Intonation Patterns in English Tag Questions of Japanese Speakers of English as a Second Language

Deborah F. Burleson,
Department of Linguistics, Indiana University &
School of Medicine, IUPUI

Ten female native speakers of American English (NS) and ten Japanese females who speak English as a second language (NNS) were recorded reading a text containing emphatically-assertive tag questions. Evidence was sought to determine whether or not native speakers can correctly identify these utterances as produced by native or non-native speakers on the basis of intonation alone. The tag questions were low-pass filtered to remove segmental content, then played to ten native speaker jurors who judged them as native or non-native. Judges correctly identified NSs at 73.5% and NNSs at 71.3%. This study further investigates what differences exist between the two groups. Utterances that were unanimously correctly identified were analyzed using the ToBI transcription system. Areas of distinct differences between the two groups were subjected to acoustic examination of frequency, duration and intensity. NS utterances were more homogeneous than were NNS utterances.

1 Introduction

1.1 Current study

The study of intonation encompasses domains ranging from descriptions of speaker physiology and voice quality to paralinguistic analyses of speakers' efforts to communicate emotion or affect. The area of interest examined here centers on the conventionalized phonological components, and their phonetic realization, in the intonation used by native speakers of American English (NS) as contrasted with that of non-native speakers of American English (NNS), specifically native speakers of Japanese who speak English as a second language. "Intonation," for the purposes of this paper is, as described by Ladd (1990), "the use of suprasegmental phonetic features to convey 'postlexical' or sentence-level pragmatic meanings in a linguistically structured way." (p. 6), though paralinguistic information is also a factor in the intonational patterns examined here.

The linguistic focus of the current study is the production of English (falling) tag questions in which speakers strongly assert a firm belief, as in (1). Typical renditions of these structures are produced with a high accent and a low boundary tone complex (H* LL%) These assertive tag questions differ from tags requesting mere confirmation, as in

(2), which are typically produced without an accent and with a rising boundary tone (LH%). Appendix A contains the structures actually examined here.

(1) It IS pretty, isn't it. (Falling)
(I'm absolute positive it's pretty, and you'd better agree with me.)
(Shaded in Appendix A)

(2) It is pretty, isn't it? (Rising)
(I'm not sure it's pretty, and I want your confirmation.)

The intonation production of such tags may appear to be obvious to native speakers of English. However, the degree to which such patterns are a peculiar aspect of English, and hence present difficulties to learners of English is unclear. The current paper examines specifically the falling tag, which is probably the more obvious of the two intonation patterns, and poses the following questions concerning items which are clearly perceived as different from native productions:

1. Do NSs produce assertive tag questions with a predictable (falling) intonational contour?

2. Do NNS who are native Japanese speakers produce English tag questions with the same (falling) intonational contour?

3. Are there additional acoustic differences between accented productions and unaccented productions.

1.2 NS intonation patterns in English tag questions

Regardless of theoretical framework of intonation assumed, there is general agreement that native speakers use intonation in a predictable fashion and that English tag questions have typical patterns. Celce-Murcia (1996), offers ESL instructors a common description of "tag questions eliciting agreement" using a description of a rising-falling contour and a sequence of 3 pitch levels in the following example:

We really ought to visit him, shouldn't we?
2---------- 3—1----- 3----1-----

Ladd (1981) describes two types of tag questions in English. Rather than distinguish the two types as "rising" and "falling" tags, as they are frequently named, he defines them in terms of nuclear and post-nuclear tags. Those that are the focus of the current study are called falling tags by many authors and are described by Ladd as follows: "Nuclear tags have a separate nucleus or nuclear pitch accent, generally preceded in the rhythm of the sentence by a noticeable pause of intonational boundary."

Beckman and Pierrehumbert (1986) describe this type of question tag in their argument for an intermediate phrase boundary before the tag. In considering whether the tag should be treated as part of one intonational phrase or two separate ones, they observe

that placing the nuclear accent in the final phrase contradicts most listeners' impression that the main stress of the utterance is in the phrase preceding the tag. They offer that "positing an intermediate phrase boundary before the tag provides a way to handle the subordination." (p. 295)

Other accounts of standard American English contours, while not explicitly referencing "the tag question," do provide descriptions which suggest the existence of a typical intonation pattern for the expression being examined here. The utterance being examined in this study has an exclamatory, non-interrogative nature and is characterized by contrast – the contrast between the speaker's belief and what the speaker thinks her listener believes. Bolinger describes the exclamatory intonation that is used in the first phrase of the tag question in this study, noting that these contours "favour a high level followed by a fall from the accent." (Hirst & deCristo, p. 51) Pierrehumbert's compositional approach to tune specifies a relationship between the content proposed in the phrase and the "mutual beliefs" of the interlocutors. A tag question may use a pattern of pitch accents, and phrase and boundary tones to describe beliefs a speaker thinks he and his listener share.

1.3 NNS intonation patterns in English tag questions

Several studies comparing NS and NNS use of intonation support the observation that NSs of English mark contrast and salience (like that in the utterances studied in this paper) with standard intonation patterns.

Kelm (1987) prompted native and non-native speakers of Spanish to produce speech which contained contrasts and comparisons. He took measurements of frequency and intensity in the focus syllable and in the syllable immediately preceding focus for three groups of speakers – one group made up of native Spanish speakers (L1), another of native English speakers who spoke Spanish as a second language, and a third (control) group of native speakers of English who did not speak Spanish. Both groups of English speakers (the controls and the Spanish L2 speakers) regularly marked contrasts with higher pitch and greater intensity than did the NSs of Spanish. In the case of intensity measurements, the two groups of English speakers significantly differed from each other as well as from the Spanish speakers in that the English speakers learning Spanish appeared to have begun to use some Spanish patterns of intensity. In each of these cases, a standard intonation pattern was identified for native English speakers describing contrast.

Wennerstrom (1994) examined NSs of English and three different NNS language groups to investigate how speakers use intonation to assign significance in discourse. Though her material concentrated, for the most part, on statements, and did not include tag questions, she reports that native English speakers consistently marked salient items with pitch contrasts. She speculates that increased exposure to English was reflected in the contour measurements of more experienced L2 learners, supporting the observation that there are standard English intonation patterns available for them to acquire.

