Prosody in the Production and Processing of L2 Spoken Language and Implications for Assessment

Christos Theodoropulos[1]
Teachers College, Columbia University

ABSTRACT

This article offers an extended definition of prosody by examining the role of prosody in the production and comprehension of L2 spoken language from mainly a cognitive, interactionalist perspective. It reviews theoretical and empirical L1 and L2 research that have examined and explored the relationships between prosodic forms and their functions at the pragmatic and discourse levels. Finally, the importance and the various dimensions of these relationships in the context of L2 assessment are presented and the implications for L2 testing employing new technologies and automated scoring systems are discussed.

INTRODUCTION

More than two and a half decades ago, Major (1987), in one of the few articles in *Language Testing* to focus solely on the assessment of speaking in terms of phonological control, wrote that the "measurement of pronunciation accuracy [was] in the dark ages when compared to measurement of other areas of competence" (p. 155). Recognizing the potentially facilitative role of computer-assisted sound-analysis applications and software—a then relatively new industry—Major called for more investigative efforts into objective, reliable, valid, and viable measurements of second language (L2) pronunciation. Since then, only a relatively few noteworthy investigations have answered his call (e.g., Buck, 1989; Gorsuch, 2001; Isaacs, 2008; Koren, 1995; van Weeren & Theunissen, 1987; Yoshida, 2004).

While L2 pedagogy has been reexamined and reconceptualized to include a focus, and an emphasis, on both local (i.e., segmental) and global (i.e., suprasegmental or prosodic) elements of pronunciation (Celce-Murcia, Brinton, & Goodwin, 1996; Derwing, Munro, & Wiebe, 1998; Morley, 1991; Pennington & Richards, 1986), and while there has been a considerable amount of research in the past decade into the nature of L2 phonetic and phonological knowledge as it relates to intelligibility and comprehensibility (e.g., Derwing & Munro, 2005; Jenkins, 2000, 2002, 2005; Kennedy & Trofimovich, 2008; Meyer & Morse, 2003; Munro & Derwing, 2001; Munro, Derwing, & Morton, 2006; Rajadurai, 2007; Zielinski, 2006, 2008), the role of pronunciation has been underrepresented in popular task-oriented guides and language standards and assessments such as the Common European Framework of Reference (Council of Europe,

[1] Christos Theodoropulos is a doctoral student in the Applied Linguistics program. His research interests include formative assessment and program evaluation. He is also the Director of the Assessment Center at the Community College of Philadelphia (CCP), where he oversees all of CCP's placement testing. He can be reached at ct2298@tc.columbia.edu.

2001; van Maele, 2009). Although, more recently, there has been a few studies that examined pronunciation from mainly a listener and "accentedness" (i.e., auditory effect) perspective (see Jamieson & Poonpon, 2013; Kang, Rubin, & Pickering, 2010; Ockey & French, 2014), pronunciation, for the most part, has been overlooked in high-stakes, large-scale L2 speaking tests and scoring rubrics (e.g., see ACTFL OPI, 1999, or the iBT TOEFL Speaking Section, 2005).

Overall, there has been a general consensus among both L1 and L2 researchers that the structure of speech prosody—an important aspect of pronunciation that generally comprises the sound duration, amplitude, and sequence of frequencies of an utterance—is an intrinsic determinant of the form and meaning of spoken language in terms of comprehensibility (see Anderson-Hsieh, Johnson, & Koehler, 1992; Cutler, Dahan, & Donselaar, 1997; Dauer, 2005; Derwing & Munro, 1997; Derwing & Rossiter, 2003; Munro & Derwing, 1995); however, prosody and how best to measure it have not been fully explored in any systematic way by L2 testing researchers.

Today, with the advent of automatic speech recognition (ASR) and scoring technologies, Major's (1987) call appears to be as relevant as it was twenty-three years ago. With the structural design of ASR systems being now fairly standardized (Elhilali, Taishih, & Shamma, 2003; Hakkinen, Suontausta, Riis, & Jensen, 2003; Janse, Nooteboom, & Quene, 2003; Richardson, Bilmes, & Diorio, 2003), many teaching and formative assessment software, such as *Connected Speech*, *Streaming Speech*, and Carnegie Speech Company's *NativeAccent*®, utilize ASR-based acoustic models to provide pronunciation feedback to students (see Cauldwell, 2002; NativeAccent®, 2010; Westwood & Kaufmann, 2001), supplementing, to a certain extent, classroom instructional settings as well as practices. Moreover, a few large-scale testing programs and speaking tests, such as the Educational Testing Service's *TOEFL® Practice Online*, the *Pearson Test of English*, and Stanford Research Institute's (SRI) *EduSpeak™*, employ ASR-based systems to score and, consequently, make claims about examinees' L2 spoken language (see Bernstein, de Jong, Pisoni, & Townshend, 2000; Downey, Farhady, Present-Thomas, Suzuki, & Moere, 2008; Franco, Abrash, Precoda, Bratt, Rao, Butzberger, Rossier, & Cesari, 2000; *Pearson*, 2008; Xi, Higgins, Zechner, & Williamson, 2008). Each of these teaching or testing programs comes with a specific and, quite often, implicit set of assumptions about the nature of L2 speaking and, in particular, the nature of phonetic and phonological knowledge. It appears that over the next decade, one of the formidable challenges for applied linguists will be to investigate the validity of instructional designs underlying many of these computerized teaching programs (Chapelle, 1998a) and to evaluate the claims and assumptions underlying the inferences and interpretative arguments made by testers using these new automated scoring technologies (Chapelle, 2010; Clauser, Kane, & Swanson, 2002).

Williamson (2010), in a recent discussion thread on the *Language and Testing Research and Practice* mailing list, advised testing practitioners who use "essay" automated scoring systems to

> Examine the way that automated scoring is designed and ensure that the [constructs] that the automated scoring system is designed to measure are the same [constructs] you want to measure in the assessment. The current state-of-the-art of automated scoring lends itself well to measuring some components . . . but not others. . . . Automated scoring doesn't do exactly the same things that human graders do in scoring. Sometimes this is good and provides an advantage and other times this is bad and overlooks important

> aspects. You will want to make sure the purpose of the test, the items in the test, and the scoring rubrics for human graders are consistent with the capabilities and strengths of automated scoring (April 5, 2010).

These remarks are equally relevant, if not more so, to testing specialists using automated scoring systems to measure the phonetic and phonological knowledge and abilities of L2 speakers. There is evidence, as Hinks (2003) argues, that one of the constraints (i.e., limitations) involved in using ASR-based systems is that these systems are very poor at handling certain prosodic information such as intonation. To date, however, there have been no substantive studies, from a phonetic or phonological perspective, that have investigated the measurement of prosody in the context of ASR-based automated scoring systems and that have considered construct validity issues related to the use of such systems.

