# A Text-free Approach to Assessing Nonnative Intonation

*Joseph Tepperman, Abe Kazemzadeh, and Shrikanth Narayanan*

Signal Analysis and Interpretation Laboratory, University of Southern California
tepperma@usc.edu, kazemzad@usc.edu, shri@sipi.usc.edu

## Abstract

To compensate for the variability in native English intonation and the unpredictability of nonnative speech, we propose a new method of assessing nonnative intonation without any prior knowledge of the target text or phonetics. After recognition of tone events with HMMs and a bigram model of intonation, we define an utterance's automatic intonation score as the mean of the posterior probabilities for all recognized tone segments. On the ISLE corpus of learners' English, we find intonation scores generated by this technique have a 0.331 correlation with general pronunciation scores determined by native listeners. In comparison, the SRI Eduspeak system's proposal for pronunciation scoring based on suprasegmental features derived from prior knowledge of the target text yields a 0.247 correlation with listener scores on a similar corpus. Because it is text-free, our approach could be used to assess intonation outside of a strictly educational application.

**Index Terms**: nonnative speech, pronunciation evaluation, prosody, intonation, text-independent

## 1. Introduction

The role of intonation in English has long been studied, both for its importance as a discourse tool and its contribution to a nonnative speaker's overall pronunciation quality. Several studies have characterized the telltale intonation of a native English speaker in terms of specific intonational events such as pitch accents and boundary tones [1]. For example, within the conventions of the ToBI system, one should expect native speakers not to put pitch accents on function words, to use the L+H* accent to denote a contrast, and to have low-rising mid-utterance boundaries.

However, to define the canonical form of an English speaker's intonation - the set of all allowable intonational variants of a phrase - is another matter. Given text as input, speech synthesis systems can generate a natural-sounding pitch contour for a target utterance, but in the absence of a true intonational transcription it is still only one of many allowable "tunes" which can be set to the text. Even for a simple sentence such as the iconic example "I didn't steal your blue hat," the choice of content words to receive pitch accents may depend largely on the speaker's intent and the context of the utterance. "**I** didn't steal your blue hat" implies that someone else did. "I didn't steal your **blue** hat" implies that the stolen hat was another color. Both are perfectly allowable intonational interpretations, and each one dramatically alters the received meaning of the sentence. As linguist Dwight Bolinger put it, the locations of pitch accents are predictable only "if you're a mind-reader" [2]. Though there exists some notion of acceptable forms, this lack of certainty in constructing an intonational reference for native speech adds to the already difficult problem of automatically assessing intonation produced by nonnative speakers, since the influence of their native language's intonation lends another degree of ambiguity to our expectations of what should be considered unacceptable.

Despite the lack of a definitive canonical form, there are several methods which language learning systems have used to automatically evaluate a student's intonation. One is to directly compare their utterances to a pre-recorded expert [3]. Though useful in certain language-learning domains, the limitations of this technique are in its failure to capture all the variability inherent in intonation, the constraint on vocabulary size, and the amount of expert input required. Another approach is to classify the student's utterance based on features extracted from salient syllables, relying on a forced alignment of the target text [4]. Though perhaps text-independent, this method does require prior knowledge of the target, and the majority of features used are selected and normalized in an ad-hoc manner. After all, we cannot predict a complete canonical form for English intonation given target text, except perhaps for a very limited and artificial sentence structure, so the motivation behind using such features will be necessarily vague - ultimately in [4] these yielded no improvement in correlation with listener scores when combined with segmental cues. Though it's reasonable to assume prior knowledge of the text in a language-learning scenario, dependence on it will make evaluation impossible in the case of spontaneous nonnative speech for foreign language practice, or in domains outside of language instruction.

Our goal in this study was to develop a method for automatically assessing nonnative intonation such that machine scores would correlate well with human perceptual evaluations of pronunciation quality. We argue that the variability in native intonation and the unpredictability in nonnative intonation demand that our method should at least be text-independent. Beyond that, so that it would work outside the domain of language learning applications, we intend to demonstrate a technique that requires no prior knowledge of the text but performs comparably with systems that do.

