur [14]. This is appropriate for Mandarin, since most words are one syllable in length and a single tone is realized over the rhyme of each syllable. For English, however, the domain of the pitch accent is the prosodic foot [3], consisting of one accented syllable and all following unaccented syllables until the next accented syllable is reached. So, instead of conditioning the models on syllable- or phoneme-level context, we conditioned them on their left and right intonation contexts. For example, a low accent surrounded by two high accents ($\mathbf{H}^* - \mathbf{L}^* + \mathbf{H}^*$) would be trained as a different acoustic model than a low accent surrounded by low accents ($\mathbf{L}^* - \mathbf{L}^* + \mathbf{L}^*$). This compensates for tone sandhi [7] - the change in tone realization based on its prosodic context, often important phonologically in tone languages like Mandarin.

To train these context-dependent tone models we tied similar HMM states together using the decision tree-based model clustering method normally used for training CD phoneme models. The "questions" for the decision tree to test in determining a model's cluster were related to groupings of tone contexts (e.g. right tone is high, right tone is a boundary, etc.). All contexts unseen in the training data were then synthesized using the tree that maximized the likelihood of the native speech given the models. As with the context-independent HMMs, these models had 3 hidden states and 16 mixtures per state. Decoding tone events was still done using the context-independent models and the theory-based finite-state intonation grammar explained in Section 4.2.3; this avoided training an overly complex bigram model for context-dependent tones. The context-dependent models were then used in calculating the posterior scores by assuming that the decoded context was correct and limiting $\sum_n P(O|M_n)P(M_n)$ in Eqn 2 to only those tone models that share the same context. For example, if a decoded tone unit had $\mathbf{L}^*$ decoded on its left and $\mathbf{H}^*$ decoded on its right, then M_n could only take the form of $\mathbf{L}^* - \mathbf{H}^* + \mathbf{H}^*$ or $\mathbf{L}^* - \mathbf{L}\% + \mathbf{H}^*$ or $\mathbf{L}^* - \mathbf{SIL} + \mathbf{H}^*$, etc.

Results for the context-dependent (CD) models are shown in Table 2, row 9. They are the best models presented in this work, and outperform the baseline decision tree method by 16.2% correlation and 1.7% error. For pronunciation scoring, correlation is a more relevant metric than mean absolute error, since giving meaningless flat scores near the test set's mean for all items could lead to very low error but also poor correlation. Using the standard statistical test, the difference in correlations between row 9 and row 2 is significant with $p \leq 0.001$.

5. Conclusion

With the baseline method from [12], the optimal decision tree for scoring English learner pronunciation based on intonation-based features achieved only 0.156 correlation with listener scores on this test set. Using HMMs representing intonation units, plus linguistic theory about prosodic structure in English and methods already developed for robust Mandarin modeling, we were able to raise that correlation to 0.398. Our method had the disadvantage of requiring specialized intonation transcripts rather than simple phrase-level pronunciation scores, but needs no prior knowledge of the target text and could potentially be used to assess intonation in spontaneous speech. As expected, the correlation of our best intonation-based scores still falls well below the inter-listener agreement in deciding pronunciation scores based on intonation and all other available cues. Future work is needed to combine these intonation-based assessments with those of pronunciation cues on other time-scales.

6. References

[1] M. Bulut, S. Lee, and S. Narayanan, "Analysis of emtional speech prosody in terms of part of speech tags," in *Proc. of Interspeech ICSLP*, Antwerp, Belgium, August 2007.

[2] E. Grabe, "Intonational variation in urban dialects of English spoken in the British Isles," in *Regional Variation in Intonation*, P. Gilles and J. Peters, eds. Linguistische Arbeiten, Tuebingen, Niemeyer, pp. 9-31.

[3] D. Hirst, "Intonation in British English," in *Intonation Systems: A Survey of Twenty Languages*. Ed. D. Hirst and A. Di Cristo. Cambridge: CUP, 1998.

[4] M.-Y. Hwang, X. Lei, W. Wang, and T. Shinozaki, "Investigation on Mandarin Broadcast News Speech Recognition," in *Proc. of Interspeech ICSLP*, Pittsburgh, September 2006.

[5] A. Ikeno and J. H. L. Hansen, "The role of prosody in the perception of US native English accents," in *Proc. of Interspeech ICSLP*, Pittsburgh, PA, 2006.

[6] G. Kochanski, E. Grabe, J. Coleman, and B. Rosner, "Loudness predicts prominence: fundamental frequency lends little," *JASA*, 118(2):1038-1054, Aug. 2005.

[7] P. Ladefoged, *A Course in Phonetics*. 5th Edition. Boston: Thomson, 2006.

[8] W. Menzel, E. Atwell, P. Bonaventura, D. Herron, P. Howarth, R. Morton, and C. Souter, "The ISLE corpus of non-native spoken English," *Proc. of LREC*, Athens, 2000.

[9] L. Neumeyer, H. Franco, V. Digalakis, and M. Weintraub, "Automatic Scoring of Pronunciation Quality," *Speech Communication*, 30(2-3):83-94,1999.

[10] M. Ostendorf, P. J. Price, and S. Shattuck-Hufnagel, "The Boston University Radio News Corpus," Boston University Technical Report No. ECS-95-001, March 1995.

[11] K. Sonmez, E. Shriberg, L. Heck, and M. Weintraub, "Modeling dynamic prosodic variation for speaker verification," in *Proc. ICSLP*, 1998.

[12] C. Teixeira, H. Franco, E. Shriberg, K. Precoda, and K. Sonmez, "Prosodic Features for Automatic Text-Independent Evaluation of Degree of Nativeness for Language Learners," in *Proc. ICSLP*, 2000.

[13] J. Tepperman, A. Kazemzadeh, and S. Narayanan, "A Text-free Approach to Assessing Nonnative Intonation," in *Proc. of InterSpeech ICSLP*, Antwerp, August 2007.

[14] Y. Tian, J. Zhou, M. Chu, and E. Chang, "Tone Recognition with Fractionized Models and Outlined Features," in *Proc. of ICASSP*, Montreal, 2004.

[15] J. Vaissiere, "Perception of Intonation," in *The Handbook of Speech Perception*. Ed. D. B. Pisoni and R. E. Remez. Oxford: Blackwell, 2005.

[16] A. Wennerstrom, *The Music of Everyday Speech*. New York: OUP, 2001.

[17] H. Wright and P. A. Taylor, "Modelling Intonational Structure Using Hidden Markov Models," *ESCA Workshop on Intonation: Theory, Models, and Applications*, 1997.