No studies were found which specifically examined NNS intonation patterns in English tag questions, but evidence does exist that NNSs, depending upon their speaking task, tend to produce regular intonation errors.

In the Wennerstrom (1994) study, speakers from all three L2 language groups (Spanish, Thai and Japanese) made less striking pitch contrasts on salient items than did native English speakers. In addition to significantly smaller pitch changes on both high and low pitch accents, NNSs failed to produce the marked intensity contrasts on low pitch accents that were used by NSs. Wennerstrom did include one interrogative in her text (though not a tag question), a Yes/No question, and noted that at the final boundary tone the NNSs of all 3 groups showed less pitch contrast than did the native English speakers. She additionally references Anderson's findings (1993) of another predictable difference between native and non-native speakers of English – that of tempo. Both NSs and the more experienced learners of English were reported to have smaller pause durations between pitch peaks.

Ueyama and Jun (1996) looked at focus realization in the speech of Korean and Japanese learners of English and explored interrogative intonation. Although they concentrated on the English interrogative pattern of L* H-H%, excluding other potential Y/N contours in English and excluding the tag question, they provided acoustic measurements describing a characteristic degree of slope in the fundamental frequency rise to the final H% for their NNSs of English.

1.4 Influence of intonation on perception of non-native speech

Second language acquisition researchers and instructors of second and foreign languages regularly attempt to identify which elements of spoken language (grammar and vocabulary aside) produce the impression that speech is non-native. Their interest includes each element's relative degree of contribution to that perception. They also address the issue of determining which of those elements contribute to greater or lesser degrees to that perception. For example, Munro (1995) asked English listeners to rate English utterances produced by native English speakers and utterances produced by native Mandarin speakers for accentedness. The utterances had been low-pass filtered to remove segmental information. Higher acceptability ratings were assigned to the utterances produced by the native speakers.

The degree of contribution of intonation toward accentedness as compared to the contribution of other variables, such as segmental deviance, is still unclear. Johansson (1978) asked listeners to rate accentedness in English utterances that had been purposefully produced with non-native intonation but contained accurate segmental content. They also rated English utterances with segmental errors but native-like intonation. Lower ratings were given to the prosodically compromised utterances. Anderson-Hsieh et al. (1992) examined the relationship between subjective ratings of NNS' oral proficiency and actual deviance in three areas – segmentals, prosody, and syllable structure, and reported significant correlations in all areas, but strongest effect

from the prosodic variable. It should be noted, however, that the authors used an impressionistic rating to measure "actual deviance in prosody."

Research on the intelligibility of deaf speech has reported significant negative correlations between prosodic elements and intelligibility by measuring erroneous speaking rhythm, slow speaking rate, and deviant accentuation, but subsequent study on the effect of correcting temporal structure on deaf speech intelligibility reports only small improvement (Maassen & Povel, 1984, p. 123, 124).

1.5 NS ability to identify NNS utterances based on intonation

It seems clear that native speakers of a language can perform at above chance level in identifying NS vs. NNS utterances on the basis of intonation alone. van Els & de Bot (1987) low-pass filtered NS and NNS utterances to remove segmental cues from their task and played the resulting sentences to NS jurors (Dutch speakers). The NNSs were English, French and Turkish individuals speaking Dutch as a second language. NS jurors correctly identified filtered NNS utterances as "not Dutch" at 79%. When the utterances' pitch alterations were replaced with an unchanging fundamental of 175 Hz, identifiability of specific language source (as opposed to merely Dutch or "not Dutch") dropped from 68% for non-altered utterances to 43% for "monotonized" utterances, supporting the importance of suprasegmental information to language source identification.

Ohala and Gilbert (1981) asked three groups of listeners (native speakers of English, Cantonese and Japanese) to listen to utterances produced by native speakers of all three languages in order to identify the native language of the speakers. The speech signal in the utterances had been converted to a buzz which retained the same frequency, amplitude and timing of the original speech. After hearing training samples, the listeners correctly identified which of the three languages were spoken at 58%, significantly above chance level of 33.3%. Identification scores increased with the use of long passages over shorter utterances.

1.6 Selection of analytic framework

While it is clear that non-native productions differ perceptibly from native productions, it is not clear what aspects of the productions support these different perceptions. The approach taken in the current study examines two ‘levels’ at which such differentiation might reside. It’s possible that there are fine differences, perhaps related to fluency, that account for the different perceptions. Alternatively, it is possible that the differences are of a more categorical nature. I.e., to the extent that certain discourse elements typically have a particular intonation pattern, NNS may differ from NS in choosing an inappropriate or unlikely pattern.

Intonation theory provides a choice of frameworks to use in examining the choice of ‘pattern’. British traditions are loosely associated with an approach sometimes described as "movement" or "prosodic" theories – ones that describe intonation in terms of chunks and of whole movements of contours and patterns that express attitude, as in the theory of O'Connor & Arnold (1961). Objections to this framework include the

difficulty of mapping the observed "attitude" onto specific contours and the issue of redundancy or the lack of a one-to-one mapping of speaker intent to a specific tune.

Phonemic theories, based on parts and sequences, pitch levels and pitch directions, offer more flexibility in breaking an analysis into components but also pose difficulties, for instance, knowing which phonetic details distinguish one level from another.

Discourse analysis theories reject generative mapping of underlying to surface representations and argue "that intonational choices speakers make are motivated by their moment-to-moment, situationally-specific decisions to add meaning to particular words or groups of words." (Chun, p. 36) Such a view is in concert with some of the recent speech synthesis science which rejects the description of contour based on components, and uses n-dimensional vectors to model continuous parameters (Cosi, et. al, 2002).

The autosegmental model of intonation introduced by Pierrehumbert in 1980 and subsequently extended by Beckman and Pierrehumbert (1986), Pierrehumbert and Beckman (1988) and Pierrehumbert (1990) seems best suited to the analyses undertaken in this study as it offers the ability to describe an utterance in terms of compositional tunes using categories which convey discourse information, and is "explicit in separating phonological constituency….from phonetic implementation…" (Bartels, p. 15).

