Automatic Labeling of Intonation and Rhythm for Scoring Speaking Proficiency

Jackson Liscombe

Columbia University
New York, New York, USA

Abstract

This paper describes work evaluating the extent to which descriptors of intonation and rhythm are useful for estimating human scores of the delivery proficiency of spoken English. It is shown that distributional measurements of syllable stress and intonational phrase boundary tones (prosodic events) correlate with human ratings in a corpus of speaking proficiency test-takers. Furthermore, automatic prediction of human ratings using these features approaches agreement of human judges. We also describe work on automatically predicting prosodic events and demonstrate that distributional features derived from such predictions still significantly correlate with human ratings.

Key words: non-native speaking proficiency, automatic scoring, intonation, ryhthm, delivery, machine learning, prosody

1 Introduction

There are thousands of human languages spoken in the world today, an indication that the world in which we live is still a multilingual one. A person's **native language** is learned very early on in life and is used to communicate with others in his or her speech community. However, one is capable, through practice, to learn other languages later in life. Such a language is usually referred to as a **second language**. In the study of language acquisition, a person's native language is referred to as **L1** and a second language is referred to as **L2**. It is an interesting and well-established fact that an L2 speaker is rarely perceived to have native-like proficiency by L1 speakers of

Email address: jaxin@cs.columbia.edu (Jackson Liscombe).

that language. Indeed, the later in life that one begins to learn a new language, the harder it becomes to sound like a native speaker of that language. It is an equally intriguing fact that people are inherently capable of instantly assessing the proficiency of a speaker of their own L1.

Though people of different L1 speech communities have always been in contact with one another, modern technology has enabled virtually every member of a speech community to easily be in contact with a member of a different speech community. It should come as no surprise then, that research using modern technology includes studies that aim to assess and teach L2 proficiency automatically using computer-based methods.

Many of today's academic and business institutions require English as the primary means of communication. These same institutions often judge the proficiency of applicants by requiring standardized testing designed to assess command of the English language. Though traditionally devised to measure listening, reading, and writing skills, many language proficiency exams now offer speaking portions as well. Computer-based automated scoring of speech has several potential advantages. It promises to reduce the cost of language assessment, has the potential to yield more consistent assessment scoring, and could also be used as an essential component of computer-assisted English language-learning software, thus providing more people with access to quality instruction.

2 Communicative Competence

Being a proficient speaker of a language demands more than proper grammar, vocabulary, and pronunciation. Though these are the three aspects of language proficiency that are usually taught in language courses, it is not uncommon for a person who has mastered these skills to nevertheless be perceived as a non-native speaker. This is because language use comprises more that syntactic and segmental properties. Discoursal, pragmatic, and sociolinguistic behavior are all critical components as well. In other words, how a language is used to accomplish different types of goals is an essential part of proficient language use.

A global view of language proficiency that includes all aspects of language use is referred to as **communicative competence** and was first put forth by Hymes 1971 and was later refined by Canale & Swain 1980 and Bachman 1990. This is the paradigm under which we are operating when we refer to spoken language proficiency in this chapter. The three dimensions of communicative competence are enumerated in Table 1. The first dimension, **language use**, includes aspects of language proficiency that have traditionally

I. Language Use			
A.	*Vocabulary*	*B.*	*Grammar*
	Diversity		Range
	Sophistication		Complexity
	Precision		

II.	Topic Development
	Coherence
	Idea progression
	Content relevance

III.	Delivery
	Pronunciation
	Fluency
	Rhythm
	Intonation

Table 1
Dimensions of communicative competence for speaking proficiency.

been emphasized when teaching an L2 and it includes grammar and vocabulary proficiency. The second dimension, **topic development**, can be thought of as the way in which a language structures discourse. The third dimension, **delivery**, includes segmental (**pronunciation**) as well as suprasegmental or prosodic behavior (**intonation**, **rhythm**, and **fluency**). We will be focusing on the delivery dimension of communicative competence in this chapter.

Pronunciation assessment concerns proficiency at the level of the phoneme. The goal of pronunciation assessment is to evaluate the spectral qualities of intended phonemes in order to determine how similar these qualities are to what a native speaker would produce. Most studies on automatic delivery assessment have focused on this sub-dimension of communicative competence, as we will show in the next section.

The term **fluency** can be defined in (at least) two ways. In its colloquial usage, it refers to global oral proficiency in all areas of speech production. As a technical term, however, it is often used by the L2 language assessment community to refer to temporal qualities and smoothness of flow in speech production—including phenomena such as speaking rate, segment duration, and pausological behavior. Filled pauses and hesitations are often included in this sub-dimension as well. Automatic assessment of the fluency proficiency of L2 learners has been studied to some degree in recent years and in this chapter we describe experiments we conducted in this area so that we might compare our findings with those studies. The remaining two sub-dimensions of delivery proficiency—**rhythm** and **intonation**—have been less studied with respect to automatic scoring techniques. These two sub-dimensions are the focus of the experiments presented in this chapter.

Rhythm—more specifically, rhythmic stress—describes the relative emphasis given to syllables in an utterance. English is known to be a rhythmically stress-timed language; that is, stressed syllables appear at a roughly constant rate. Syllable stress is conveyed in English through a variety of means: pitch, loudness, duration, and spectral information. Rhythmic stress is not equivalent to lexical stress. In English, lexical stress refers to the fact that each word has underlying stress on at least one syllable in a word. Where the stress lies is unpredictable and must be learned. Clearly, lexical stress is an important part of communicative competence but is more appropriately grouped with pronunciation. This is not to say that there is no relation between lexical and rhythmic stress. In spontaneous speech, many syllables that bear underlying stress become destressed; however, the converse is not generally true: underlying unstressed syllables rarely become stressed.

Intonation describes how pitch varies in speech. Intonational languages, such as English, use pitch variation to signal several types of meaning change. In such languages, intonation can be used syntactically; for example, to distinguish questions, which may exhibit phrase-final rising intonation, from statements, which tend to have phrase-final falling intonation. It can also be used to structure discourse flow by signaling to the listener when to expect a change in topic or turn-taking. Intonation can even be used to distinguish between paralinguistic meaning such as emotion (a topic addressed in depth in Part I). Clearly, intonation is an important aspect of communicative competence and yet it is rarely explicitly taught to language learners and is, in fact, one of the last skills L2 speakers acquire. This should not indicate that it is not important for an L2 language learner to neglect this area of language proficiency, though. As Gass & Selinker 1994, p. 184 state, "Miscommunication resulting from native speaker perceptions of relatively proficient non-native speakers as opposed to learners with low level comprehension and productive skills is more often the more serious in terms of interpersonal relations because the source of the difficulty is attributed to a defect in a person (or a culture)." In other words, though the literal content of the message may be correctly communicated, low proficiency in suprasegmental delivery may yield negative perceptions of the speaker.

3 Previous Research

Most studies on automatic delivery assessment have focused on pronunciation, with much success (Bernstein et al., 1990; Neumeyer et al., 1996; Franco et al., 1997; Cucchiarini et al., 1997, inter alia). In fact, there are even a few automatic proficiency assessment products on the market today, including EduSpeak Franco et al. (2000) and PhonePass Bernstein et al. (1990). With respect to the other sub-dimensions of delivery proficiency, however, not nearly

as much work has been done. Table 2 presents a comparison of notable recent studies on the automatic assessment of L2 fluency, rhythm, and intonation.