In part, to address these construct concerns, and in part to take up the call for more investigative efforts into the measurement of pronunciation and, hence, begin to bridge the language pedagogy and testing literature gap, this paper offers a theoretical examination of one important feature of phonetic and phonological knowledge, namely, prosody, and attempts to frame the discussion of prosody within the context of assessment. To this end, this paper first will define the construct of prosody and examine its role within a general speaking and discourse framework. It will then review related theoretical approaches and empirical studies that have investigated the function of prosody in the production and comprehension of L1 and L2 spoken language. Finally, these claims and findings from the speech-prosody literature will provide a context in which the underlying assumptions about "pronunciation" and prosody that have guided a few L2 measurement efforts will be examined.

TOWARDS A DEFINITION OF PROSODY IN THE CONTEXT OF SPOKEN DISCOURSE

In order to examine phonetic and phonological knowledge and, in specific, to have a better understanding of the role of prosody, a theoretical model of speaking that includes an articulatory/auditory phonetic and phonological component must first be examined and the relationship among prosody, other linguistic components, and discourse defined.

Bachman and Palmer's (1996) influential communicative language ability (CLA) framework, however, does not provide adequate guidance. In fact, the pronunciation and the related phonetic and phonological knowledge a speaker has in order to produce and comprehend an utterance is deemphasized in Bachman and Palmer's framework and conveniently grouped with "graphology" under the grammatical knowledge category. One reason for this might lie in the fact that these subcomponents of language knowledge are inextricably linked to aspects of channel and modality and, therefore, to "language skills," which Bachman and Palmer dismiss as the "contextualized realization[s] of the ability to use language in the performance of specific language use tasks" (pp. 75-76) rather than "language ability" per se.

However, from a speaking assessment perspective (see Fulcher, 2003; Luoma, 2004), there is a general recognition that the language "channel,"—that is, the sound of speech or what Fulcher refers to the "outer manifestation of speech" (p. 25)—is not only an important part of speaking ability but also an integral component in its definition. The phrase "sound of speech" is often used interchangeably with the term "pronunciation," which, according to Luoma (2004)

and Celce-Murcia et al. (1996), refers not only to individual sounds (i.e., segmentals) but also to pitch, volume, speed, pausing, stress, and intonation. According to Celce-Murcia et al. (1996) and Roach (2000), the latter group of sound characteristics, which affect whole sequences of syllables, utterances, and discourse segments, can be referred to and are collectively known as prosody.

The important role of pronunciation in speaking is also acknowledged by Bygate (1987), who outlined a model of "oral skills" that makes a clear distinction between grammar and lexical resources, on the one hand, and articulation and pronunciation, on the other. Later, drawing on Levelt's (1989) speech model, Bygate (1998) called for more investigations into both the processes and components of L2 oral production. Endorsing this distinct component view, and also using Levelt's framework, Douglas (1997) presented a speech model for the purposes of assessment that included a phonological component within a framework of five speech processors, which together account for the production and comprehension of the verbal message. The processors he outlined are a) the formulator, which handles grammatical and phonological encoding, b) the lexicon, which stores lexical units containing morphosyntactic and phonological forms as well as semantic and pragmatic information, and works in consort with the formulator, c) the articulator, which converts chunks of internal speech from the formulator and lexicon into actual speech, d) the auditor, which takes phonetic strings and inputs them into a speech comprehension system, which in turn parses the speech in terms of its phonological, morphological, syntactic, and semantic composition, and e) the conceptualizer, a processing system that handles a set of cognitive and metacognitive activities (e.g., conceiving of an intention, selecting the relevant information to be expressed for the realization of an intention, planning and ordering the information for expression, attending to and monitoring the expression, and so on). It should be noted that according to Levelt's model (1989), in order to encode or decode a message, a speaker must have access to two kinds of knowledge: procedural, which is an integral and inherent part of the processors listed above, and declarative knowledge, which holds a speaker's structured knowledge of the world, of other speakers, and of him or herself (e.g., encyclopedic, situational, contextual, propositional knowledge, etc.). Douglas understood Levelt's model to be compatible with Bachman and Palmer's (1996) "Language Use" and "Language Ability" framework and, in specific, equated Levelt's declarative knowledge and conceptualizer to Bachman and Palmer's "knowledge" and "strategic" components, respectively.

Levelt's (1989) model allows one to view the hypothesized relationships between procedural and declarative knowledge and to investigate the interactions among grammatical, lexical, and phonological processes in the context of these relationships. Furthermore, it sketches out a psychologically plausible outline of the phonological encoding and decoding processes engaged in the production and comprehension of connected speech. Part of this system is the so-called "prosody generator," which, according to Levelt, receives informational input from various sources (e.g., the speaker's intentionality, attentional state, world view), facilitates the procedural encoding and decoding of certain types of morpho-syntactic, lexical, phonological, and conceptual representations, superimposes on segmental sequences, and creates along with the formulator a phonetic plan that is externalized by the articulator or taken in by the auditor for parsing. The operations of the prosody generator, as well as the operations and interactions of all the main processors in Levelt's model, are executed in a parallel yet incremental fashion, accounting, thus, for the automaticity evidenced in the production and comprehension of L1 spoken language. Regarding L2 processing, there is an indication from second language acquisition studies that some of the problems that L2 learners face are due to a dependence on

mental representations and categories associated with their L1 sound system (Flege, 1987) and a reliance on less developed realization rules (i.e., faulty phonological encoding) and motor skills (see Leather & James, 1991).

In contrast to Douglas's (1997) contention, there are a few structural differences between Levelt's (1989) speech model and Bachman and Palmer's framework (1996; see also Bachman, 1990, and Bachman, 2002a), and these might affect the inferences and interpretations made about a speaker's linguistic knowledge and linguistic ability as they pertain to prosody from an assessment perspective. In Bachman and Palmer's framework, language ability, or, in other words, the capacity to produce and comprehend discourse, is conceptualized as the interaction between language knowledge and strategic competence, which is understood to be a set of metacognitive strategies such as goal setting, monitoring, and planning. In contrast, from Levelt's perspective, the capacity to produce and comprehend discourse can be viewed as the interaction among processors that are automatically and incrementally executed, on the one hand, and the interaction between this knowledge (i.e., procedural knowledge) and declarative knowledge (e.g., propositional knowledge, contextual knowledge of the world, etc.), on the other. What activates and regulates this process is the conceptualizer or, more specifically, the speaker's or hearer's intentional activity and attentional state. This is in agreement with certain postulations from the field of pragmatics (see Brown & Yule, 1983; Levinson, 1983) in which, in the "use" of language, an utterance's propositional content (i.e., meaning) and linguistic form are constrained by the communicative intent of (Austin, 1970) and the situational, contextual, ritual, and conversational norms held by the conversational participants (Goffman, 1974; Grice, 1975; Gumperz, 1982, 1992; Hymes, 1972). In relation to ethnographic approaches to communication, the knowledge a speaker has of a speech community, speech situation, and/or speech event could be viewed, within Levelt's schema, as declarative knowledge, and speech acts as the conduits between this knowledge and procedural knowledge.