A brief summary of the model and a short inventory of the components of tune are included below and in Section 3.2.1.

> "S(peaker) chooses an intonational contour to convey relationships between (the propositional content of) the current utterance and previous and subsequent utterances – and between (the propositional content of) the current utterance and beliefs H(earer) believes to be mutually held. These relationships are conveyed compositionally via selection of pitch accent, phrase accent, and boundary tone. Pitch accents convey information about the status of discourse referents and of relationships specified by accented lexical items. Phrase accents convey information about the relatedness of intermediate phrases, particularly whether one intermediate phrase forms part of a larger interpretive unit. Boundary tones convey information about whether the current intonational contour is 'forward-looking' or not." (Pierrehumbert & Hirschberg, p. 308)

The present study, then, examines NS and NNS productions which clearly differ perceptibly in terms of categorical encoding, using the ToBI framework.

2 Methods

2.1 Subjects

Three sets of subjects were used in this study – two groups of talkers and one set of listeners as jurors. One of the two groups of speakers was composed of ten females who are native speakers of American English from the Midwest portion of the US. All of these talkers were drawn from the community of a Midwestern, university-based city, reported normal speech and hearing and were between the ages of 21 and 45 years. The second group of talkers was made up of ten females who are native speakers of Japanese and speak English as a second language, ranging in age from 22 to 42 years. All were graduate students at Indiana University from various disciplines who reported normal speech and hearing capabilities. All but three had been in the US for at least a year, and all but one possessed a proficiency level beyond the requirements of the university's Intensive English Program, an instructional program designed to help non-native English speakers develop the skills required for admission at a North American university. (One speaker was currently enrolled in the university Intensive English Program.)

The NS juror subject group contained five males and five females, solicited from e-mail calls to the Indiana University undergraduate and staff communities. All were naïve to linguistic study, ranged in age from 20 to 41 years, and reported normal speech and hearing. Each completed a questionnaire designed to screen out subjects with bias regarding foreign-accented speech.

2.2 Material recorded and preparation of stimuli

Both sets of talkers (NS and NNS) were asked to read a 250-word text into which three rising and three falling tag questions had been embedded and which narrated a conversation between two females whose remarks included angry and insistent content. The design was intended to elicit intonation on tag questions which would capture a speaker's intent to demand agreement. The material included no complex syntactical structures. (The text of the four paragraph task is included as Appendix A.)

Recordings were made in a quiet room in the homes or work settings of the speakers. Speakers wore a Shure SM10A low-impedance, noise-canceling, unidirectional dynamic headset microphone to record their speech onto digital audio tape using a Sony TCD-D8 Digital Audio Tape Corder. The recordings were digitized (16-bit, 22,050 Hz) using Sound Forge Audio software, and the six tag questions from each talker’s recording were excerpted and subjected to low-pass filtering in Sound Forge to remove frequencies above 300 Hz, then saved as digital .wav files. To verify that the filtering had removed segmental cues, excerpts of ten, two-word portions were subjected to the same filtering process and were played for two NS listeners (not members of the NS listening jury) who were asked to write down the "two words" they understood the talkers to say. These individuals were unable to determine what was being said in the filtered speech. Pitch tracking of the two-word stimuli showed that the vocalic portions of the intonational contours were unchanged after filtering.

2.3 Jury evaluation

Native listeners performed their task one–by-one in a sound-proofed, quiet room, wearing digital binaural stereo output earcup headphones. They were instructed that both native speakers and non-native speakers of English had recorded a short story and that excerpted portions of that story had been degraded in quality so that they would not be able to hear individual sounds, syllables or words, but would be able to hear the intonation "or melody" of each phrase.

Their task was to listen to each numbered phrase, which would be repeated one time, and to mark the corresponding phrase number on their score sheet to indicate whether or not the speaker was a native speaker of English. Five of the jurors were allowed to make their selection from four choices (probably a native speaker, definitely a native speaker, probably not a native speaker, definitely not a native speaker). The other five jurors selected from only two choices (native speaker, not a native speaker). They were informed that they would not be allowed to go back to listen to any phrases again and to make their best judgment in cases of uncertainty.

The listeners were given scoring sheets that provided the text of each phrase they would hear so that they would know what was being said. Before beginning their listening task, they were also asked to follow along with a written copy of the narrative as they listened to a non-native speaker of American English (not a member of the NNS subject group) read the entire story from which the excerpts were taken. That reading was not filtered or distorted in any way. Each juror judged 240 tag question tokens ordered randomly (20 talkers x 6 tag questions each x 2 repetitions).

Average within-juror reliability was 78% (across both groups of speakers); average inter- juror reliability was 73% (across both groups of speakers). The decision to give five jurors 4 choices as opposed to 2 choices was deemed not to provide useful information. Thus, the responses that were marked as "probably" a native speaker were counted as "native speaker" judgments; those marked "probably not" a native speaker were counted as "not a native speaker."

2.4 Selection of tokens to be analyzed

The reading was designed to produce both rising and falling tag questions. Situations were embedded into the text which, pragmatically, would elicit two different renderings of a tag question – (1) a speaker tests a belief in his first phrase and adds the tag in order to request confirmation from his listener (It's sweet, don't you think?), and (2) a speaker *asserts* a belief in his first phrase and adds the tag in order to insist on agreement. (It *is* sweet, *isn't* it.) The second example produces a falling tag and is a "question" in syntactic form only, but not in function. The utterances in the first example which were produced merely to request confirmation, "unaccented post-nuclear tags" (Ladd, 1981) are not considered in the current study.

Nine NS utterances (representing four different speakers) and 11 NNS utterances (representing five different speakers) were selected from the jurored recordings for ToBI and acoustic analyses on the basis of having received "unanimously" correct scores (as either native or non-native) by the NS jurors. The text for these phrases has been divided into the designations, "Phrase 1" and "Phrase 2," for reference in all analyses. Table 1 shows how the utterance is divided into Phrase 1 and Phrase 2 for the purposes of description and provides the text in each phrase.