Study	Delivery			L1	L2	Speech
	flu	rhy	int			
Neumeyer et al. 1996	✓			English	French	read
Franco et al. 2000	✓			Japanese	English	read
Bernstein et al. 2000	✓			various	English	read/spontaneous
Teixeira et al. 2001	✓	✓	✓	Japanese	English	read
Cucchiarini et al. 2002	✓			various	Dutch	read/spontaneous
Herry & Hirst 2002	✓		✓	French	English	read
Toivanen 2003			✓	Finnish	English	read
Kormos & Dénes 2004	✓	✓		Hungarian	English	spontaneous
Morgan et al. 2004	✓			unreported	Mandarin	read
Hincks 2005	✓			Swedish	English	planned
Xi et al. 2006	✓			various	English	spontaneous
Present study	✓	✓	✓	various	English	spontaneous

Table 2
Previous research on the automatic assessment of three sub-dimensions of the delivery proficiency of L2 speech: fluency (**flu**), rhythm (**rhy**), and intonation (**int**).

All studies but one Toivanen (2003) have investigated frequency assessment, and most studies have looked at the delivery of spoken English (notable exceptions are Neumeyer et al. (1996); Cucchiarini et al. (2002); Morgan et al. (2004)). Additionally, most studies have used data from one L1 language community, though these communities have ranged from study to study and have included Japanese, Finnish, Hungarian, and Swedish. Another important variable to consider is the type of speech data used for analysis. Overwhelmingly, researchers have chosen to use read speech for the reason that it is highly constrained and can be automatically segmented into words easily. Two studies Kormos and Dénes (2004); Xi et al. (2006) examined only spontaneous speech while two others Bernstein et al. (2000); Cucchiarini et al. (2002) have looked at both read and spontaneous speech. Hincks 2005 used data from oral presentations wherein the speakers knew ahead of time what they planned to say, though they did not read from scripts.

There is a consensus among the studies that have explored fluency. It has been found that proficient speakers have a faster speaking rate, longer stretches of

speech between pauses (run length), and shorter pause lengths than do non-proficient speakers, regardless of L1 and L2. The reason proposed for these findings is usually described in terms of cognitive load. Non-proficient speakers have a higher cognitive load when accessing vocabulary and grammar rules than do proficient speakers and this increase in load is manifested as speech that is slower and that has longer and more pauses.

Most studies that have examined the automatic assessment of delivery proficiency have first extracted features from the speech signal and then have correlated these values with human assessment scores. Cucchiarini et al. 2002 found significant correlations between human raters and speaking rate (r = 0.6), run length (r = 0.5), pause length (r = -0.5), and pause rate (r = -0.3). Neumeyer et al. 1996 reported a correlation coefficient of 0.4 for speaking rate; Teixeira et al. 2001 observed a correlation coefficient of 0.4 for speaking rate and 0.5 for pause length; and Hincks 2005 reported a correlation coefficient of 0.4 for speaking rate and 0.5 for run length. Kormos & Dénes 2004 observed the highest correlation of fluency features and human assessment scores: 0.8 for speaking rate, -0.6 for pause length, and 0.9 for run length. It is difficult to compare absolute correlation values across studies because the factors between one study and the next can vary greatly by sample size, feature extraction method, L1 and L2, and the nature of human assessment scores. Nevertheless, it is clear that relatively simple fluency measurements can be extracted from the speech of non-native speakers that significantly correlate with human assessment scores.

The number of studies exploring reliable rhythm and intonation features that correlate well with human assessment scores have been few, and no consensus with respect to these sub-dimensions has yet emerged. Of the two, rhythm has been shown to be the most important. Kormos & Dénes 2004 found a correlation coefficient of 0.9 between the number of stressed words per minute and human assessment scores for Hungarian L1 speakers of English L2. Teixeira et al. 2001 approximated syllable stress by identifying the vowel of longest duration, the vowel with underlying lexical stress, and the vowel with the highest pitch. Using these stressed vowels as reference points, the author computed rhythm features by recording the stressed vowel length, the time between a stressed vowel and the previous and following intra-word vowels, and the time between the consecutive stressed vowels. No results were reported for the rhythm features alone, but they were shown not to increase correlation when combined with fluency features.

Teixeira et al. 2001 also examined intonation proficiency by extracting fundamental frequency ($f0$) measurements from their speech data as an approximation of global pitch behavior. Several measurements were calculated, including maximum $f0$, $f0$ slope, and $f0$ variation. The pitch feature set was found to have the weakest correlation with human raters but was found, when com-

bined with the fluency features, to increase correlation by 7.7% over using fluency features alone. [1]

Herry & Hirst 2002 also approximated intonation by using a direct pitch modeling approach, but instead of using absolute values they compared $f0$ statistics of non-native speakers with productions of two native English speakers. The features included $f0$ range, $f0$ standard deviation, and $f0$ slope variation. They found that the intonation features were not as critical, as were the fluency features on all assessment tests except one that measured the "repetition quality" of read speech, in which case they found $f0$ range to be most useful. [2]

Toivanen 2003 took a different approach to the analysis of intonation proficiency of non-native speakers by adopting the "British school" of abstract intonation analysis. Tonic syllables were identified by their pitch and loudness peaks and then labeled as one of the following perceptual tones: fall, rise-fall, rise, fall-rise, fall-plus-rise, or level. The distribution of these tones were compared between two corpora of read British English: one produced by L1 speakers of British English and the other by L1 speakers of Finnish. The author found that the Finnish speakers produced far fewer rising tones, and far more falling and level tones than did the British speakers. It was also found that Finnish speakers produced rising intonation with yes/no sentence types and falling intonation with declarative sentence types far more often than did British speakers, indicating the importance of intonation in the pragmatic nature of communicative competence. Toivanen 2005 described work on further annotation of the data in terms of word-level stress, voice quality, and ToBI intonation labels, though no results have been reported on this to-date.

The approach we have taken to the automatic assessment of fluency, rhythm, and intonation proficiency builds on the fluency work of Xi et al. 2006 and is most similar to the rhythm approach of Kormos & Dénes 2004 and the intonation analysis of Toivanen 2003. Our work benefits from the fact that our corpus of non-native speech was unconstrained and spontaneously generated and that the speakers came from different speech communities. In this way our findings aimed to generalize across language backgrounds and to emphasize the communicative competence of non-native speakers in real-life speaking scenarios. The following sections detail our corpus, annotation scheme, human rater correlation, and prosodic event detection techniques.

[1] Correlation using pitch features was 0.23, when using fluency features was 0.32, and when using both in combination was 0.34.

[2] The authors reported no correlation tests for these results.

4 The DTAST Corpus

The Test of English as a Foreign Language™ Internet-based Test (TOEFL iBT), developed and administered by the Educational Testing Service (ETS), measures the ability of non-native speakers of English to use and understand English as it is spoken, written, and heard in college and university settings. Today, there are more than 6,000 institutions in 110 countries that use TOEFL scores. The assessment test is designed for non-native English speakers at the 11th-grade level or above and is most frequently used to assess English proficiency before beginning academic work at an institution of higher education.

The TOEFL Academic Speaking Test (TAST) evaluates English speaking proficiency at intermediate to advanced proficiency levels. TAST presents integrated speaking tasks that simulate academic settings, just like the speaking portion of TOEFL iBT. TAST consists of six tasks. The first two tasks are independent tasks that require test-takers to talk about familiar topics. The other four tasks are integrated tasks that ask test-takers to listen to a short talk or to read a passage, and to then respond verbally to questions regarding these talks or passages. In total, the test takes about 20 minutes to complete and the data is recorded over the telephone. The audio data is recorded digitally at an 8 kHz sampling rate with 8 bit resolution. TAST is not an official test but a subset of the TOEFL iBT's speaking portion, which is used to evaluate test-takers' ability to communicate orally in English and prepare them for the TOEFL iBT and, more generally, for course work conducted in English.