In Levelt's (1989) model, prosodic knowledge can be seen as one type of procedural knowledge or, more aptly, a device, which does not carry with it propositional content but, rather, helps participants signal or narrow down (in conjunction with other phonological, lexical, morpho-syntactic, conceptual, and declarative knowledge) the interpretation of an utterance in terms of what is referenced, implied, presupposed, and/or inferred by the participants. In other words, prosody can be seen as a process that adjusts or maintains the "accessibility of meaning" on both the semantic and pragmatic level (Wichmann & Blakemore, 2006), playing an important role in the structuring of information within discourse. This is not unlike the speaking component of "key" that Hymes (1974) identified and presented as part of his SPEAKING mnemonic (see Hymes, 1974, pp. 53-66).

This emphasis on form-meaning associations at the pragmatic and discourse levels has also been outlined from a pedagogical perspective in the teaching of grammar (Larsen-Freeman, 2001) and from a CLA approach in the context of assessing grammatical ability (Purpura, 2004). In particular, Purpura defined grammatical knowledge within a framework that specified the relationships between grammatical forms, their literal meanings, and their implied meanings at both the sentential and discourse levels. Elaborating on the notion of implied grammatical meaning, Purpura argued that grammatical knowledge encompasses not only knowledge of grammatical form but also the knowledge of these forms as they relate to contextual (e.g., interpersonal), sociolinguistic (e.g., register and social norms), sociocultural (e.g., figurative speech), psychological (e.g., attitude and affect), and rhetorical (e.g., genres) meaning.

Furthermore, the aspects of intentionality and attentional states, which are key components to Levelt's (1989) model, have been included in recent "interactionalist" approaches (Chapelle, 1998b) to construct definitions in which the focus is on both knowledge of language (i.e., trait) and ability to put language to use (i.e., context). From this perspective, the definitions of a language construct considers not only internal trait and environmental variables but also a processing component such as Levelt's conceptualizer or Bachman & Palmer's (1996) strategic competence component, which brings about the interaction between the two. One such approach holds that language can be seen as complex adaptive system (CAS) (Beckner, Blythe, Bybee, Christiansen, Croft, Ellis, Holland, Ke, Larsen-Freeman, & Schoenemann, 2009; Larsen-Freeman & Cameron, 2008) consisting of a complex linguistic code and involving multiple speakers whose behaviors are based on their past and present interactions and are the result of attentional constraints and social motivations. According to Beckner et al., language structures and knowledge "emerge from interrelated patterns of experience, social interaction, and cognitive mechanisms" (p. 2). Here, it should be noted that the term "social interaction" is not really analogous to what some see as "interactional competence" (Kramsch, 1986), "ability – in language user – in contexts" (Chalhoub-Deville, 2003), "situated competence" (McNamara, 1997), or "co-constructed discursive practices" (He & Young, 1998; Young, 2000, 2002). Instead, "social interaction" in the context of CAS can be characterized more accurately as a shared cooperative activity that is dependent on "shared cognition," which, Beckner et al. point out, is a set of shared beliefs and intentions. From this perspective, language can be viewed as a system of conventional signaling devices that coordinates the shared activity of communication and that operates on four levels: "producing and attending to the utterance; formulating and identifying the proposition; signaling and recognizing the communicative intention; and proposing and taking up the joint action" (p. 4).

Finally, from a cognitive pragmatics approach, certain theories of discourse production and comprehension are aligned with the basic sectors of Levelt's (1989) model. For example, centering theory (Grosz & Sidner, 1986; Grosz, Joshi, & Weinstein, 1995; Walker, Joshi, & Prince, 1998) hypothesizes that discourse structure is a composite of three interacting components: a) the linguistic structure, which is the utterance that is shared by conversational participants, b) the intentional structure, which is the intended purpose (i.e., the speakers' intentions) underlying the discourse, and c) the attentional state, which is the participants' focus of attention as the discourse unfolds. According to Walker et al. (1998), these three constituents create a "focal space" that helps the conversation partners determine how an utterance fits with the rest of the discourse and helps the participants figure out why something was said and what it means. And within this theory of discourse, there have been a few studies that have investigated the relationship between prosody and discourse structure in both L1 and L2 (e.g., Hirschberg & Nakatani, 1996; Wennerstrom, 1998).

As argued in this section, prosodic features (i.e., sound frequency, intensity, and duration) are essential aspects of speaking and an integral part of its very definition. In specific, prosody can be viewed as a component of the articulatory process that impacts whole utterances and discourse segments and that helps speakers/hearers access and share several levels and nuances of meaning. The way that these features function in the production and comprehension of L1 and L2 spoken language and discourse, along with a list of various approaches to the study of prosody and supporting research, will be reviewed in the next section.

THE FUNCTION OF PROSODY: THEORETICAL APPROACHES AND EMPIRICAL RESEARCH

From a phonetic and phonological perspective, the components and relationships outlined by Levelt's (1989) model and by certain approaches to discourse and pragmatics discussed above have been, more or less, the focal points of various approaches to and empirical investigations into the study of prosody. The following approaches and studies, although quite diverse in terms of methodology and underlying assumptions about language use, are unified in their overall goal, which is to understand the role that prosody plays in the production and comprehension of spoken language and discourse. Representing only a small portion of the L1 prosody research, the approaches presented here focus mainly on the prosodic features of stress and prominence (i.e., sound duration and amplitude/intensity) and intonation (i.e., pitch and pitch variation) and have been chosen because they have been applied to L2 investigations.

Many of these approaches to speech-prosody have been grouped into one of two distinct schools of thought (see Couper-Kuhlen, 2001; Gumperz, 1982; Roach, 2000)—one being the British school, with its lineage traced to the work of "tone units" by H. E. Palmer in the 1920's, and the other, the American tradition, linked to the work of Kenneth Pike (1945) on "pitch phonemes." Within the British tradition, Halliday (1970, 1985, 2009) outlined the study of prosody in the context of language as a semiotic system. According to Halliday (2009), language is a meaning generating system, the function of which is to make sense of one's world (i.e., ideational function) and to get along with others (i.e., interpersonal function). The text or discourse produced or comprehended, he argued, and still argues, is the mapping of these functions on to each other and on to the context in which meanings are being exchanged (i.e., textual function). Within this "textualization" framework, Halliday (1985) argues that rhythm (i.e., syllabic stress patterns in terms of sound amplitude and vowel duration), tone (i.e., the combination of falling and rising pitch frequencies at the syllable level), and tone groups or units (i.e., the modulations of pitch forming intonational contours across many syllables) are used and manipulated by speakers to signal whether information within a text, as realized in grammatical constituents, is recoverable (i.e., "given") or not recoverable (i.e., "new" or salient) to the hearer by reference to the surrounding text or discourse.