The lexical content of the tag questions examined differs in only a minor way; the lexical content of the third word of the phrase is sometimes the word "sweet," "easy," or "dark" (See Table 1 for a list of all phrases). Although the difference in the lexical content in phrases allows differing effects of adjacent consonants and inherent vowel frequency differences, both of which are present in the instrumental measurements used here, that effect is considered insignificant in terms of juror perceptions. "In terms of pitch differences that are perceived by hearers, it is only duration and loudness, not vowel type or adjacent consonant, that carry intonational functions and thus influence perception." (Chun, 2002, p. 5)

3 ToBI Analyses and Results

3.1 Tune types: ToBI Analyses

Table 2 provides a summary of the ToBI analyses for both sets of speakers and both phrases. Its content is discussed in the sections that follow.

3.1.1 ToBI analyses of Phrase 1

There are three different areas of observation in the patterns.

Number of Pitch Accents:

Three of the four NSs who produced the nine NS utterances used only one pitch accent in Phrase 1, always H* or L+H*, and always on Syllable 2. H* accents convey items to be added to the mutual belief space of speaker and hearer. Syllable 2, the word "is," would not be a normal candidate for accent except in a context such as the one used here where a speaker is insistent about the existent quality of Syllable 3 (sweet, dark, easy).

The NS who used two pitch accents in Phrase 1 placed H*+L accents on Syllable 2, followed by downstepped !H* accents on Syllable 3. This is a pattern NSs may employ in a "finger-wagging lecturing style where the clear intent of the style is to indicate that 'you should know this by now'." (de Jong, 2001).

NNS-J	TAG QUESTION	
	Phrase 1	**Phrase 2**
NNS-J 1.1	It is sweet	isn't it
NNS-J 1.2	It is dark	isn't it
NNS-J 1.3	It is easy	isn't it
NNS-J 2.1	It is sweet	isn't it
NNS-J 2.2	It is dark	isn't it
NNS-J 2.3	It is easy	isn't it
NNS-J 3	It is dark	isn't it
NNS-J 4.1	It is dark	isn't it
NNS-J 4.2	It is easy	isn't it
NNS-J 5.1	It is sweet	isn't it
NNS-J 5.2	It is dark	isn't it
NS-AE	**TAG QUESTION**	
	Phrase 1	**Phrase 2**
NS-AE 1	It is easy	isn't it
NS-AE 2.1	It is sweet	isn't it
NS-AE 2.2	It is easy	isn't it
NS-AE 3.1	It is sweet	isn't it
NS-AE 3.2	It is dark	isn't it
NS-AE 3.3	It is easy	isn't it
NS-AE 4.1	It is sweet	isn't it
NS-AE 4.2	It is dark	isn't it
NS-AE 4.3	It is easy	isn't it

Table 1. Representative utterances of NNS and NS. Speakers are identified within each group by number. Multiple excerpts for any speaker are indicated by sub-numbering. (Thus, NNS-J 2.3 = third utterance by second Japanese NNS.)

Utterance	Number of BI=2	Final BI	Number of Accents	Syllables with Accents	Accents	Final Tones	Syll 2 BI	# of Accents	Syllables with Accents	Accents	Final Tones
NS-1.1	0	3	1	2	L+H*	L-	0	1	1	L+H*	L-L%
NS-2.1	0	3	1	2	H*	L-	0	1	1	L+H*	L-L%
NS-2.2	0	3	1	2	L+H*	L-	0	1	1	H*	L-L%
NS-3.1	0	3	2	2+3	H*+L !H*	L-	0	1	1	H*	L-L%
NS-3.2	0	3	2	2+3	H*+L !H*	L-	0	1	1	L+H*	L-L%
NS-3.3	0	3	2	2+3	H*+L !H*	L-	0	1	11	H*	L-L%
NS-4.1	0	3	1	2	H*	L-	0	1	1	H*	L-L%
NS-4.2	0	3	1	2	L+H*	L-	0	1	1	H*	L-L%
NS-4.3	0	3	1	1+3	L+H*	L-	0	1	1	H*	L-L%
NN-1.1	2	3+	2	1+3	H* !L+H*	L-	1	1	1	H*	L-L%
NN-1.2	2	3+	2	1+3		L-	1	1	1	H*	L-L%
NN-1.3	2	3+	2	1+3	H* !L+H*	L-	1	1	1	H*	L-L%
NN-2.1	2	3	2	2+3	H* !L+H*	H-	1	1	1	H*	L-L%
NN-2.2	1	4	2	2	L+H* L+H*	L- L%	1	1	1	H*	L-L%
NN-2.3	1	3	1	2	L+H*	L-	1	1	1	H*	L-L%
NN-3.1	0	1	1	2+3	L+H*		0	1	1	L+H*	L-L%
NN-4.1	0	3	2	2+3	L+H* !H*	L-	1	1	1	L+H*	L-L%
NN-4.2	0	3	2	2+3	L+H* L+H*	L-	1	1	1	L+H*	L-L%
NN-5.1	1	3	3	All	H* L+H* !H*	L-	1	1	1	L+H*	L-L%
NN-5.2	1	3	2	2+3	L+H* !H*	L-	1	1	1	H*	L-L%

Table 2. Summary of ToBI analysis of intonation patterns found in unanimously juried utterances.

All but two of the NNSs Phrase 1 utterances were produced with two, and in one case, three pitch accents. With one exception, these were either all H* accents or some variation on H* (H*+L, L+H*, !H*). Two of the speakers who used two pitch accents placed them on Syllables 1 and 3. Another two speakers placed them on Syllables 2 and 3. Or course, when *three* pitch accents were used, every syllable in the phrase was accented, creating an abnormal pattern of cues for the hearer who might struggle to distinguish salient from non-salient items. It is likely that the NNS overuse of pitch accents arose from imperfect fluency rather than from an intention to suggest that each of the three syllables be treated as new information. More discussion of disfluency effects is provided in the next section.