Xi et al. 2006 described the TAST 2003/180 corpus, a set of responses from 180 test-takers of TAST collected in 2003. Each response was scored holistically by two trained human raters on a scale from 1 (poor proficiency) to 4 (excellent proficiency). Additionally, responses from a subset of the subjects (80) were rated separately for delivery and language use proficiency. The scoring of different dimensions of communicative competence was undertaken to separate different aspects of proficiency. For example, a test-taker might have great delivery skills but poor language use skills, resulting in a mediocre holistic score but assessment ratings that differ on these two dimensions.

The corpus we used for the analyses reported in subsequent chapters was a further subset of the 80 test-takers of TAST 2003/180 corpus that were rated separately on delivery and language use dimensions. We refer to this corpus as DTAST because we were most interested in assessing the spoken delivery proficiency of the test-takers. DTAST comprised 87 responses distributed among 61 test-takers. We used only one task response from the majority of the speakers in our corpus (67.2%), though for some speakers we used two (9.8%) or three (23.0%) task responses. The self-reported native language (L1) of the

test-takers varied considerably. The native languages of the test-takers were distributed among 12 distinct language families, though the most frequent L1s by far were Chinese dialects, which constituted 36% of all L1 speakers in DTAST.

Each task response in the corpus was rated by a trained human judge using the rubrics discussed above. For this analyses, we used only the delivery scores because doing so allowed us to study fluency, intonation, and rhythm proficiency while normalizing for language use and topic development dimensions of communicative competence. A subset of the responses (55) were rated by two raters. On this subset, we measured the rate of inter-rater agreement over chance using the quadratically-weighted Kappa statistic (κ_{qw}, defined in Section **??**). Inter-rater κ_{qw} was found to be 0.72 and the correlation between human ratings was also 0.72. The mean ratings of both raters were 2.29 and 2.92; the standard deviations of both rater were 0.96 and 0.90. In all experiments we conducted that relied on human raters, we used only the ratings from the person who scored all of the DTAST corpus.

Each response in DTAST was hand-segmented into words and syllables. In total, there were found to be 9,013 words and 14,560 syllables. The shortest response was 19.5 seconds and the longest was 60 seconds—the maximum time allowed for an answer in TAST. Average response length was 48.0 seconds with a standard deviation of 10.7 seconds.

5 Annotation

Intonation and rhythm were hand-labeled as sequences of abstract prosodic events aligned to syllables under the Tones and Break Indices (ToBI) framework Beckman et al. (2005). ToBI is a system for transcribing intonation patterns and other aspects of the prosody of English utterances, though it has been adapted for other languages as well. Refer to Chapter **??** for a more thorough description of ToBI phrase accents and boundary tones.

5.1 Intonation

Since intonation is phonological in nature it is by definition a language-specific phenomenon. In other words, there are well-governed rules for the distribution of prosodic events that native speakers adhere to. Knowledge of these rules are utilized by the human labelers under the ToBI framework. In this sense, it is particularly difficult to label non-native speech with an abstract labeling system designed for native productions.

However, it has been shown that inter-labeler agreement is high for intonational phrase boundary tones and we hoped it would be the same when labeling non-native data. [3] Full intonation phrase boundaries were marked when strong perceptual prosodic juncture was detected. [4] The syllable immediately preceding each prosodic juncture was assigned one of the following phrase accent + boundary tone labels: `L-L%`, `L-H%`, `H-L%`, `H-H%`. [5]

Labeling of boundary tones was accomplished through quasi-consensus labeling by two labelers. One labeler annotated each of the files in the corpus. A second, more experienced annotator then checked each annotation and indicated where there was disagreement. The two labelers then conferred and reached a mutually agreed-upon label. We observed a κ value of 0.77 between the annotations before consensus, indicating that disagreements were acceptably rare. It should be noted that a much stricter test would be to measure κ between annotations that were truly independently labeled.

5.2 Rhythm

Syllable stress is not incorporated into the standard set of labels under the ToBI framework. This is due largely to the fact that it can be predicted in English based on the underlying lexical stress of each word and whether it bears a nuclear pitch accent or not (which is labeled in the ToBI framework). However this assumes, again, that speakers adhere to the rhythmic stress rules of English. With non-native speakers we could not assume this, so explicit labeling of syllable prominence was necessary.

Two levels of syllable prominence were labeled independently of phrase accents and boundary tones. Primary stress was marked with a `1` and secondary stress (less prominent than primary stress) was marked with a `2`. Unstressed syllables were left unmarked. In addition to listening to the speech, labelers also had access to visualization of the fundamental frequency, intensity, and waveform of the speech, as they did when labeling phrase accents and boundary tones.

One annotator labeled all of the stressed syllables in the corpus. A second annotated a subset of these (5 test-taker responses totaling 590 syllables). We found κ on the subset of the data labeled by two annotators to be 0.53 for full rhythmic stress and 0.71 when primary and secondary stress were conflated to a single category `stressed`. It has been noted that labeling of secondary (and

[3] Yoon et al. 2004 found inter-rater κ of phrase accent + boundary tones to range from 0.7 to 0.8.

[4] A "strong perceptual prosodic juncture" corresponds to a break index of value 4 in the ToBI framework.

[5] Refer to Section **??** for a description of the ToBI annotation scheme.

	Label	Distribution	
stress	¬stress	11099	(76.2%)
	stress	3451	(23.8%)
tone	¬tone	12976	(89.1%)
	L-L%	878	(6.0%)
	L-H%	377	(2.6%)
	H-L%	243	(1.7%)
	H-H%	86	(0.6%)

Table 3
Distribution of stress (top) and phrase accents + boundary tones (bottom) in the DTAST corpus.

tertiary) stress in spontaneous speech can be quite difficult and unreliable. Some linguists claim it may not even exist in spoken English (Ladefoged, 1993, e.g.,). For these reasons, we decided to adopt a binary annotation scheme for stress; a syllable was considered to be be either stressed (stress) or unstressed (¬stress).

Table 3 shows the label distribution in our corpus for stress and tone at the syllable-level. Most (76.2%) syllables bore no stress and most (89.1%) did not end at a phrase boundary. The most common phrase accent and boundary tone combination was L-L% (6.0%) and L-H% was observed nearly half as often (2.6%) as this tone. The remaining intonation labels—H-L% and H-H%—were observed less often (1.7% and 0.6%, respectively).

6 Automatic Scoring Metrics

In this chapter we describe the features that were extracted based on the intonation and rhythm annotation. In order to determine whether these features were useful for scoring the English delivery skills of non-native speakers of DTAST, we used the Pearson product-moment correlation coefficient (r) to measure the linear relationship between the features and the delivery scores provided by the human raters.

6.1 Fluency

The fluency metrics calculated for the TAST 2003/180 corpus Xi et al. (2006) were recalculated for the DTAST corpus and are described in Table 4 on page 27. A **silence** was defined to be any non-speech segment greater than 150

milliseconds, a **long pause** was identified as a silence greater than 500 milliseconds, and a **chunk** was any segment of speech without internal long pauses. Significant correlations between fluency measurements and human ratings are shown in Table 5 on page 27. All features except `longpstddv`, `silpsec`, and `dpsec` were found to significantly correlate with human ratings. These findings were consistent with other studies of the fluency of non-native speech: proficiency ratings increase as speaking rate (`wpsec`, `wdpchk`, `secpchk`) increases and both the rate and duration of silences and long pauses (`silpwd`, `silmean`, `longmn`) decrease. We also found that proficient speakers tended to be more consistent in their length of silences (`silstddv`) than less proficient speakers (though this was not found to hold for long pauses).

6.2 Intonation

Intonation features were extracted in a way that quantized the distribution of prosodic boundary tones observed in the DTAST corpus. Table 6 on page 28 describes the intonation features under consideration. Recording the present or absence of phrase accent + boundary tone labels (`tonedist`) was the same as recording the distribution of intonational phrases, which we hypothesized to be an important delivery skill; however, this metric may be somewhat redundant with the fluency metrics that take note of pausing information, considering that most intonational phrase boundaries occur before a pause. However, we also hypothesized that metrics that accounted for the distributional properties of specific tones would also be a significant indicator of delivery proficiency.