Stress & Prominence

From a cognitive perspective, the relationship between prosodic features and prominence has been examined in the work of Chafe (1994), where it has been proposed that, in specific, intonation units, as identified by not only pitch levels but also pauses, speed, and voice quality (i.e., affective and attitudinal vocal effects), help interlocutors distinguish between information that is active in the consciousness of the speaker and hearer and information that is semi-active or inactive. More recently, and in line with Chafe's work, Wilson and Wharton (2006) extended the functions of prosody by distinguishing among two types of information conveyed by prosodic input: unintentional, which tend to be natural signs and signals, and intentional, which are mainly natural (i.e., paralinguistic) and linguistic signals. In both cases, prosodic signals, they argue, are gradient rather than categorical, which means that these signals, although not carrying any propositional content, are procedural in the sense that they guide the hearer to infer meaning in the comprehension process. Within this "relevance theory" perspective, prosody is seen as guiding the interpretation of an utterance not only in terms of prominence and textual features

such as grammar and referencing (as was the case with Halliday, 1985) but also in terms of processing effort and cognitive effects. According to Wilson and Wharton, different types of prosodic input help the hearer "follow a path of least effort, deriving whatever effects are made most accessible . . . [and] stopping when [the hearer] has enough effects to justify the extra effort caused by the departure from neutral (or 'expected') prosody" (p. 1567).

In psycholinguistic studies of L1 speech processing, investigations of prosody and prominence have focused on the "given-new distinction" (Prince, 1981) between unstressed and stressed words signaling new and contrastive information. These investigations found that when native speakers violated the "new-given" principle, speech processing and comprehension for the hearer became more difficult (e.g., see Terken & Hirschberg, 1994).

In terms of prosodic prominence and L2 processing and comprehension, Hahn (2004), in an experimental control and two-group design, found that a group of native speakers (NSs) who listened to a monologic passage spoken by a non-native speaker (NNS) using target-like prominent stress patterns comprehended the content of the passage better and evaluated the speech more favorably than did two other groups of native speakers who listened to the same NNS passage with misplaced or without primary stress patterns. However, processing difficulty as measured by the listeners' reaction time to background tone signals during the listening task was not significantly different among the three groups, indicating that the listeners did not allocate more processing resources to the speech with deviant stress patterns. Nevertheless, Hahn argues that the sizable difference in effect size between the target-like stress and misplaced stress groups, although not statistically significant, could be viewed as substantive when considered in light of the comprehension results.

Field (2005) also examined the prosodic feature of amplitude in an L2 context but focused mainly on the relationship between syllabic stress, vowel quality, and intelligibility (i.e., a hearer's word level recognition) at the lexical level. In Field's study, NSs and NNSs were asked to listen to a list of two syllable words that had either target like or non-target like stress patterns and whose accented syllables were manipulated in terms of vowel quality. In this study, intelligibility was measured as the extent to which the participants were able to transcribe the isolated words accurately. Field found that for both NSs and NNSs, intelligibility was reduced significantly by rightward stress shifts without vowel quality modifications. This finding, according to Field, is important given that NNS place great reliance on meaning representations at the lexical level in order to form expectations and make interpretations at the utterance and discourse levels.

Intonation

Another approach to prosody that comes from the British school is illustrated in the work of David Brazil (1997) and his associates (Brazil, Coulthard, & Johns, 1980). This approach looks at the communicative value of intonation, and makes the case that tones (i.e., falling, rise-falling, rising, or fall-rising tones), and the choice of tones, express whether a speaker's assertion can be shared with a hearer, creating thus a "state of convergence" between speaker and hearer (Brazil, 1997, p. 70). Brazil also proposed an interpretation of larger prosodic units, which he viewed as sound "paragraphs" that are marked by pitch sequences (i.e., "key" and "termination" tones). According to Brazil, a low pitch sequence, for example, ending one discourse segment followed by a high pitch sequence starting another discourse segment indicates a semantic and

structural separation between the two discourse segments, while mid pitch termination followed by a mid pitch opening can be interpreted as an indication of topic continuation.

Brazil's (1997) framework has been used by Pickering (2001) and Pickering (2004) to investigate the difference between NS and NNS use of tone choices and intonational "paragraphs" in the context of extended discourse. In one study, Pickering (2001) analyzed and compared the tone choices, pitch levels, and discourse structures used by NS teaching assistants and those used by NNS international teaching assistants (ITAs) in the context of instructional interactions. Using a discourse analytic approach and employing auditory and computerized sound analysis instruments to extract pitch function and fundamental frequency traces, she found that all of the NSs used a combination of rising and falling tones as rapport-building strategies in their responses to students, while most NNSs used a series of level and co-occurring falling tones. According to Pickering, this showed that while the NSs were able to use tone choice to create "common ground" with their students, the NNSs lacked this ability and adapted a tone choice strategy that indicated withdrawal and disengagement. In another qualitative research study, Pickering (2004), again following Brazil's model, investigated and compared the use of tone "keys" and tone "terminations" between NS and NNS to signal intonational paragraph boundaries. She found that while the NSs were able to match prosodic signals with discourse structural cues (e.g., phrasal discourse markers) to mark the information in discourse units and the relationship between discourse units, the NNSs were unable to do so, creating a discord in the internal structure of the discourse and confounding attempts by NS hearers to use these prosodic features as organizational cues.

Another approach to the meaning of intonation comes from the American tradition. In particular, Pierrehumbert (1980) and Pierrehumbert and Hirschberg (1990), differentiating between prosodic features of pitch, duration, amplitude, and other vocal characteristics of speech, have put forward a theory of intonational meaning that asserts that only two tones (i.e., high and low pitch frequencies) are perceptually distinct. Moreover, they argue that the combination of these tones at the lexical level (i.e., pitch accents), at the phrasal level (i.e., phrase accents), and at the discourse level (i.e., boundary tones) are used by speakers to specify a particular relationship between the meaning of an utterance and the shared beliefs and intentions of the participants. The composition of these three types of tone groups create what Pierrehumbert and Hirschberg call "tune" meanings, which can convey, among other things, informational status (e.g., new and salient versus new but not salient) and attitude (e.g., uncertainty, politeness, surprise) and can mark relationships between conjoined clauses (e.g., temporal versus causal relationships) as well as between discourse segments (e.g., anaphoric versus cataphoric referencing). Finally, this approach is based on Pierrehumbert's (1980) "autosegmental" model of intonation, which sees pitch levels as phonemes (cf. Pike's notion of "pitch phoneme," 1945) and which has been supported by biophysical articulatory and auditory research (see Xu, 2005).