NNS Use of Break Index "2":

Eight of the 11 NNS utterances included at least one occurrence of Break Index 2. A brief description of the ToBI assignment of break index values follows (Pierrehumbert & Hirschberg, 1990) with a description of the less common Break Index 2 at its conclusion:

"Break indices represent a rating for the degree of juncture perceived between each pair of words and between the final word and the silence at the end of the utterance."

Break Index 0:
"…the lowest level break index (0) is defined in terms of connected speech processes, such as the flapping of word-final /t/…

Break Index 1:
…"the label to be used for "most phrase-medial word boundaries"

Break Indices 3 and 4:
"…are equated with the intonational categories of intermediate (intonation) phrase and (full) intonation phrase."

Break Index 2
"devised to mark cases of …two types of 'mismatch' between the subjective boundary strength and the intonational constituency. These two types are described in the ToBI Annotation Conventions as follows:

> (1) a strong disjuncture marked by a pause or virtual pause, but with no tonal marks; i.e. a well-formed tune continues across the juncture.
> OR
> (2) a disjuncture that is weaker than expected at what is tonally a clear intermediate or full intonation phrase boundary."

Though not devised to account for issues of disfluency or hesitation, (1) above was determined to be the most appropriate description of the juncture between syllables in many of the NNS Phrase 1 utterances. This usually occurred when both Syllables 1 and 3 were pitch-accented and the high tone continued across the juncture. Break Index 2 was also used following Syllable 1 in cases where Syllable 2 did carry a tonal mark but where the degree of juncture between Syllables 1 and 2 was greater than the strength of a typical word boundary. The disfluency conveyed by the NNS tendency to overuse pitch accents is heightened by Break Index 2 frequency which, itself, causes a subjective perception of pitch accents, even in cases where a pitch "event" has not occurred.

<u>Boundary Tones Between Phrase 1 and Phrase 2</u>:

Characterization of the juncture between the two intermediate phrases in the NNS utterances is somewhat like the issue of the mismatch present in observations of Break Index 2. NNS 1.1, 1.2 and 1.3 utterances offer the subjective perception of a juncture stronger than what would be expected at an intermediate phrase but provide no contour cues to a full intonational phrase boundary between Phrases 1 and 2. These occurrences are marked in Table 2 with the Break Index 3+. (This is not a break index under the

current ToBI labeling guidelines but is invented here to indicate the excessive juncture at these locations.)

3.1.2 ToBI analyses of Phrase 2

Native and non-native speakers used near identical components in Phrase 2 but differed in the break index values between Syllables 2 and 3 (between the words "isn't" and "it"). No NSs produced a word break between the contraction "isn't" and the following word "it" (Break Index=0), resulting in their pronunciation of Phrase 2 as the continuous form "IsnIt," while all but one of the NNSs assigned a typical word break (Break Index=1) at this location. With this as the only distinctive difference between the speaker groups, it would seem, from a ToBI analysis standpoint, that Phrase 1 contributes most dramatically to the perceptual distinction between native speaker and non-native speaker intonation.

3.2 Acoustic differences

3.2.1 Acoustic measurements in Phrase 1

As fundamental frequency is one of several correlates of intonational prominence, F0 measurements were taken for syllables of interest in Phrase 1. NSs unanimously placed the highest frequency in Phrase 1 on Syllable 2, while five of the NNS utterances placed the highest F0 for Phrase 1 on either Syllable 1 or Syllable 3.

Also at odds with the NS frequency pattern was the degree of pitch excursion from Syllable 2 to Syllable 3. NS dropped an average of 102 Hz from the peak value of Syllable 2 to the peak value of Syllable 3 while non-native speakers dropped an average of only 48 Hz. When frequency values are normalized as percentages of the F0 range over the entire utterance for each speaker, NSs frequency excursion between Syllable 2 and 3 was a drop of 60% of their total pitch range, while NNSs dropped only 23% of their total pitch range. *T*-tests show that the differences between NS and NNS frequency excursion values are significant at $p<.01$.

Figure 1 shows the distribution of individual speakers' values for this measurement. NNSs showed a greater degree of within group variance than did the NSs-AE and a range of values that had only slight overlap with the NS values.

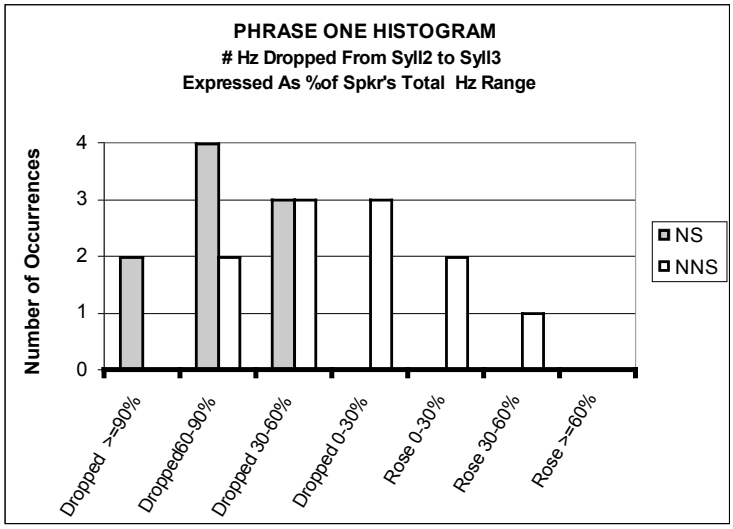

Figure 1. Frequency distribution of degree of pitch excursion between Syllable 2 and Syllable 3 in Phrase 1.

In a series of experiments investigating the existence of a categorical boundary between normal and emphatic intonational emphasis (in native speakers of English), Ladd and Morton (1997) found results that suggested "that listeners are predisposed to interpret … utterances as being categorically either 'normal' or 'emphatic'." If this observation is correct, it might be that the NS jurors in the current study, being aware of the emotive and lexical content of the filtered utterances they judged, expected that a NS would employ a pitch change beyond the boundaries of "normal" excursion and within the category boundaries of "emphatic". If so, the characteristics of the NNS smaller pitch excursions between Syllables 2 and 3 in Phrase 1 could contribute to their being judged as NNS utterances.