Table 7 on page 28 shows significant correlations ($p < 0.05$) between human delivery scores and intonation features. We observed that the strength of the correlations were on par with the correlations observed for the fluency features; the average absolute correlation strength with human scores using the fluency features was 0.39, while it was 0.35 when using the intonation features. We also note that the features that described phrase-final rising intonation—`L-H%` and `H-H%`—made up most (62.5%) of the significant correlations. The message here appears to be that judicious use of rising intonation correlated with higher human proficiency scores. More specifically, the greater the distance between `L-H%` and `H-H%` tones (`LHdist` and `HHdist`, respectively) the higher the delivery score. Conversely, the lower the rate of `L-H%` and `H-H%` tones per syllable (`LHrate` and `HHrate`, respectively) the higher the score. It was also found that if there were too many `H-H%` tones relative to other tones (`HH2tone`) then this indicated a low delivery proficiency score. One possible explanation for the observation that frequent phrase-final rising intonation correlated with low delivery proficiency could be pragmatic in nature. Given that student answers to TAST questions should be affirmative and assertive, and that phrase-final rising intonation is thought to convey non-commitment

and/or non-assertiveness in English (cf. Chapter **??**), using rising intonation too often in this context might convey poor proficiency. These findings motivate future exploration of features that measure the relationship of phrase accent and boundary tone to syntactic or pragmatic sentence type, as has also been suggested by Toivanen 2003.

Another pattern emerged from the observed significant correlations as well: the segmentation of a stream of speech into too many intonational phrases (`tonerate`) that were too close together (`tonedist`) indicated lower proficiency scores. Due to the predominance of `L-L%` tones to other tone types in the corpus, we can say that the `LLrate` feature was indicative of this pattern as well, rather than of something inherent to the `L-L%` tone type in particular. In other words, the frequent display of strong prosodic juncture was indicative of low proficiency in spoken Standard Academic English.

6.3 Rhythm

Rhythm features, described in Table 8, were extracted in a way very similar to the intonation features. The two rhythm features recorded the distance between successive stressed syllables (`stressdist`) and the rate of stressed syllables (`stressrate`) in each task response. Table 9 shows significant correlations between human delivery scores and the rhythm features. The observed strength of the correlations here was much lower than that observed for the fluency and intonation features, though we can remark on a pattern similar to one already discussed. To some extent, proficient speakers were distinguished from non-proficient speakers by the fact that the former stressed syllables less frequently than the latter did. This pattern is analogous to the pattern found for phrase accent + boundary tone distribution, as discussed in Section 6.2.

6.4 Redundancy

Above, we established that there were distribution fluency, intonation, and rhythm metrics that significantly correlated with human-assigned delivery proficiency scores of the spoken answers of test-takers in the DTAST corpus. In this section we describe the redundancy of each feature set, as indicated by the significant pairwise inter-correlation of all features, in order to identify the strongest and least redundant features. We hypothesized that many features would encode redundant information.

Table 10 shows all significant inter-correlations ($p < 0.001$) within the fluency feature set. Of the 45 possible pairwise tests, 15 (33.3%) were found to significantly correlate. In other words, the features in the fluency feature set were redundant to some extent and, in fact, the average absolute corre-

lation strength was found to be quite high (M = 0.74, SD = 0.18). Among the strongest correlations were those that measured information about silences and long pauses (e.g., the correlation coefficient for `silstddv` and `longpstddv` was 0.97). This implied here that it was not necessary for fluency assessment to distinguish between silences of different lengths. Another redundancy seemed to exist between measurements based on seconds and those based on linguistic units, such as words (e.g., the correlation coefficient for seconds per chunk (`secpchk`) and words per chunk (`wdpchk`) was 0.89). Interestingly, the rate of disfluencies (`dpsec`) was not found to correlate with any of the other fluency features, indicating that it modeled fluency information that was truly independent of the other fluency features.

It was not surprising that many of the fluency features were found to be somewhat redundant, and it is not presented here as a way to be controversial. If the cost of calculating the features is minimal, which is true for the fluency features, then there is no harm in having redundant features. The analysis is presented as a way to compare subsequent examination of the redundancy of the intonation and rhythm feature sets, which have been far less studied.

We calculated correlations for all pairs of intonation features and noticed first and foremost that there was near perfect correlation between the features that measured specific phrase accent + boundary tone labels per syllable and the features that measured the ratio of specific phrase accent + boundary tone combination to all boundary tones: (`LLrate`, `LL2tone`) = 0.83, (`LHrate`, `LH2tone`) = 0.89, (`HLrate`, `HL2tone`) = 0.94, (`HHrate`, `HH2tone`) = 0.95. This can be explained by the distribution of the phrase accents and boundary tones observed in DTAST. Since the overwhelming majority of phrase accent + boundary tone labels were `L-L%`, both sets of features effectively measured the frequencies of phrase accent + boundary tone labels in a response. Thus, we have excluded `LL2tone`, `LH2tone`, `HL2tone`, `HH2tone` from subsequent discussion.

Of the remaining intonation feature pairs, shown in Table 11, we note that there were the same number of significant correlations as there were with the fluency features (15/45 = 33.3%), though the mean correlation (M = 0.55, SD = 0.18) was quite a bit lower than they were for the fluency features (M = 0.74, SD = 0.18). Taken together, we can state that the intonation feature set had at least some internal degree of redundancy as did the commonly used fluency feature set, though it was possibly more diverse in terms of representing the intonation sub-dimension.

In Table 11 we also note two very strong correlations: (`HHdist`, `HLdist`) = -0.97 and (`tonerate`, `tonedist`) = -0.95. The latter makes sense if we consider both the observed boundary tone rate and average distance between tones to be relatively stable. The former correlation is more perplexing, though. It

is not entirely clear why the distance between H-H% and H-L% labels would correlate so strongly, though it was likely due to the sparsity of the tones in the corpus. We required there to be at least two of the same tone type in a response in order to calculate these measures and in all likelihood this happened rather infrequently for these tones, so this potentially skewed our calculations of these measures. This reasoning is further supported by the fact that we did not see similarly strong correlations for L-L% or L-H%, of which we had more instances.

The remaining correlations tended to mirror the correlation we saw between tonerate and tonedist to a much lesser degree, but for the same reason. For example, the correlation between LLrate and LLdist (-0.72) indicated that L-L% labels were distributed at regular intervals in the data. However, the fact that the correlation strength was weaker than the correlation strength between tonerate and tonedist implies that recording the distributional properties of individual boundary tones was justified because their behavior was less predictable and, hopefully, more indicative of speaker proficiency.

The rhythm features (stressrate and stressdist) were highly correlated with one another (r = -0.97) indicating that they were redundant. However, when examining the extent to which each feature set was redundant with one another, we found that no rhythm feature significantly correlated with any other feature in the fluency or intonation feature sets. In other words, the rhythm features can be considered to have modeled delivery information independently of both fluency and intonation.