## 2. Corpora

The ISLE Corpus [5] consists of read passages by native speakers of Italian and German (23 of each) who are students of British English, with widely varying degrees of proficiency. Uniquely nonnative pronunciation phenomena are elicited by the inclusion of difficult consonant clusters, shifts in lexical stress between noun and verb forms, and minimal phone pairs. Certain passages - in particular the many sentences of the form "I said X, not Y" - help elicit some clearly nonnative tone behavior in terms of pitch accents and boundaries. The corpus, however, is not transcribed for intonation, so mere ToBI label detection or classification is not really an option as a method of automatic assessment. But even if we had some type of intonational transcription, we argue that those labels alone would not

really inform our assessment of the speaker's degree of nonnativeness, for reasons outlined in Section 1.

An appropriate reference for nonnative students of British English is, of course, native British English - the best intonational sampling of which comes in the form of the IViE corpus [6], designed to highlight intonational variants across dialects and to spearhead the development of a labeling system that captures the subtleties of these variants perhaps more astutely than ToBI. However, for continental language learners, only the more Southern dialects - those of London, Cambridge, and perhaps Leeds - are really representative of the British English taught in foreign schools. But the prosodically-transcribed reading passages from these regions (1 male and 1 female speaker, each) total only about 0.1 hours of speech, which may not be enough to train very robust intonational models. Fortunately, many linguists agree that American and Southern British English share a common intonational vocabulary. According to Bolinger [7], the differences between them are "not in the configurations. . .but in frequency and pragmatic choice." In other words, the same general shapes of pitch contours are seen in both types of English, but are used in differing contexts and possibly also differ in their grammatical structure. This legitimizes the use of American intonation, in the form of the Boston University Radio News Corpus (BURNC) [8], as training data in combination with the IViE recordings. The BURNC is composed entirely of professionally read radio speech, with one speaker completely annotated with the ToBI system, providing another 1.2 hours of intonation training.

## 3. Intonational Representation

Our chosen representation of intonational behavior in terms of low and high target accents and boundaries comes less from a bias against continuous representations - the tune or tilt approaches to intonational analysis [9] - but rather from the constraints of our training corpora and their respective transcription conventions. However, some modifications are necessary to make these transcriptions consistent and useful. The IViE transcription convention is slightly different from traditional ToBI annotation in that it both omits from and adds to ToBI's tiers of prosodic annotation. IViE's phonological tier - the one which most resembles ToBI's tone tier - collapses phrase accents and boundary tones into boundary tones alone, though a finer-grain representation is encoded in IViE's phonetic tier. For these reasons, we have mapped all phrase accents in the BURNC to their corresponding high or low pitch accents; though a phrase accent technically describes a type of boundary behavior, we wanted to keep it dinstinct from true phrase-final boundary motion.

Due to the number of variants in accent and boundary, and the relatively poor agreement among ToBI transcribers in assigning these labels to pitch events, most studies which use this representation to model intonation collapse all accents into one class, and all boundaries into another. From a classification or detection point of view, this minimizes uninformative confusions among the classes, but also loses many of the subtle variants in pitch contour shape which may help characterize a native or nonnative pronunciation. We chose to train two sets of acoustic models - one which collapsed accents and boundaries (**AB**), and one which allowed for high or low distinctions within each tone category and distinguished between initial and final boundaries (**HL**). Both sets also required intonational "Silence" models which were defined as any ToBI break of 3 or higher, and were expected immediately following every IViE boundary as well as at the beginning and end of each intonational phrase.

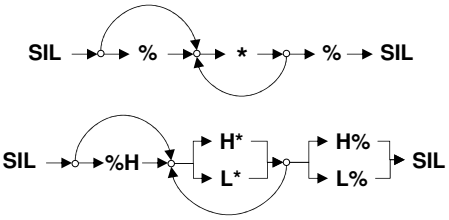


Figure 1: Our finite-state intonation recognition grammars for each acoustic model set, using a modified form of ToBI notation. Top: **AB**. Bottom: **HL**.