3.2.2 Intensity measurements in Phrase 1

Using the intensity values provided by PitchWorks software, which converts the average RMS of the points in a window (30 ms. window, sampled at 11,025 Hz) to dB values, measurements of peak intensity were taken on Syllables 2 and 3 in Phrase 1. NSs dropped an average of 8.0 dB from Syllable 2 to Syllable 3 while NNSs dropped an average of only .73 dB. Because Syllable 3 might be from one of three different words ("sweet," "dark," or "easy"), correction for inherent vowel intensity values (Lehiste & Peterson, 1959) were calculated, resulting in minimal corrections, to report an 8.5 dB average drop for NSs and a .24 dB average drop for NNSs. When normalized by being expressed as a percentage of the highest intensity used in each speaker's utterance, the NS average drop in intensity from Syllable 2 to Syllable 3 is 18% of their total intensity range while NNSs averaged a drop of one-half of a percent of their total intensity range. Differences between NS-AE and NNS-J intensity excursion values are significant at $p < 0.01$.

3.2.3 Syllable duration measurements in Phrase 1

Syllable duration measurements in Phrase 1 were also of interest. In an examination of duration and intensity as physical correlates of stress, Fry (1955) concluded that the duration ratio between two syllables in isolated words is a stronger predictor of stress than is the same ratio for intensity. Using spectrographic analysis of the utterances before they had been filtered to remove segmental information, measurements were made of the vowel portion of Syllable 2 of Phrase 1. As Syllable 2 was always preceded by either a flap or an aspirated or unaspirated /t/ (end of preceding word "it") and was always followed by the /z/ segment (end of the word "is"), the measurement boundaries for the vowel were clear.

In terms of absolute duration, the vowel of NSs was 49% longer than that of NNS-J. When normalized as a percentage of the total utterance duration, the vowel in Syllable 2 of Phrase 1 is 95% longer for NSs than for NNSs. "Total utterance" is made up of both Phrase 1 + Phrase 2, and measurements were made from onset of voicing at outset of utterance through and including any aspiration of released final consonant on utterance final word. Differences between NS-AE and NNS-J Syllable 2 duration values are significant at $p<.01$.

Figure 2 shows the distinction between NSs' and NNSs' use of duration and pitch excursion. NS datapoints, with longer Syllable 2 and more dramatic F0 drop from Syllable 2 to 3, are located in the upper left quadrant of the plot.

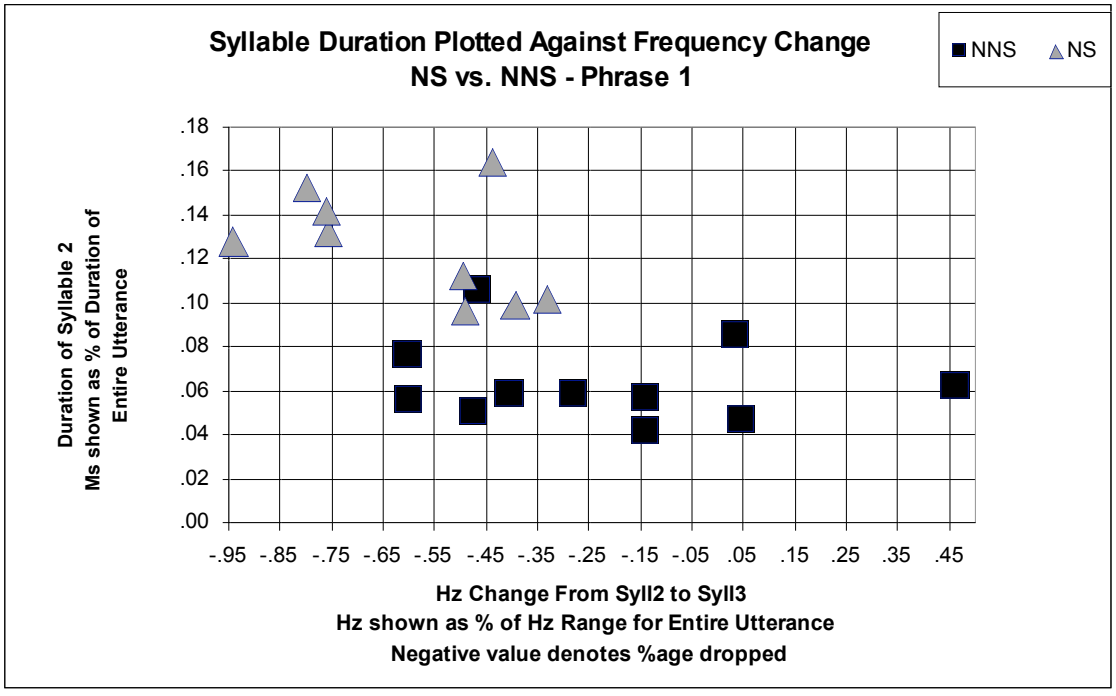

Figure 2. X-axis shows change in frequency from Syllable 2 to Syllable 3, expressed as a percentage of each speaker's highest intensity value. Y-axis shows duration of Syllable 2, expressed as percentage of total ms. in each speaker's utterance.

The distinctive use of intensity by NSs vs. NNSs is added to this information in Figure 3 where drop in Hz value from Syllable 2 to 3 and drop in intensity from Syllable 2 to 3 are plotted on the x- and y-axes, respectively, and duration of Syllable 2 is indicated by size of bubble marker.