There were 140 correlations we might have observed between the fluency measurements and the intonation measurements (10 fluency features · 14 intonation features). We observed only 16 significant correlations, though, which are shown in Table 12. The fact that the overwhelming majority of correlations were *not* significant suggested that there was some general degree of independence between the fluency measurements and those we developed to measure intonation. This notwithstanding, speaking rate—as defined as the number of words per second (wpsec)—correlated with five of the intonation features. Furthermore, information about general boundary tone distribution (tonedist and tonerate) correlated with five fluency measurements as well. [6] Taken together, and noting that speaking rate is known to be one of the best correlates of fluency, this would seem to imply, as we hypothesized, that some intonation information need not be separately modeled from fluency when correlating with human raters. A closer look at the correlations of specific phrase accent + boundary tone labels seems to indicate otherwise, though. We note that wpsec negatively correlated with tonerate and LHrate, but positively

[6] Since tonerate and tonedist were highly inter-correlated (r = -0.95) the number of correlations here is less dramatic than would appear.

correlated with HLrate. Forgoing intonation measurements might run the risk of discarding important distributional aspects of boundary tones. Also, most correlations, though significant, were relatively weak (from 0.3 to 0.4) considering that here we were using correlation to determine whether feature sets were redundant or not. Clearly, intonation and fluency features are redundant *to some extent*, but not so dramatically that we can say that they model the same information.

The strongest correlations, by far, concerned the average number of syllables between successive high rising boundary tones (HHdist). It is very intriguing to note that the distance between H-H% boundary tones positively correlated with the rate of filled pauses (dpsec) and negatively correlated with the rate of silences (silpwd). However, upon closer examination, we discovered that there were only 12 task responses that had more than one H-H% label in the DTAST corpus; in other words, in most of our data HHdist was actually undefined. With such a small sample it would be unwise to conjecture about the relationships concerning H-H% except to keep it in mind as a future avenue of exploration.

To summarize, many of the metrics we used to evaluate speaking proficiency encoded redundant information. Foremost among such redundancy was discrimination between the mean distance between phrase accent + boundary tone labels and the rate of these labels. Also redundant was recording the relationship between different phrase accent + boundary tone labels to one another. However, there were few significant correlations found between the fluency, intonation, and rhythm features sets, strengthening our belief that each feature set was representative of complementary aspects of delivery proficiency.

7 Automatic Estimation of Human Scores

As has been shown by other studies for fluency, we demonstrated in preceding chapters that measurements of intonation and rhythm correlated with human ratings of the delivery of non-native spoken English. Furthermore, analysis indicated that the feature sets postulated as reflecting the intonation and rhythm sub-dimensions of delivery proficiency were independent from one another as well as from the commonly-proposed fluency metrics. Though these findings may shed light onto the sometimes intangible human rating process, the larger goal of our research was to estimate with high accuracy and reliability the human rating scores themselves so that the assessment of the communicative competence of non-native English speakers can be done automatically. This section describes experiments conducted to assess the degree to which human rating scores could be automatically assigned.

We chose to use the support vector regression method Smola and Schoelkopf (1998) as implemented in the WEKA software package because we wanted to classify human ratings as a continuous variable ranging from 1 to 4. We used a quadratic kernel and ran 5-fold cross validation. We report average performance on the five sets of held-out data in terms of both correlation coefficient (r) and quadratically-weighted Kappa (κ_{qw}) after rounding the predictions. Table 13 lists the performance of support vector regression using all combinations of feature sets. Our gold standard was the agreement of human raters (r = 0.72, κ_{qw} = 0.72).

We noticed somewhat different performance depending on whether correlation or κ_{qw} was used as the measurement of agreement between the automatically estimated ratings and those assigned by the human rater, though a global trend did emerge. When each feature set was considered in isolation, rhythm performed much worse (r = 0.26, κ_{qw} = 0.20) that both fluency and intonation. When considering κ_{qw}, fluency and intonation feature sets were comparable to one another (κ_{qw} = 0.51 and κ_{qw} = 0.52, respectively). However, when considering correlation, performance of the fluency feature set was substantially better than than performance of the intonation feature set (r = 0.60 and r = 0.50, respectively).

Both performance measures showed an increase in human rating estimation as feature sets were combined, though the exact nature of these improvements differed slightly. For all pair-wise feature set combinations, the sets containing the fluency features performed the best, and equally as well for both correlation (fluency+intonation = 0.69, fluency+rhythm = 0.70) and κ_{qw} (fluency+intonation = 0.67, fluency+rhythm = 0.67). The performances of the paired fluency+intonation and fluency+rhythm feature sets were substantially higher than intonation+rhythm (r = 0.65, κ_{qw} = 0.64) which, in turn, was higher than when using each feature set in isolation. Finally, The combination of all feature sets (fluency+intonation+rhythm) was shown to perform the best when measured by κ_{qw} (0.74), though was not shown to be better than other feature sets when measured by correlation.

Despite the observed differences between correlation and κ_{qw} performance, the findings supported the following general assessment. We observed estimation of human delivery ratings at or above what we had observed for human agreement on the task. The fact that this was corroborated by two performance measures was reassuring because it signified that reaching human performance levels on the assessment of the delivery dimension of spoken English is indeed obtainable. Furthermore, we observed that the addition of either intonation or rhythm features increased performance over the use of fluency features alone. This supported our earlier hypothesis that using intonation and rhythm metrics is an important (in fact, essential) part of spoken language assessment. It remains to be seen how critical the rhythm features are. On their own, they

were quite poor predictors of human ratings, though in combination with the other features they were considered useful. These findings warrant further investigation into the role of rhythm in spoken language proficiency.

A final caveat is in order. Unlike classification tasks in some other domains, when machine learning is used in the context of language assessment, care must be taken to ensure that the automatically-learned models reflect a pedagogical rubric, in this case the construct of communicative competence shown in Table 1. For example, before our models could be used to assess the delivery proficiency of non-native English speakers using TAST or TOEFL-iBT, we would need to verify that no single feature dominated the prediction algorithm, and that features were not used to drive scores in a direction opposite of what we would expect. Despite this cautionary message, we are optimistic that intonational and rhythmic assessment are critical for this task, and we feel that our experimental results have supported this claim. Future research would be to validate the models given these considerations.

8 Automatic Detection of Prosodic Events

As the motivating force behind our research was to automate proficiency scoring of non-native English speakers, we must ultimately automate every part of the process. In the previous chapter we demonstrated how human ratings might be estimated using a machine learning approach utilizing fluency, intonation, and rhythm features. However, these features themselves must be automatically derived from the speech stream. Work on automatically calculating fluency features is quite robust. Most of the information necessary for these features can be obtained from state-of-the-art continuous speech recognizers (Franco et al., 2000; Cucchiarini et al., 2002; Xi et al., 2006; Zechner and Bejar, 2006, inter alia). For the automatic detection of prosodic boundary tones and stressed syllables we can also utilize speech recognition technology to segment and transcribe the data. However, one must go beyond the capabilities of this technology to fully model prosodic events.

The general framework we chose to use for automatic detection of prosodic events was the binary decision tree, a type of machine classifier especially good at predicting categorical data, such as boundary tones and stress.[7] Decision trees allow for non-linear learning and the ability to easily understand the types of decisions made in the learning process. For the experiments we describe here, some of our features relied on hand-labeled information; specifically, orthographic transcription and, as well as word and syllable alignments. Thus, our prosodic prediction results should be taken as upper bounds as a

[7] We used the Weka J4.8 implementation of C4.5 decision tress.

fully implemented automatic proficiency scorer must use predictions for this information as well.

8.1 Feature extraction

From each DTAST response, features were extracted based on several known cues to prosodic events. These features included acoustic, segmental, lexical, and syntactic information and are shown in Table 14 on page 32. The basic unit of analysis was the syllable, but many features comprised information about the word containing the syllable in question. For most features, information from surrounding syllables was taken into account as well, extending up to five syllables in the future and the past.