In L1 corpus-based empirical work, Pierrehumbert and Hirschberg's (1990) model of intonational meaning has been used to investigate the relationship between prosodic variation and discourse structure. In specific, Grosz and Hirschberg (1992), employing a method of discourse analysis based on Grosz and Sidner (1987) (see previous section), examined the prosodic features of pitch range, amplitude, and timing used in audio-recorded news stories by having two groups of trained L1 participants label the discourse structure of the news stories either from written text alone or from text plus speech. They found high inter-labeler reliability among the participants in the text plus speech group and also statistically significant associations

between prosodic features and discourse structure at both the utterance and extended discourse levels. These findings have also been supported by Hirschberg and Nakatani (1996), who explored a corpus of direction-giving monologues and found that the listeners in similar types of labeling tasks were exploiting prosodic cues to increase their accuracy of discourse segmentation.

Pierrehumbert and Hirschberg's (1990) approach has also guided a few L2 research studies (e.g., Wennerstrom, 1994,,1998; Wennerstrom and Siegel, 2003). In particular, Wennerstrom (1998) examined the ability of Chinese NNS of English to use pitch accents, phrase accents, and boundary tones in order to signal boundaries between lexical items, adjacent phrasal constituents, information units, and discourse segments in the context of academic discourse. The participants were asked to each give a 10-minute presentation on an academic topic of interest. The presentations were videotaped, scored by NS raters in terms of comprehensibility (i.e., the listener's ease of understanding), and transcribed and analyzed for pitch differences and range using computerized sound-analysis instruments. The comprehensibility scores were then regressed on the intonation variables. Wennerstrom found that the only significant predictor of comprehensibility scores was the boundary tone component, which suggests, she argued, that intonation at the macro-discourse level, which helps the speaker indicate topic shifts, seems to be an important function of intonation and crucial in terms of a hearer's ability to follow and comprehend the content of discourse.

Social & Cultural Perspectives

Finally, and quite distinct from the approaches discussed so far, two other views on the function of prosody have been advanced. These views, in general, see language use predominately as a social act and prosody, in specific, as a phonetic resource in the construction of interpersonal meaning. The first was expressed by Gumperz (1982, 1992), who, from a sociolinguistic perspective, viewed prosody as essential to the interpretation of meaning in conversational exchanges. Gumperz understood prosodic features to be types of cues (i.e., contextualization cues) that help participants make conversational inferences and contextual presuppositions based, in part, on their shared social and cultural assumptions about communicative intents and interpersonal relations. Structurally, however, Gumperz methodological approach was similar to Halliday's (1970) in that it viewed different dimensions of prosody as the "conceptual conflations" (Gumperz, 1982, p. 100) of three distinct characteristics: sound frequency, amplitude, and duration. Moreover, the three analytical components that Gumperz identified (i.e., "tone groups," "nucleus placement," and "melodic shape") were, more or less, identical to those presented by Halliday.

The second view comes from conversation analysis (CA). CA is an approach that sees language use as consisting of a series of turns, which participants interactively and cooperatively build one after another to create different types of sequences of talk that are done for various pragmatic ends (Schegloff, 2007). Within this framework, prosodic features such as intonation and pausing are seen as important to the composition of the turn (i.e., utterance) and to the organization of turn-taking (Sacks, Schegloff, & Jefferson, 1974). Using a corpus of naturally occurring conversations, Ford and Thompson (1996) investigated the relationship among intonational, syntactic, and pragmatic completion points in turn-taking and found that turns tended to shift when syntactic completions, final intonations, and pragmatic closures co-occurred. They also found that intonational boundaries were more reliable as "turn-yielding

cues" (see Levelt, 1989) and speaker change than as syntactic completion points. These findings have been corroborated by Wennerstrom and Siegel (2003), who, employing logistic regression analysis, found that syntax, intonation, as well as pause duration, were good predictors of turn shifts. It should be noted that while Ford and Thomspon's qualitative study distinguished between two categories of intonation, Wennerstrom and Siegel's quantitative study, which used Pierrehumbert and Hirschberg's (1990) methodological approach, found six. Other CA guided studies have focused on the relationship between prosody and the meanings expressed in turns. In specific, Ogden (2006) found that, given the right prosodic shape, a turn structured lexically or syntactically to convey agreement may in fact project disagreement. Also, Curl, Local, and Walker (2006) examined how certain prosodic features such as speed, sound intensity, and pitch can alter the pragmatic function of a repeated utterance within a sequence of talk and signal closure. What should be emphasized with CA is that its focus is not on what the participants' perceived intentions or motives are, as was the case with all of the above approaches, but, rather, on how the participants themselves orient to a given utterance; that is, how the participants demonstrate understanding toward talk.

Overall, the approaches and empirical studies that have been presented and outlined in this section have shown that prosodic features such as fundamental frequency, sound intensity, and sound duration function in multiple and complex ways in the production and comprehension of L1 and L2 spoken language and discourse. In short, these functions can be listed as accentual, lexical/grammatical, attitudinal, socio-cultural, organizational/discourse, and interactional. What underscores all of these approaches, however, and what can be seen as a common denominator across all prosodic functions, is an understanding that prosody, in combination with other manifestations of speech (e.g., morphosyntactic, lexical, and other phonological forms), is primarily used by speakers to signal and by listeners to perceive finer shades of both intrapersonal and interpersonal meanings. Finally, it is this paper's position that these functions and uses, along with the forms, of prosody need to be considered in assessing and measuring L2 spoken language.

MEASURING PROSODY IN L2 SPOKEN LANGUAGE

While there have been several L1 and L2 studies that investigated the nature of prosody in spoken language, the relationships between prosodic forms and their functions have not been explored adequately in the context of L2 assessment. This section will review the manner in which the measurement of certain prosodic features have been examined in L2 assessment and consider the constructs as well as the instruments used in light of the findings listed in the previous section.

The first major step in measuring L2 prosody was taken by Robert Lado (1961), who applied Pike's (1945) theoretical postulations to L2 testing methods. In specific, Lado emphasized the importance of testing both the production and perception of prosodic features such as stress and intonation and outlined a set of possible test item types for these purposes. In this sense, his work foreshadowed Major's (1998) claim that while speech perception and production are related, they are not necessarily identical processes (see also Buck, 1989). Several of Lado's test items were designed to measure stress and intonation productively in terms of their associated meanings and functions (e.g., prominence, attitude, question type, phrasal and grammatical relations). His focus, though, was on measuring prosodic abilities in the perceptive

mode. Regarding intonation, Lado argued that although direct approaches were "highly desirable," the relationship between intonation and meaning were "extremely vague and slippery" (p.130), and pointed out a few, but major, problems with the measurement of intonation as it related to the design of test production techniques. These concerns, which include content relevance and scoring objectivity, have been, and remain, central issues surrounding not only pronunciation testing (e.g., van Weeren & Theunissen, 1987) but also direct and semi-direct oral performance testing (e.g., see Bachman, 1988, 2002a, 2002b; Brown & Hudson, 1998; Clark, 1979; Johnson, 2000; Jones, 1979a, 1979b, 1985; Lazaraton, 1996, 2002; McNamara, 1996; O'Loughlin, 2001; Perren, 1968; Shohamy, 1994; Swain, 2001; Wilds, 1975). Finally, because of practical issues (e.g., lack of trained raters, time constraints, etc.), Lado reasoned that the "testing of intonation productively [was] destined to remain a restricted activity in language testing with the consequent drop of interest in teaching intonation" (p. 137). Indeed, many of the pronunciation tests, as well as related investigations of testing techniques and scoring procedures, throughout most of the 60's and 70's, focused primarily on segmentals at the lexical and/or sentential level (e.g., Briere, 1967; de Jong, 1977; Whiteson, 1978).