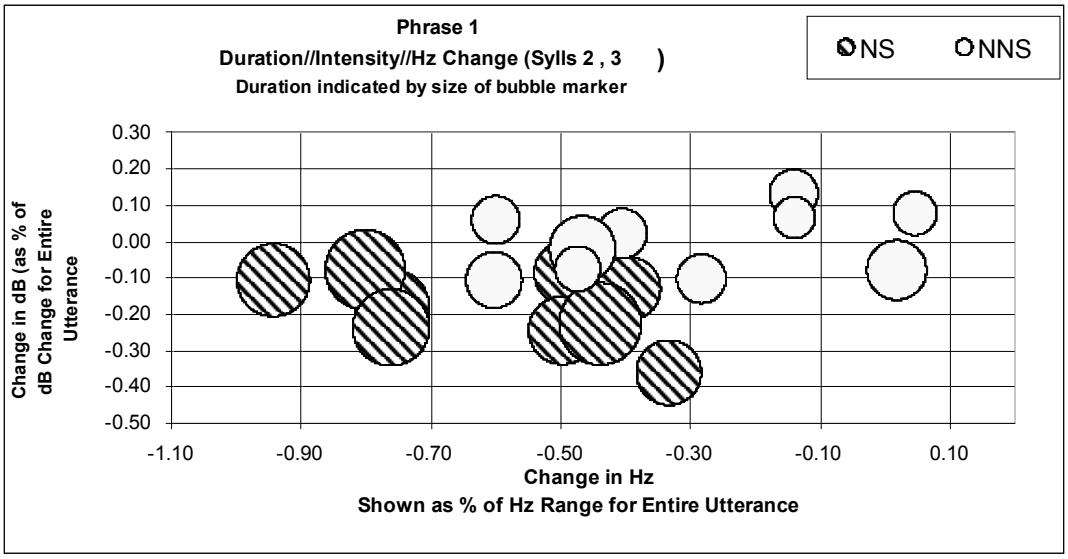

Figure 3. NSs have a larger drop in frequency (x-axis) and a larger drop in intensity (y-axis) between Syllables 2 & 3 in Phrase 1 than do NNSs. Duration of Syllable 2 in Phrase 1 is longer for NSs than for NNSs (size of bubble marker).

3.5 Comparison of acoustic measurements in Phrase 2

3.5.1 Intensity measurements in Phrase 2

The only acoustic comparison of interest in Phrase 2 was that of the degree of intensity drop between Syllables 2 and 3 (between the words "isn't" and "it"). NSs dropped an average of 2.4 dB from Syllable 2 to Syllable 3 in Phrase 2 while NNSs dropped an average of 8.1 dB for the same location. When normalized as percentages of intensity ranges over entire utterance for each speaker, NSs' drop in intensity from Syllable 2 to Syllable 3 averaged 5% of speakers' total intensity ranges while NNSs' averaged a drop three times as great, 17.5% of their total intensity ranges. The difference was significant at $p<.05$.

This observation supports the differences in ToBI break index values in Phrase 2 for the two groups of speakers (discussed in Section 3.3.2) and emphasizes the NNS tendency to retain individual word identity in Phrase 2.

4 Discussion

4.1 Summary of Findings

The ToBI transcription labeling system was used to describe the intonation of utterances which had been perceived by native speaker jurors as produced by either non-native or native speakers of English. Comparison of the transcriptions of the two groups of speakers showed that despite having been identified by jurors as produced by NNSs, the NNS-J utterances were not *compositionally* at dramatic odds with those of the NSs.

Phrase 1 of the tag utterance contained most of the differences between groups, and the NS group was more homogeneous than was the NNS group. NNSs used more pitch accents in Phrase 1 and inserted breaks between words in both phrases more often than did the NSs. NNS breaks between words in Phrase 1 were ones that are described by a mismatch between the strength of disjuncture between two words and the tonal events expected for such a disjuncture and were designated with Break Index 2. A similar observation of "mismatch" was evident at the end of a small number of NNS intermediate phrases (end of Phrase 1) where tonal events did not prescribe an intonational phrase boundary but strength of juncture subjectively seemed greater than the break normally occurring with Break Index 3.

Subsequent acoustic measurements in each phrase affirmed notable differences between speaker groups, particularly in Phrase 1 where NSs produced a H* (or variant of H*) pitch accent on Syllable 2 that was of significantly greater duration than that of NNSs. Additionally, NSs' downward excursions of both frequency and intensity between Syllable 2 and Syllable 3 were significantly larger than those of NNSs. In Phrase 2 NSs treat the lexical components as one chunk while NNSs allow the words to retain discrete identities.

In addition to the results cited above regarding examinations of each separate phrase of the utterance, two *utterance-level* differences between NS and NNS groups were noted.

NSs assigned the highest frequency value for the entire utterance within Phrase 1 (specifically on Syllable 2). Two of the NNSs violated this pattern by placing their absolute frequency high value for the utterance on a syllable in Phrase 2, disallowing the declination of the second phrase of the utterance, a pattern common to NSs.

NSs located the absolute high for vocalic intensity over entire utterance either in Phrase 1 Syllable 2 or in Phrase 2 Syllable 1. NNSs were not as consistent in their location of highest vocalic intensity. Five of the 11 utterances followed the NS pattern, five selected Phrase 1 Syllable 3 (eg, It is *SWEET*), and the remainder selected Phrase 2, Syllable 2 (is*N'T* it).

4.2 Limitations of the study

4.2.1 Text narrative vs. free speech

An examination of intonational meaning in discourse, whether in the speech of native speaker or non-native speaker, is optimally made in free discoursal settings. This study attempted to elicit speech patterns consistent with the emotive nature of one form of the English tag question by developing a narrative with context that adequately provided speaker attitude. It cannot be said, however, to have reproduced free speech patterns which allow extended discourse functions.

4.2.2 Interactions between acoustic variables

Except for the observations regarding differences in NS and NNS patterns of syllable duration and pitch excursion, no attempt has been made within this study to examine the interaction of the components of prominence, nor has this study attempted to determine to what degree each of these parameters contributed to NS judgments that an utterance was produced by a NNS.

4.2.3 Sample size

Only 20 utterances were examined, and they were produced by a total of nine speakers. Though the patterns found within these samples showed consistent tendencies, particularly, and as expected, within the NS utterances, a larger sample size would more reliably indicate patterns.