The largest group of features encapsulated acoustic information automatically extracted from the speech signal. As a known correlate of pitch, fundamental frequency (f0) was extracted using the speech analysis software Praat Boersma (2001) and the minimum (`f0-min-`σ), mean (`f0-mean-`σ), maximum (`f0-max-`σ), standard deviation (`f0-stdv-`σ), and range (`f0-range-`σ) were calculated for each syllable. Additionally, two approximations of pitch contour shape were calculated using linear regression of the pitch values against time. The first calculated a straight line through the data and the feature value recorded was the slope of this line (`f0-rslope-`σ). The second calculated a second-degree polynomial through the data and the feature value recorded was the coefficient of the quadratic term (`f0-curve-`σ). Finally, the ratio of voiced frames to unvoiced frames was calculated as an approximation of speaking rate (`f0-voiced-`σ). [8]

A second set of acoustic features were computed based on the intensity (measured in decibels) of the sound waveform; again, as estimated by Praat. This is known to be a correlate of perceptual loudness of speech. The statistics recorded for intensity were the mean (`db-mean-`σ), minimum (`db-min-`σ), and maximum (`db-max-`σ) over each syllable.

Segmental information of each syllable and word containing that syllable was also recorded. For syllables, this was the the phonemic transcription of the vocalic nucleus, including lexical stress, as obtained from the CMU Pronouncing Dictionary based on the hand-labeled word transcriptions of each task response (`trans-`σ). [9] Note that orthographic transcriptions included demarcation of silences and filled pauses. Both the syllable and word durations, in

[8] Whether a frame was voiced or not was determined by the pitch detection algorithm in Praat.

[9] The CMU Pronuncing Dictionary can be accessed here: `http://www.speech.cs.cmu.edu/cgi-bin/cmudict`

milliseconds, were noted as well (ms-σ and ms-w, respectively). Other features in this set recorded the relationship between the syllable and word that bore the syllable: the number of syllables in the word (count-σ2w), the ratio of the duration of the syllable to the duration of the word (ms-σ2w), and the position of the syllable in the word (loc-σ2w).

The part of speech of each word, as automatically determined by a part of speech tagger Ratnaparkhi (1996) trained on Switchboard transcripts, was recorded to encapsulate syntactic information. Assigned to each syllable was the predicted part of speech for the word containing that syllable (pos-w).

Included as additional features were the corresponding values of all aforementioned features for of each of the five preceding and following syllables. For example, db-mean-σ_1 recorded the intensity of the syllable following the syllable under consideration. In this way, we hoped to contextualize information of the current syllable by those surrounding it.

A final set of features recorded information about stress and tone labels occurring previous to the syllable in question. In particular, these features measured the number of syllables and words since the observations of the last stress and tone boundaries (span-stress and span-tone). Also, the stress and phrase accent + boundary tone labels (as well as simply the presence or absence of each) were recorded for each of the previous five syllables (pstress and ptone). For example, pstress$_{-1}$ recorded whether the previous syllable bore stress or not.

Given the features enumerated in Table 14 and the features derived from exploiting a 5-syllable contextual window, there were 200 features associated with each syllable. Real-valued features were further contextualized by z-score normalizing per task response.

8.2 Performance

We trained separate classifiers for stress and tone detection. Performance of each classifier is reported as the average of 5-fold cross-validation and is shown in Table 15. Expectedly, classification of the absence of stress ($\neg$stress) and/or tone ($\neg$tone) was quite reliable, in part because of the high frequency of these labels. Though not as good, the detection of stressed syllables (stress) was quite robust (F = .72). Tone detection did not perform as well. L-L% was the only boundary tone that was predicted reasonably well (F = 0.51). L-H% prediction had an F-measure of only 0.21 and both H-L% and H-H% were essentially not modeled with any degree of accuracy.

An examination of the induced decision trees provided some insight into the

usefulness of our features. For stress detection, all features types were represented but varied with respect to which of the surrounding syllables were most useful. Intensity features of the current and surrounding syllables were heavily used, as was information about past stress and tone labels. Somewhat surprisingly, pitch information of surrounding syllables was never used. The duration of current and surrounding syllables was deemed important, as were the transcriptions of the current (but not surrounding) words and syllables. Finally, the part of speech of the current and following words were used often, as were the number of syllables in the current and five preceding words.

The features used for boundary tone detection were much more restricted. Little contextual information was used beyond the tone and stress labels of preceding syllables. In a more limited capacity, other useful features were the intensity and pitch measurements of the syllable.

Despite the fact that we did not achieve perfect performance when automatically detecting prosodic events, we still considered it important to observe whether the automatic intonation and rhythm scoring metrics of Chapter 6, based on the prosodic event predictions that were made, correlated with human delivery scores. Table 16 lists the significant correlations of predicted rhythm and intonation features with human delivery ratings. Also shown (reprinted from Tables 7 and 9 from Chapter 6) are significant correlations using hand-labeled intonation and rhythm and intonation features, for comparison.

It is somewhat surprising to have found that the strength of the observed correlations did not differ as much as might be expected given the imperfect prediction of prosodic events. The mean absolute correlation strength using predicted prosody was 0.26 (SD = 0.05) and was 0.33 (SD = 0.11) when considering hand-label prosody. Though, as the standard deviations indicate, the correlations for predicted prosody were all relatively the same, whereas when using the hand-labeled prosody, there was more variation. In other words, it seems that we lost some of the stronger correlations when predicting prosodic events. We will examine the possible impact of this shortly.

Though there was less variation with respect to the correlations based on predicted prosodic events, the relative correlation strength was nevertheless maintained. The `tonedist` and `tonerate` features correlated most strongly with the human ratings; `stressdist` and `stressdist` showed the weakest correlations. It should also be noted that the significance factor (p) for the correlations using predicted prosody were slightly higher, on average, than they were for the correlations using hand-labeled prosody. This indicates that the correlations themselves were slightly less significant.

In the hand-labeled scenario, correlations involving boundary tones other than

L-L% were found to correlate with human ratings, whereas this was largely not the case with the prediction scenario. In fact, HHrate and HH2tone, based on predicted prosody, were the only metrics that referred to specific tone labels and that were found to have a significant correlation coefficient. Additionally, L-L% boundary tones were represented more often in the prediction scenario than they were in the hand-labeled scenario. These two findings were most likely due to the fact that boundary tones other than L-L% were not reliably predicted in the prosodic event classification task. This was probably the most egregious shortcoming of our prosodic event detection, considering our earlier suggestion that non-L-L% boundary tones might be an essential aspect of proficiency assessment of non-native English speakers.

In keeping with our previous experiments using hand-labeled prosodic events, we conducted machine learning experiments to automatically predict human delivery ratings given predicted prosodic events. We ran 5-fold cross-validation using support vector regression for all combinations of feature sets. The results of these experiments can be seen in Table 17. Also shown are the results of the same experiments using delivery features derived from hand-labeled prosody, for comparison (reprinted from Table 13). Fluency feature values were the same across scenarios because we used hand-labeled word segmentation for all experiments.

Human score estimation using intonation and rhythm features (in isolation) in the predicted scenario suffered greatly compared with the same features in the hand-labeled scenario. Score estimation, as measured by κ_{qw} between estimated and actual human ratings, dropped from 0.20 to 0.09 when using the rhythm features and from 0.52 to 0.30 when using intonation features. Combining the two feature sets (intonation+fluency) increased κ_{qw} to 0.46, but this was still well below the 0.64 value we saw when using hand-labeled prosodic events. Reassuringly, we note that score estimation when using all features combined (fluency+intonation+rhythm) was better than when using only the fluency features. In other words, though predicting prosodic events degraded our ability to predict human delivery ratings, intonation and rhythm features still contributed to the improvement of human score estimation to a level almost comparable to the agreement we observed between the two human raters on the task.