In the 90's, an equal emphasis on both segmental and prosodic features reemerged in the form of both assessment tasks designed to diagnose perception and production in the context of the classroom (Celce-Murcia et al., 1996) and alternative scales that profiled a learner's pronunciation as it impacted communication (Morley, 1991). Along these lines, a refocusing of L2 pronunciation testing was also proposed by Koren (1995), who outlined and evaluated a test of pronunciation that focused on the production of segmental sounds, stress, and intonation in different, according to Koren, "speech situations." In specific, the test tasks that Koren designed and used followed Lado's (1961) production tests and Tarone's (1983) interlanguage continuum model (which states that speech production varies with elicitation technique). These ranged from controlled techniques (i.e., repetition of words, reading passages) to less controlled techniques (i.e., describing a story or acting out a role play while following a script), and the scoring was done by two raters using a 5-point scale ranging from "very heavy non-native pronunciation" to "very native-like pronunciation." While the inter-rater reliability for each component was acceptable (i.e., ranging from .61 for "stress" to .94 for "phonemic sounds"), and while there was evidence that the test was able to discriminate among the testees' different levels of pronunciation as determined by "native-like" control, Koren did not provide any additional information about how the constructs of stress and intonation were defined and operationalized (i.e., how the samples of performance elicited by these tasks could relate to or reflect the underlying claims made about prosodic knowledge and ability) or how the prosodic forms measured in the extended response tasks related to either their discourse or interactional functions. These, it should be noted, are issues with Celce-Murcia et al.'s "free speech" tasks as well. Moreover, the scripted nature of the story and role play tasks undermined, to a great extent, the essential element of communicative purposefulness (Bachman & Palmer, 1996) or intentionality (Levelt, 1989) that is assumed to mobilize, and be reflected in, prosodic ability.

In the last decade, there have been a few noteworthy studies that looked at the measurement of prosodic features either from an overall pronunciation ability perspective (e.g., Gorsuch, 2001; Isaacs, 2008; Yoshida, 2004) or within a general speaking framework (e.g., Iwashita, Brown, McNamara, & O'Hagan, 2008). However, similarly to Koren's (1995) investigation, these studies and tests focused mainly on the forms of prosody with little, if any, consideration of form-function-meaning associations in the definitions of the constructs, in the structure of the tasks, and/or in the design of the rating rubrics. In specific, the list of prosodic

forms examined in these studies was quite extensive, yet the only prosodic function included was the intonational function of "yes/no" questions (see Gorsuch, 2001). In terms of test methods, the tasks used in these studies were in the form of either dialogue/prose readings or SPEAK/iBT (Educational Testing Service, 2000, 2005) question types eliciting monologic speech, and the testees' performances were evaluated through either binary judgments (e.g., native-like or not), scalar judgments (e.g., ease of understanding), or frequency counts of instances of target-like forms. Overall, none of the researchers in these studies provided any substantive rationales for the use of these assessment task types or for the appropriateness of their scoring rubrics in terms of prosodic knowledge and ability. They did not explain, for instance, how these tasks or rubrics related to any theoretical views of prosodic ability, and it is not clear how the observed performances on these tasks revealed these prosodic abilities. Finally, the only attempt to relate prosodic forms to meaning was in Isaacs' study. Similarly to Morley (1991), Isaacs used the rather elusive concepts of intelligibility and comprehensibility as criterion measures, which is also a common practice in studies of L2 pronunciation (e.g., see Rajadurai, 2007; Zielinski, 2006, 2008). However, one should note that relying solely on, for example, intelligibility to measure phonetic and phonological ability might be problematic since, according to Kim (2009), this concept may relate more to attributes residing within the listener or rater than to the phonetic and phonological knowledge and abilities of the speaker. One can also extend this argument to studies of accentedness (e.g., see Jamieson & Poonpon, 2013; Kang, Rubin, & Pickering, 2010), which, although include stress and intonation, can be defined as an auditory effect of pronunciation.

Finally, form-oriented approaches to construct definitions of prosody have also been the prevailing paradigm to the design, development, and evaluation of speaking tests that utilize automated scoring systems. This is mainly due to the systems' technological limitations. An essential part of such systems is the mechanism involved in phonetically decoding human speech. This includes two components: a speech recognition component (i.e., ASR), which is based on phone recognition (i.e., spectral matches of sound), and an articulatory statistical model (e.g., a Hidden-Articulator Markov Model), which approximates the characteristics and physical constraints of human articulatory processes (Richarson, Bilmes, & Diorio, 2003). So far, the pronunciation scoring structure used by these systems, according to Franco et al. (2000), is very limited and consists of a combination of phone scores (i.e., segmental sound accuracy), phone duration scores, and speech rate (i.e., the mean number of phones per unit of time in a sentence). Interestingly, neither sound amplitude (i.e., stress) nor frequency variations (i.e., intonation) are measured. In fact, these prosodic features are ignored as erroneous signals that interfere with and confound the measurement of the phone accuracy component (Hinks, 2003).

In a recent set of validity studies, ETS's automated scoring system, SpeechRater^{SM}, was investigated (Xi et al., 2008; Zechner, Higgins, Xi, & Williamson, 2009). Its "delivery" construct and scoring structure followed, more or less, the one outlined by Franco et al. (2000). While Xi et al. provided evidence to support the accuracy of the automated scores and their generalizability across different tasks, they admitted that the features used in the scoring model were only a subset of those included in the construct of speaking used in the TOEFL iBT. This, the researchers acknowledged, reduced the test's explanatory power. One could argue that the adequacy of the test's power in explaining a test-taker's performance would be challenged even further by an approach that considers not just a sufficient set of prosodic features but also their accentual and discourse functions. Such an approach, it seems, would be imperative in guiding the development of ASR-based speaking tests in that it could provide these systems with a

linguistically and psychologically plausible blueprint. More importantly, an understanding of prosodic functions can help guide the formulation of claims that need to be incorporated in validity arguments put forth to support the use (Bachman, 2005) of such systems. And this, of course, applies to all efforts investigating the measurement of prosody in L2 spoken language.

Overall, since the 90's, although research in the measurement of pronunciation has shifted to include suprasegmentals, most studies have simply focused on what one could label as form - form associations rather than the more subtle form-meaning-use relationships outlined in the previous section of this article.