## 4. Feature Processing and Model Training

Though the phonetic and lexical content of a phrase can affect the intonation produced with it - particularly the discontinuities imposed by unvoiced segments or stop consonants - a pitch contour useful for training canonical native models should reflect gradual trends in contour shape and be functionally independent of the phonetics over which it is set. To do this, we adopted the method outlined in [9] for processing the pitch contour by first linearly interpolating all unvoiced segments, then smoothing the resulting curve. With the ESPS pitch estimation method and no phonetic segmentation, an unvoiced segment suitable for interpolation was defined as any section of adjacent frames with zero probability of voicing and of less than 300 msec in length. Longer unvoiced sections were assigned a pitch of 0 Hz. After interpolation, smoothing was performed with 7-point median filtering, and finally the processed pitch contour was mean-normalized.

A pitch accent is the perceptual phenomenon of word-level emphasis within a phrase and is linked to a major pitch change or iconic contour shape, though some argue that the energy (or loudness) of an accented word or syllable is a better indicator [10]. Because of its link to the change in pitch, we chose to follow [9] and include delta and acceleration estimates for the pitch contour as features. These were calculated using the standard regression formula with four neighboring frames as context for the estimates of the derivatives. The mean-normalized RMS energy contour was also included as a feature, due to its link to accent - in theory, low energy and zero pitch slope seen together would be a dead giveaway for the presence of silence.

Many studies in detection of tone events have used Hidden Markov Models (HMMs) to capture the changes in intonation encoded by the ToBI system [9]. However, extraction of the features used to train these HMMs often relies on prior knowledge of syllabic or phonetic segmentation of the target utterance. Durational features and, for example, the pitch range or slope over a syllable or syllable nucleus have been shown to be meaningful indicators of tone behavior [4], while other studies have simply used the EM algorithm over continuous pitch features (such as those stated above) to estimate model parameters [9]. We take our lead from these latter studies, on the assumption of no prior knowledge about the spoken text or phonetic alignment. Furthermore, ToBI transcriptions denote only the center of a perceived target tone, so EM training is desireable because it is not really possible to define where the perceptually relevant segment of the pitch contour begins or ends. With a flat-start initialization and Baum-Welch re-estimation, we employed a standard 5-state left-right HMM topology to capture the overall long-term shape of each tone event in our modified transcripts, similar to the way phonetic HMMs will model gradual spectral

| | *mean inter-eval. corr.* | *intonation corr.* |
|---|---|---|
| *all utterances* | 0.657 | 0.331 |
| *speaker-level* | 0.798 | 0.280 |
| *Sentence 1* | 0.640 | 0.308 |
| *Sentence 2* | 0.760 | 0.511 |
| *Sentence 3* | 0.584 | 0.181 |
| *Italian* | 0.707 | 0.233 |
| *German* | 0.238 | 0.156 |

Table 1: Inter-evaluator correlation compared with the performance of the best intonation models from Table 2, over various subsets of the test set.

shifts within a phoneme. The number of Gaussian mixtures per state we left as an experimental variable. One would expect the **AB** models to require more mixtures to describe the many different types of accents and boundaries collapsed into only two models, but then for a pronunciation evaluation task such as this it may be better not to let the models encompass too much variability that may be strictly speaker-dependent.

Though the real goal of this project was not only to classify or detect intonational events, an intermediate recognition step was necessary to estimate an intonational nativeness score for each utterance. To that end, we trained four different intonation grammars for each set of models. The first was a finite-state network allowing only recognition results of the form expected given the chosen intonational representation. These are illustrated in Figure 1 and are based on the theory behind the ToBI system as well as the grammars used in [11]. The other three grammars were backoff bigram models for intonation events, trained on just the British transcriptions, just the American transcriptions, or both. If it is true that American and British speakers of English use the same types of intonation but with different frequencies, chances are the British-trained language model will work best for assessing students who are learning specifically British pronunciation. As a check of model suitability at this point, we performed intonation recognition on a held-out set of ten files from the BURNC using **AB** models with 16 mixtures per state and the finite-state grammar in Figure 1. This achieved 71.08% tone detection accuracy (where the number of reference tone labels was 930); this figure is consistent with the results reported in [9], which used similar methodology.