9 Discussion

Intonation, rhythm, fluency, and pronunciation are believed to be independent factors in the delivery dimension as set forth by the construct of communicative competence. While there has been substantial empirical research concerning fluency and pronunciation of non-native English speakers within

the assessment community, little research has been conducted with respect to intonation and rhythm. Our work has provided valuable insight into these dimensions by validating their usefulness and independence in the automatic scoring of delivery proficiency. In the DTAST corpus, we showed that these phenomena, for the most part, modeled distinct information from that encoded by well-known fluency correlates, and yet correlated equally strongly with human scores. Furthermore, though fluency measurements were somewhat better at automatically estimating human scores, using a combination of all delivery features together approached the agreement of human raters on the task. What remains, and is planned for future work, is to validate that automatically-learned assessment models use features in a way consistent with the human scoring rubric.

It was also shown that prediction of prosodic events in non-native speech can be achieved to some extent using some automatically derived information, such as pitch and energy. Furthermore, it was reassuring to observe that significant correlations of delivery features derived from these predicted prosodic events still significantly correlated with human ratings in a similar (albeit weaker) manner. However, low prediction accuracy of some prosodic events, especially boundary tones, indicated that there is still more work to be done in this area. A decreased ability to predict human rating scores also supported this position. One solution may be to model boundary tones independently or use graphical models that find the best sequence of tones given acoustic and segmental information. Also, with respect to segmentation, our feature set relied heavily on hand-labeled segmentation and transcription. It remains future work to evaluate our method using speech recognition technology that automatically predicts this information.

We should strive in particular to predict phrase-final rising intonation more accurately, as it appeared that distributional features quantifying these tones dominated the correlations with human raters. We conjecture that these correlations are due to the inadvertent signaling of non-commitment or non-assertiveness by test-takers through the use of rising intonation at inappropriate times. One of our future goals is to encode such information into the delivery features by recording the rate of tones given pragmatic and syntactic sentence types. We also believe that a similar technique might yield more informative rhythmic features as well by recording the rate of different parts of speech that are stressed (e.g., function words should not generally be stressed).

There are two distinctly different directions one could take with this research, each motivated by a different goal. The first direction would be to continue the line of research we have set forth in this paper. Though our focus was rating the communicative competence of non-native speakers of English, we were also very interested in modeling behavior that could be used for diagnostic purposes in language-learning software. The delivery features we have

described here could be of use to a foreign language learner by informing them of general tendencies that lead to low proficiency perception. For example, students could be informed that they are pausing too often or stressing too many words. However, our approach of modeling prosodic events actually allows us to be even more specific by, ideally, pinpointing areas of poor proficiency. For example, a student might be told that the use of a particular boundary tone on a specific utterance is not what a proficient speaker might do. Of course, before reaching this point one would need to ensure reliable prediction accuracy of prosodic events, and this remains a major challenge.

The second direction would forgo diagnosis and strive only to increase scoring accuracy. There may be a more direct way of measuring intonation and rhythm that does not rely on segmentation or transcription of any kind, thus decreasing potential sources of noise. Instead of modeling discrete prosodic events, one might directly model the intensity and fundamental frequency of speech in a fashion similar to acoustic modeling for pronunciation evaluation. Though the output of such models might be more predictive of human ratings, they would most likely be of little use to the language-learner.

References

Bachman, L., 1990. Fundamental Considerations in Language Testing. Oxford U. Press.

Beckman, M. E., Hirschberg, J., Shattuck-Hufnagel, S., 2005. The original ToBI system and the evolution of the ToBI framework. In: Jun, S. A. (Ed.), Prosodic Typology – The Phonology of Intonation and Phrasing. Oxford, OUP.

Bernstein, J., Cohen, M., Murveit, H., Rtischev, D., Weintraub, M., November 1990. Automatic evaluation and training in english pronunciation. In: Proceedings of ICSLP. Kobe, Japan.

Bernstein, J., DeJong, J., Pisoni, D., Townshend, B., 2000. Two experiments in automatic scoring of spoken language proficiency. In: Proceedings of InSTIL 2000. Dundee, Scotland.

Boersma, P., 2001. Praat, a system for doing phonetics by computer. Glot International 5 (9/10), 341–345.
URL `http://www.praat.org`

Canale, M., Swain, M., 1980. Theoretical bases of communicative approaches to second language teaching and testing. Applied Linguistics 1 (1), 1–47.

Cucchiarini, C., Strik, H., Boves, L., 1997. Using speech recognition technology to assess foreign speakers pronunciation of Dutch. In: Proceedings of New Sounds. Klagenfurt, Austria, pp. 61–68.

Cucchiarini, C., Strik, H., Boves, L., 2002. Quantitative assessment of sec-

ond language learners' fluency: Comparisons between read and spontaneous speech. Journal of the Acoustical Society of America 111 (6), 2862–2873.

Franco, H., Abrash, V., Precoda, K., Bratt, H., Rao, R., Butzberger, J., Rossier, R., Cesari, F., 2000. The SRI Eduspeak system: Recognition and pronunciation scoring for language learning. In: Proceedings of InSTIL 2000. Dundee, Scotland.

Franco, H., Neumeyer, L., Kim, Y., Ronen, O., 1997. Automatic pronunciation scoring for language instruction. In: Proceedings of ICASSP. Los Alamitos, CA, USA.

Gass, S., Selinker, L., 1994. Second Language Acquisition: An Introductory Course. Publications of the Department of Phonetics, University of Turku.

Herry, N., Hirst, D., April 2002. Subjective and objective evaluation of the prosody of English spoken by French speakers: The contribution of computer assisted learning. In: Proceedings of Speech Prosody. Aix-en-Provence, France.

Hincks, R., 2005. Computer support for learners of spoken English. Ph.D. thesis, Speech and Music Communication, The Royal Institute of Technology in Sweden (KTH).

Hymes, D. H., 1971. On Communicative Competence. Philadelphia: University of Pennsylvania Press.

Kormos, J., Dénes, M., 2004. Exploring measures and perceptions of fluency in the speech of second language learners. System 32 (2), 145–164.

Ladefoged, P., 1993. A Course in Phonetics, third edition Edition. Harcourt Brace Jovanovich, New York.

Morgan, J., Gallagher, M., Molholt, G., Holland, M., LaRocca, S., 2004. Towards the automatic evaluation of fluency. In: The 12th Annual Army Research Laboratory/United States Military Academy Technical Symposium. West Point, New York, USA.

Neumeyer, L., Franco, H., Weintraub, M., Price, P., 1996. Automatic text-independent pronunciation scoring of foreign language student speech. In: Proceedings of ICSLP. Philadelphia, PA, USA, pp. 1457–1460.

Ratnaparkhi, A., May 1996. A maximum entropy model for part-of-speech tagging. In: Proceedings of the Conference on Empirical Methods in Natural Language Processing. Somerset, New Jersey, pp. 133–142.

Smola, A. J., Schoelkopf, B., 1998. A tutorial on support vector regression. In: In NeuroCOLT2 Technical Report Series.

Teixeira, C., Franco, H., Shriberg, E., Precoda, K., Soenmez, K., 2001. Evaluation of speaker's degree of nativeness using text-independent prosodic features. In: Proceedings of the Workshop on Multilingual Speech and Language Processing. Aalborg, Denmark.

Toivanen, J., 2003. Tone choice in the English intonation of proficient non-native speakers. In: Proceedings of Fonetik. Umeå/Lövånger, Sweden, pp. 165–168.

Toivanen, J., 2005. Pitch dynamism of English produced by proficient non-native speakers: Preliminary results of a corpus-based analysis of second

language speech. In: Proceedings of FONETIK. Stockholm, Sweden.

Xi, X., Zechner, K., Bejar, I., 2006. Extracting meaningful speech features to support diagnostic feedback: An ECD approach to automated scoring. In: Proceedings of NCME. San Francisco, CA, USA.