CONCLUSION

With the challenges and opportunities that new scoring technologies bring, a different perspective and approach to the measurement of phonetic and phonological knowledge is called for—one that not only includes prosodic forms but also considers how they are used by participants to access meaning in L2 spoken discourse. By providing an extended definition of prosody in the context of spoken discourse, by reviewing related empirical research, and by examining the implications for testing, it is hoped that this article, essentially a literature review, has added to a better understanding of the various issues involved.

REFERENCES

American Council on the Teaching of Foreign Languages. (1999). *ACTFL Oral Proficiency: Interview tester training manual*. Yonkers, NY: Author.

Anderson-Hsieh, J., Johnson, R., & Koehler, K. (1992). The relationship between native speaker judgments of nonnative pronunciation and deviance in segmentals, prosody, and syllable structure. *Language Learning, 42,* 529-555.

Austin, J. L. (1970). *Philosophical Papers.* Oxford: OUP.

Bachman, L. F. (1988). Problems in examining the validity of the ACTFL Oral Proficiency Interview. *Studies in Second Language Acquisition, 10*, 149-164.

Bachman, L. F. (1990). *Fundamental considerations in language testing*. New York: OUP.

Bachman, L. F. (2002a). Alternative interpretations of alternative assessments: some validity issues in educational performance assessments. *Educational Measurement: Issues and Practice, 21*, 5-18.

Bachman, L. F. (2002b). Some reflections on task-based language performance assessment. *Language Testing, 19,* 453-476.

Bachman, L. F. (2005). Building and supporting a case for test use. *Language Assessment Quarterly, 2, 1-34.*

Bachman, L. F., & Palmer, A. S. (1996). *Language testing in practice.* New York: OUP.

Beckner, C., Blythe, R., Bybee, J., Christiansen, M. H., Croft, W., Ellis, N. C., Holland, J., Ke, J., Larsen-Freeman, D., & Schoenemann, T. (2009). Language is a complex adaptive system: position paper. *Language Learning, 59 (Suppl. 1)*, 1-26.

Bernstein, J., De Jong, J., Pisoni, D., & Townshend, B. (2000). Two experiments on automatic scoring of spoken language proficiency. In P. Delcloque (Ed.), *Proceedings in InSTIL2000* (pp. 57-61). Dundee, Scotland: University of Abertay.

Brazil, D. (1997). *The communicative value of intonation in English.* New York: CUP.

Brazil, D., Coulthard, M., & Johns, C. (1980). *Discourse intonation and language teaching.* New York: Longman.

Briere, E. J. (1967). Phonological testing reconsidered. *Language Learning, 17,* 163-171.

Brown, G. & Yule, G. (1983) *Discourse analysis*. Cambridge: CUP.

Brown, J. D., & Hudson, T. (1998). The alternatives in language assessment. *TESOL Quarterly*, *32*, 653-675.

Buck, G. (1989). Written tests of pronunciation: do they work? *ELT Journal, 43*, 50-56.

Bygate, M. (1987). *Speaking*. New York: OUP.

Bygate, M. (1998). Theoretical perspectives on speaking. *Annual Review of Applied Linguistics, 18,* 20-42

Carnegie Speech Assessment™ [CD-ROM] 2010. Pittsburgh, PA: Carnegie Speech Company.

Cauldwell, R. (2002). *Streaming Speech.* [CD-ROM]. Birmingham, England: speechinaction.

Celce-Murcia, M., Brinton, D. M., & Goodwin, J. M. (1996). *Teaching pronunciation: a reference for teachers of English to speakers of other languages.* New York: CUP.

Chafe, W. L. (1994). *Discourse, consciousness, and time: the flow and displacement of conscious experience in speaking and writing.* Chicago: University of Chicago Press.

Chalhoub-Deville, M. (2003). Second language interaction: current perspectives and future trends. *Language Testing, 20*, 369-383.

Chapelle, C. A. (1998a). Multimedia call: lessons to be learned from research on instructed SLA. *Language Learning & Technology, 2,* 21-39.

Chapelle, C. A. (1998b). Construct definition and validity inquiry in SLA research. In L. F. Bachman & A. D. Cohen (Eds.), *Interfaces between second language acquisition and language testing research* (pp. 32-70). New York: CUP.

Chapelle, C. A. (2010). Does an argument-based approach to validity make a difference. *Educational Measurement: Issues and Practice, 29*, 3-13.

Clark, J. L. D. (1979). Direct vs. semi-direct tests of speaking ability. In E. J. Briere & F. B. Hinofotis (Eds.), *Concepts in language testing: some recent studies* (pp. 35-49). Washington, DC: TESOL

Clauser, B. E., Kane, M. T., & Swanson, D. B. (2002). Validity issues for performance-based tests scored with computer-automated scoring systems. *Applied Measurement in Education, 15*, 413-432.

Council of Europe (2001). *A Common European Framework of reference for language learning, teaching, assessment*. Cambridge: CUP

Couper-Kuhlen, E. (2001). Intonation and discourse: current views from within. In D. Schiffrin, D. Tannen, & H. E. Hamilton (Eds.), *The handbook of discourse analysis* (pp. 13-34). Malden, MA: Blackwell Publishing Ltd.

Curl, T. S., Local, J., & Walker, G (2006). Repetition and the prosody -- pragmatics interface. *Journal of Pragmatics, 38,* 1721-1751.

Cutler, A., Dahan, D., & van Donselaar, W. (1997). Prosody in the Comprehension of Spoken Language: a literature review. *Language and Speech, 40,* 141-201.

Dauer, R. M. (2005). The lingua franca core: a new model for pronunciation instruction? *TESOL Quarterly, 39*, 543-550.

de Jong, W. N. (1977). On validating a pronunciation test. *ELT Journal, 31*, 233-239.

Derwing, T. M. & Munro, M. J. (1997). Accent, intelligibility, and comprehensibility: evidence from four L1s. *Studies in Second Language Acquisition, 20,* 1-16.

Derwing, T. M. & Munro, M. J. (2005). Second language accent and pronunciation teaching: a research-based approach. *TESOL Quarterly, 39*, 379-397.

Derwing, T. M., & Rossiter, M. J. (2003). The effects of pronunciation instruction on the accuracy, fluency, and complexity of L2 accented speech. *Applied Language Learning, 13*, 1-17.

Derwing, T. M., Munro, M., & Wiebe, G (1998). Evidence in favor of a broad framework for pronunciation instruction. *Language Learning, 48*, 393-410.

Douglas, D. (1997). *Testing speaking ability in academic contexts: theoretical considerations* (TOEFL Monograph Series). Princeton, NJ: Educational Testing Service.

Downey, R., Farhady, H., Present-Thomas, R., Suzuki, M., & Moere, A. V. (2008). Evaluation of the usefulness of the Versant for English Test: a response. *Language Assessment Quarterly, 5*, 160-167.