4.3 Possible future investigations

The differences between NS and NNS acoustic measures of syllable duration, pitch and intensity excursions were large enough to suggest a future investigation of whether or not the perception of native vs. non-native speech is categorical or continuous, i.e. whether perceptions of difference change continuously throughout the range from NS to NNS patterns or whether NNS patterns are perceived as such until their variance from NS speech is of a particular magnitude . A related investigation could be undertaken to determine the degree of variance in each measure that is tolerated before a speech sample crosses from the perception category of NS to NNS. Finally, the possible influence of L1 on intonation errors is an important area to consider in future examinations because of its possible impact on ESL teaching methods.

References

Anderson-Hsieh, Janet, Ruth Johnson and Kenneth Koehler. (1992). The relationship between native speaker judgments of nonnative pronunciation and deviance in segmentals, prosody, and syllable structure. Language Learning, 42:4, December, 529-555.

Bartels, Christine. (1999). The Intonation of English Statements and Questions: A Compositional Interpretation, Garland Publishing, Inc., New York.

Beckman, Mary E. and Gayle Ayers Elam. (1997). Guidelines for ToBI Labelling, v. 3. Columbus, OH: The Ohio State University.

Beckman, Mary E. and Janet B. Pierrehumbert. (1986). Intonational structure in Japanese and English. Phonology Yearbook, 3, 255-309.

Bolinger, D. (1998). Intonation in American English. In D. Hirst & A. DiCristo (eds.) Intonation Systems: A Survey of Twenty Languages , UK: Cambridge University Press, 1-56.

Celce-Murcia, M., D. Brinton and J. Goodwin. (1996). Teaching Pronunciation, A Reference for Teachers of English to Speakers of Other Languages. UK: Cambridge University Press.

Chun, Dorothy. (2002). Discourse Intonation in L2. From theory and research to practice, Philadelphia, PA: John Benjamins Publishing Company.

Cosi, Piero, Fabio Tesser, Roberto Gretter and Fabio Pianesi. (2002). A modified "PaIntE" model for Italian TTS. IEEE Workshop on Speech Synthesis, Santa Monica, CA, September 11-13, 2002.

de Jong, Kenneth (2001). Notes on English Intonation. (Class notes distributed to L541, Spring 2001).

Fry, D. B. (1967). Duration and intensity as physical correlates of linguistic stress. In Ilse Lehiste (ed.) Readings in Acoustic Phonetics, Cambridge, MA: MIT Press, 155-158.

Kelm, Orland R. (1987). An acoustic study on the differences of contrastive emphasis between native and non-native Spanish speakers. Hispania, 70, 3, 627-633.

Johansson, S. (1978). Studies of error gravity: Native reactions to errors produced by Swedish learners of English. Gothenburg Studies in English, No. 44, Goteborg, Sweden: Acta Universitatis Gothoburgensis.

Ladd, D. R. and R. Morton. (1997). The perception of intonational emphasis: continuous or categorical? Journal of Phonetics, 25, 313-342.

Ladd, D. Robert. (1996). Intonational Phonology, UK: Cambridge University Press.

Ladd, D. Robert. (1986). A first look at the semantics and pragmatics of negative questions and tag questions. Papers from the Regional Meetings, Chicago Linguistic Society, 17, April-May, 164-171.

Lehiste, Ilse and Gordon E. Peterson. (1967). Vowel amplitude and phonemic stress in American English. In Ilse Lehist (ed.) Readings in Acoustic Phonetics, Cambridge, MA: MIT Press, 183-190.

Maassen, Ben and D. J. Povel. (1984). The effect of correcting temporal structure on the intelligibility of deaf speech. Speech Communication, 3, 123-135.

Munro, Murray (1995). Nonsegmental factors in foreign accent: ratings of filtered speech. Studies in Second Language Acquisition, 17, 1, 17-34.

O'Connor, J. D., and G. F. Arnold, G. F. (1961). Intonation of colloquial English (2nd ed., 1973). London: Longman.

Ohala, John J. and Judy B. Gilbert. (1981). Listeners' ability to identify languages by their prosody. In P. Leon and M. Rossi (eds.) Problemes de Prosodie, Vol II Experimentations, Modeles et Fonctions. Paris: Didier, 123-131.

Pierrehumbert, Janet, and Julia Hirschberg (1990). The meaning of intonational contours in the interpretation of discourse. In P. R. Cohen, J. Morgan and M.E. Pollock (eds.), Intentions in Communication, Cambridge, MA: MIT Press, 271-311.

Ueyama, M. and S. A. Jun (1996). Focus realization of Japanese English and Korean English intonation. UCLA Working Papers in Linguistics, 94.

Van Els, Theo and Kees De Bot. (1987). The role of intonation in foreign accent. The Modern Language Journal, 71, 147-155.

Wennerstrom, Ann. (1994). Intonational meaning in English discourse: a study of non-native speakers. Applied Linguistics, 15, 4, 399-420.

Appendix A
Narrative Text

Jane and Karen were *not* good friends. They annoyed each other and often argued. Tonight, as they walked down the road munching fresh fruit, Jane grew frightened by the night's shadows. She turned to Karen and said nervously, "It's dark, don't you think?" Karen answered by saying, "I guess so, but don't worry. Just eat your apple. It's sweet. You'll see." Karen waited while Jane tasted the apple and then asked…"it's sweet, don't you think?"

Jane smiled at the apple's pleasant taste. Karen was smugly happy that Jane was enjoying the apple and said, "I knew it, I *knew* you would like it! You *can't* deny it! Karen grew more insistent and said, "Now *say* it! It *is* sweet, *isn't* it!"

"Yes," agreed Jane, "but I am still worried about walking here at night." Growing even more frightened, she shouted, "It's dark, it's dark, it's so very dark, and I won't stop worrying until you admit that it's dark. Now *say* it…It *is* dark, *isn't* it!"

Karen hesitated. "Maybe a little bit, but if you try to be brave, it's easy, don't you think?" "No," insisted Jane. "I just think you are the kind of person who is never scared, so to tell someone else to be brave is *easy* for you. For you, it's always *so* easy. She became angry and shouted, "It's easy, it's easy, it's easy! Now tell me that you agree. *Tell* me! It *is* easy, *isn't* it!

Karen didn't answer, and the two girls walked on in silence.