## 5. Score Calculation

Our approach to calculating an overall utterance-level score for intonation is modeled after the SRI Eduspeak system's method of calculating a sentence-level pronunciation score with phone models [12]. After recognition of our intonational events as described in Section 4, for each tone or boundary segment we calculate its posterior probability, defined as

$$P(M_t|O) = \frac{P(O|M_t)P(M_t)}{\sum_i P(O|M_i)P(M_i)} \quad (1)$$

where $O$ is the speech observation in suprasegmental features, $M_t$ is the recognized model and $P(M_t)$ is its prior, and $i$ takes values over all HMMs. The motivation behind using posterior probilities instead of likelihoods is that it is a measure of how close the student's speech is to a native model, given the features - and indeed posteriors have been shown to correlate more closely with human perception of pronunciation than likelihood scores do. Then we define an overall utterance score $\rho$ as the mean of the posteriors of the $T$ intonation segments recognized in that utterance:

$$\rho = \frac{1}{T}\sum_{t=1}^{T} P(M_t|O) \quad (2)$$

## 6. Perceptual Evaluations

Our test set consisted of 138 read sentences from the ISLE corpus - the same three sentences, read by all 46 speakers. Sentence 1 was "I said white, not bait," designed to artificially induce vowel contrasts. Sentence 2 was part of a more natural restaurant dialogue: "Could I have chicken soup as a starter and then lamb chops?" Sentence 3 exemplified a standard travel agency interaction: "This summer I'd like to visit Rome for a few days." The diversity in the structure of these sentences can help validate our method's applicability to many types of utterances, without restrictions on the text or syntax.

The setup of our listener evaluations of these test recordings is modeled after that of [4], since we wanted our results to be roughly comparable (though their test set was larger, and came from native speakers of Japanese). Six native listeners were asked to evaluate the pronunciation of each nonnative test set utterance on a 1 to 5 scale, taking into account not just intonation but also fluency, articulation, rhythm, and other cues. The ordering was randomized so as to minimize the influence of a speaker's overall pronunciation on the perception of any one of that speaker's utterances. The set of reference scores for comparison with our automatic results was decided by taking the median score over all evaluators for each test sentence. For purposes of inter-evaluator comparison, we also calculated each listener's median score for every speaker as an overall rating of that speaker's pronunciation.

Inter-evaluator correlation in scoring over various subsets is reported in Table 1. We see first that the level of correlation varied depending on the sentence - so, evidently some sentences had more obvious cues to pronunciation or reading quality than others (e.g. the silent 'b' in "lamb" in Sentence 2). Secondly, speaker-level correlation was much higher than the overall utterance level scores, which indicates that the evaluators were self-consistent and used inter-speaker comparisons to inform their judgments. The most startling finding here is that the correlation in judgment of the native Italian speakers is dramatically higher than for that of the native German speakers. This is perhaps because the Italian speakers in the corpus have overall much lower English proficiency than the German speakers [5], and so their reading and pronunciation mistakes would be more obvious to some listeners; moreover, German is more closely related to English than Italian is, perhaps also causing some listeners to rate them less harshly than others. These levels of correlation seem reasonable considering the reported 0.7 maximum agreement among ISLE corpus transcribers in deciding the location of a pronunciation mistake [5].