Yoon, T. J., Chavarría, S., Cole, J., Hasegawa-Johnson, M., October 2004. Intertranscriber reliability of prosodic labeling on telephone conversation using ToBI. In: Proceeding of the International Conference on Spoken Language Processing – Interspeech. Jeju, Korea.

Zechner, K., Bejar, I., June 2006. Towards automatic scoring of non-native spontaneous speech. In: Proceedings of the Human Language Technology Conference of the NAACL. Association for Computational Linguistics, New York City, USA, pp. 216–223.
URL http://www.aclweb.org/anthology/N/N06/N06-1028

Variable	Description
silmean	Mean silence duration.
silstddv	Standard deviation of silence duration.
silpsec	Silences per second.
longpmn	Mean long pause duration.
longpstddv	Standard deviation of long pause duration.
silpwd	Number of silences per word.
wpsec	Words per second.
wdpchk	Words per chunk.
secpchk	Seconds per chunk.
dpsec	Disfluencies per second.

Table 4
Descriptions of the fluency features computed for DTAST, as proposed in Xi et al. 2006.

Variable	r	p
silpwd	-0.54	0.000
wpsec	0.51	0.000
wdpchk	0.48	0.000
silmean	-0.40	0.008
secpchk	0.33	0.000
longpmn	-0.25	0.008
silstddv	-0.25	0.009

Table 5
Significant correlations between delivery scores and fluency measurements.

Variable	Description
LLdist	Average number of syllables between successive L-L% tones.
LHdist	Average number of syllables between successive L-H% tones.
HLdist	Average number of syllables between successive H-L% tones.
HHdist	Average number of syllables between successive H-H% tones.
tonedist	Average number of syllables between any successive tones.
LLrate	Number of L-L% tones per syllable.
LHrate	Number of L-H% tones per syllable.
HLrate	Number of H-L% tones per syllable.
HHrate	Number of H-H% tones per syllable.
tonerate	Number of any and all tones per syllable.
LL2tone	Number of L-L% tones / number of all tones.
LH2tone	Number of L-H% tones / number of all tones.
HL2tone	Number of H-L% tones / number of all tones.
HH2tone	Number of H-H% tones / number of all tones.

Table 6
Descriptions of the intonation features computed for DTAST.

Variable	r	p
HHdist	0.53	0.025
LHdist	0.47	0.000
tonerate	-0.46	0.000
tonedist	0.38	0.000
HHrate	-0.29	0.002
HH2tone	-0.24	0.012
LHrate	-0.22	0.024
LLrate	-0.22	0.024

Table 7
Significant correlations between delivery scores and intonation features.

Variable	Description
stressdist	Average number of syllables between successive stressed syllables.
stressrate	Number of stressed syllables / number of syllables.

Table 8
Descriptions of the rhythm features computed for DTAST.

Variable	r	p
stressrate	-0.26	0.007
stressdist	0.23	0.017

Table 9
Significant correlations between delivery scores and rhythm features.

	silmean	silstddv	silpsec	longpmn	longpstddv	silpwd	wpsec	wdpchk	secpchk
silstddv	0.85								
silpsec									
longpmn	0.81	0.96							
longpstddv	0.71	0.97		0.95					
silpwd	0.33		0.70						
wpsec			-0.54			-0.47			
wdpchk						-0.80	0.66		
secpchk			-0.73			-0.78		0.89	
dpsec									

Table 10
Significant ($p < 0.001$) correlations of fluency features.

	LLdist	**LHdist**	**HLdist**	**HHdist**	**tonedist**	**LLrate**	**LHrate**	**HLrate**	**HHrate**
LHdist									
HLdist									
HHdist			-0.97						
tonedist	0.52	0.43							
LLrate	-0.72				-0.67				
LHrate		-0.57			-0.44				
HLrate						-0.44	-0.33		
HHrate	0.33								
tonerate	-0.46	-0.44			-0.95	0.63	0.51		

Table 11
Significant ($p < 0.001$) correlations of intonation features.

	dpsec	longpmn	silpwd	silstddv	wdpchk	wpsec
HHdist	0.85		-0.73			
HHrate				0.34		
HL2tone					0.36	0.49
HLrate						0.39
LHrate						-0.36
LLrate			0.31			
tonedist	0.39		-0.45			0.40
tonerate	-0.38	0.37	0.52	0.36		-0.40

Table 12
Significant correlations ($p < 0.001$) between intonation and fluency measurements.

Task	**r**	κ_{qw}
human agreement	0.72	0.72
fluency+intonation+rhythm	0.69	0.74
fluency+intonation	0.69	0.67
fluency+rhythm	0.70	0.67
intonation+rhythm	0.65	0.64
fluency	0.60	0.51
intonation	0.50	0.52
rhythm	0.26	0.20

Table 13
Performance measured as correlation and κ_{qw} of estimated and human delivery ratings.

	Feature	Description
Acoustic	f0-min-σ	minimum f0 of syllable
	f0-max-σ	maximum f0 of syllable
	f0-mean-σ	mean f0 of syllable
	f0-stdv-σ	standard deviation of f0 of syllable
	f0-range-σ	range of f0 of syllable
	f0-rslope-σ	slope of regression line through all f0 points
	f0-curve-σ	coefficient of quadratic regression curve of f0 against time
	f0-voiced-σ	percent of voiced frames in syllable
	db-min-σ	minimum intensity of syllable
	db-max-σ	maximum intensity of syllable
	db-mean-σ	mean intensity of syllable
Segmental	trans-σ	transcription of syllable nucleus
	ms-σ	duration of syllable
	ms-w	duration of word bearing syllable
	ms-σ2w	ratio of syllable duration to word duration
	count-σ2w	number of syllables in word
	loc-σ2w	location of syllable in word: {initial, medial, final, entire}
	pos-w	part of speech of word bearing syllable
Prosodic	pstress	stress label of previous syllables
	ptone	phrase accent + boundary tone of previous syllables
	span-stress	number of syllables between current σ and last with stress
	span-tone	number of syllables between current σ and last with tone

Table 14
Features extracted from each syllable in the DTAST corpus.

Label	Precision	Recall	F-measure
¬stress	0.91	0.92	0.91
stress	0.73	0.71	0.72
¬tone	0.95	0.99	0.97
L-L%	0.54	0.49	0.51
L-H%	0.33	0.16	0.21
H-L%	0.08	0.01	0.02
H-H%	0.06	0.01	0.02

Table 15
Precision, recall, and F-measure for each prosodic label in the corpus.

		Predicted		Hand-labeled	
	Feature	r	p	r	p
Intonation	tonedist	0.34	0.000	0.38	0.000
	tonerate	-0.31	0.001	-0.46	0.000
	HH2tone	-0.29	0.002	-0.24	0.012
	HHrate	-0.29	0.002	-0.29	0.002
	LLrate	-0.20	0.034	-0.22	0.024
	LL2tone	0.27	0.005	—	—
	LLdist	0.23	0.016	—	—
	HHdist	—	—	0.53	0.025
	LHdist	—	—	0.47	0.000
	LHrate	—	—	-0.22	0.024
Stress	stressrate	-0.19	0.050	-0.26	0.007
	stressdist	0.20	0.038	0.23	0.017

Table 16
Comparison of delivery correlations between predicted and hand-labeled intonation/rhythm correlations.

	Predicted		Hand-labeled	
Task	**r**	κ_{qw}	**r**	κ_{qw}
human agreement	0.72	0.72	0.72	0.72
fluency+intonation+rhythm	0.69	0.74	0.66	0.66
intonation+rhythm	0.65	0.64	0.48	0.46
fluency	0.60	0.51	0.60	0.51
intonation	0.50	0.52	0.28	0.30
rhythm	0.26	0.20	0.09	0.09

Table 17
Performance of predicted and human delivery ratings using predicted and hand-labeled prosody labels.