## 7. Results and Discussion

To compare our acoustic model sets - **AB** and **HL** - and the tone grammars proposed in Section 4, we calculated the correlation of our automatic posterior-based scores with the native listener scores over all combinations of acoustic models, recognition grammars, and number of mixtures per state (16, 32, or 64). These results are reported in Table 2. The best overall cor-

| **HL** | *16* | *32* | *64* |
|---|---|---|---|
| *FSG* | -0.030 | 0.070 | -0.049 |
| *bigram: British* | 0.221 | 0.244 | 0.247 |
| *bigram: American* | 0.210 | 0.262 | 0.203 |
| *bigram: both* | 0.246 | **0.331** | 0.248 |
| **AB** | *16* | *32* | *64* |
| *FSG* | -0.156 | -0.171 | -0.203 |
| *bigram: British* | 0.026 | 0.012 | -0.045 |
| *bigram: American* | 0.034 | 0.025 | 0.022 |
| *bigram: both* | -0.010 | 0.012 | -0.027 |

Table 2: Correlation of automatic and native listener scores, over all four proposed tone grammars and both acoustic model sets (**HL** and **AB**) with 16, 32, or 64 mixtures per state.

relation, 0.331, was achieved by using the **HL** acoustic models, the bigram tone model trained on both British and American tone transcripts, and 32 mixtures per state. This correlation result exceeds the SRI Eduspeak system's performance for a similar test set - the correlation of their "continuous score" derived from all suprasegmental features (including those derived adhoc from knowledge of the phonetic segmentation) was 0.247, and these features did not yield an improvement in correlation when combined with segmental features [4].

It is apparent that the **HL** acoustic models proposed in Section 3 perform much better than our **AB** models; the **AB** set's high negative correlation is curious considering the machine scores are derived from posteriors. We can conclude that **HL**'s allowance to high and low tone models within the boundary and accent categories resulted in better characterization of native intonation shape than simply collapsing all accent and boundary tones into two respective HMMs. For the **HL** models, a bigram trained on all available transcripts gave the best results, which legitmizes the use of American intonational data for assessing nonnative speakers of British English, and suggests the finite-state grammar derived from theory does not really apply to nonnative speech. If the best performance is achieved with 32 mixtures per state, this indicates there is a critical point in the number of mixtures beyond which the acoustic models begin to account for too much variation in native intonation.

For this best combination of acoustic model and grammar, we calculated the correlation with listener scores over the same subsets used to analyze inter-evaluator agreement. These subset results are reported in Table 1. We found that the general trends agree with inter-evaluator correlation found in that table's adjacent column: Sentence 2 had better correlation than Sentence 1, which had better correlation than Sentence 3; likewise, correlation for the Italian speakers was better than for the German speakers. This suggests that our automatic scores follow the same subtle differences in perception elicited from the listeners. However, correlation of speaker-level scores was not better than those on the sentence level, probably because our algorithm had no way of comparing speakers or utterances, though that is what the native listeners must have done.

## 8. Conclusion

With this text-free approach to intonation assessment, our automatic scores achieved correlation with listener perception comparable to that of a state-of-the-art pronunciation system that does rely on textual knowledge in addition to our proposed suprasegmental cues. In the future, these intonation-based scores can be combined with ones derived from segmental and rhythmic analysis to improve correlation with general pronunciation scores. With some small changes, our method may also be applied in other domains, such as speaker identification or assessment of spontaneous speech. Possible performance improvements may come from accounting for declination when normalizing the pitch contour, using a higher-order n-gram model for intonation recognition, or training nonnative models of intonation for comparison.

## 9. References


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

[2] D. Bolinger, "Accent Is Predictable (If You're a Mind-Reader)", *Language*, 48(3):633-644, Sept. 1972.

[3] R. Delmonte, "SLIM prosodic automatic tools for self-learning instruction," *Speech Communication*, vol. 30, pp. 145-166, 2000.

[4] 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.

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

[6] 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.

[7] D. Bolinger, "Intonation in American English," in *Intonation Systems*, D. Hirst and A. Di Cristo, eds. Cambridge: CUP, 1998.

[8] 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.

[9] P. Taylor, "Analysis and Synthesis of Intonation using the Tilt Model," *JASA*, 107(3):1697-1714, 2000.

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

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

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