Educational Testing Service. (2000). *TSE and SPEAK score user guide.* Princeton, NJ: Author

Educational Testing Service. (2005). *Standard Setting: iBT/Next Generation TOEFL*. [CDROM]. Available: ETS

Elhilali, M., Taishih, C., & Shamma, S. A. (2003). A spectro-temporal modulation index (STMI) for assessment of speech intelligibility. *Speech Communication, 41,* 331-348.

Field, J. (2005). Intelligibility and the listener: the role of lexical stress. *TESOL Quarterly, 39,* 399-423.

Flege, J. E. (1987). The instrumental study of L2 speech production: some methodological considerations. *Language Learning,* 37, 285-296.

Ford, C. E. & Thompson, S. A. (1996). Interactional units in conversation. In E. Oches, E. A. Schegloff & S. Thompson (Eds.) *Interaction and grammar* (pp. 135-184). New York: Cambridge University Press.

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.* Menlo Park, CA: SRI International. Retrieved from www.speech.sri.com.

Fulcher, G. (2003). *Testing second language speaking.* New York: Longman.

Goffman, E. (1974). *Frame analysis.* Boston, MA: Northeastern University Press.

Gorsuch, G. J. (2001). Testing textbook theories and tests: the case of suprasegmentals in a pronunciation textbook. *System, 29,* 119-136.

Grice, H. P. (1975). Logic and conversation. In P. Cole and J. L. Morgan (Eds.), *Syntax and semantics 3: speech acts* (pp. 41-58). New York: Academic Press.

Grosz, B. J., & Hirschberg, J. (1992). Some intonational characteristics of discourse structure. *Proceedings of the Second International Conference on Spoken Language Processing* (429-432), Banff, Canada.

Grosz, B. J., & Sidner, C. L. (1986). Attention, intentions, and the structure of discourse. *Computational Linguisitics, 12,* 175-204.

Grosz, B. J., Joshi, A., & Weinstein, S. (1995). Centering: a framework for modeling the local coherence of discourse. *Computational Linguistics, 2,* 203-225.

Gumperz, J. J. (1982). *Discourse strategies.* New York: CUP.

Gumperz, J. J. (1992). Contextualization and understanding. In A. Duranti & C. Goodwin (Eds.), *Rethinking context: language as an interactive phenomenon* (pp. 229-252). New York: CUP.

Hahn, L. D. (2004). Primary stress and intelligibility: research to motivate the teaching of suprasegmentals. *TESOL Quarterly, 38*, 201-223.

Hakkinen, J., Suontausta, J., Riis, S., & Jensen, K. J. (2003). Assessing text-to-phoneme

mapping strategies in speaker independent isolated word recognition. *Speech Communication 41,* 455-467.

Halliday, M. A. K. (1970). *A course in spoken English: Intonation*. London: OUP

Halliday, M. A. K. (1985). *An introduction to functional grammar.* New York: Edward Arnold.

Halliday, M. A. K. (2009). Methods - techniques - problems. In M. A. K. Halliday & J. J. Webster (Eds.), *Continuum companion to systemic functional linguistics* (pp. 59-86), New York: Continuum International Publishing Group.

He, A. W., & Young, R. (1998) Language proficiency interviews: a discourse approach. In R. Young & A. W. He (Eds.), *Talking and testing: Discourse approaches to the assessment of oral proficiency* (pp. 1-24). Philadelphia: John Benjamins.

Hinks, R. (2003). Speech technologies for pronunciation feedback and evaluation. *ReCall*, *15*, 3-20.

Hirschberg, J., & Nakatani, C. (1996) A Prosodic Analysis of Discourse Segments in Direction-Giving Monologues *Proceedings of the Association for Computational Linguistics*, Santa Cruz, June, pp. 286–293.

Hymes, D. H. (1972). On Communicative competence. In J.B. Pride & J. Holmes (Eds.), *Sociolinguistics: Selected Readings* (pp. 269-293). Middlesex, UK: Penguin.

Hymes, D. H. (1974). Ways of speaking. In R. Bauman and J. Sherzer (Eds.), *Explorations in the ethnography of speaking* (pp. 433-451). New York: CUP.

Isaacs, T. (2008). Towards defining a valid assessment criterion of pronunciation proficiency in non-native English-speaking graduate students. *The Canadian Modern Language Review, 64,* 555-580.

Iwashita, N., Brown, A., McNamara, T., & O'Hagan, S. (2008). Assessed levels of second language speaking proficiency: How distinct? *Applied Linguistics, 29*, 1, 24-49.

Jamieson, J., & Poonpon, K. (2013) Developing analytic rating guides for TOEFL iBT integrated speaking tasks. (TOEFL iBT Research Report 20). Princeton, NJ: Educational Testing Service.

Janse, E., Nooteboom, S., & Quene, H. (2003). Word-level intelligibility of time-compressed speech: prosodic and segmental factors. *Speech Communication 41,* 287-301

Jenkins, J. (2000). *The phonology of English as an international language.* Oxford: OUP.

Jenkins, J. (2002). A sociolinguistically based, empirically researched pronunciation syllabus for English as an international language. *Applied Linguistics, 23*, 83-103.

Jenkins, J. (2005). Implementing an international approach to English pronunciation: the role of teacher attitudes and identity. *TESOL Quarterly, 39,* 535-542.

Johnson, M. (2000). Interaction in the oral proficiency interview: Problems of validity. *Pragmatics, 10,* 215-231.

Jones, R. L. (1979a). Theoretical and technical considerations in oral proficiency testing. In R. L. Jones & B. Spolsky (Eds.), *Testing language proficiency* (pp. 10-28), Arlington, VA: Center for Applied Linguistics.

Jones, R. L. (1979b). Performance testing of second language proficiency. In E. J. Briere & F. B. Hinofotis (Eds.), *Concepts in language testing: some recent studies* (pp. 50-57), Washington, DC: TESOL.

Jones, R. L. (1985). Second language performance testing: an overview. In P. C. Hauptman, R. LeBlanc, & M. D. Wesche (Eds), *Second language performance testing* (pp. 15-24). Ottawa, Canada: University of Ottawa Press.

Kang, O., Rubin, D., & Pickering, L. (2010). Suprasegmental measures of accentedness and judgments of language learner proficiency in oral English. *The Modern Language Journal*, *94*, 1-13.

Kennedy, S. & Trofimovich, P. (2008). Intelligibility, comprehensibility, and accentedness of L2 speech: the role of listener experience and semantic context. *The Canadian Modern Language Review, 64,* 459-489.

Kim, H-J. (2009). *Investigating the effects of context and task type on second language speaking ability.* Unpublished doctoral dissertation, Teachers College, Columbia University, New York City.

Koren, S. (1995). Foreign language pronunciation testing: a new approach. *System, 23,* 387-400.

Kramsch, C. (1986). From language proficiency to interactional competence. *The Modern Language Journal, 70,* 366-372.

Lado, R. (1961). *Language testing: The construction and use of foreign language tests.* New York: McGraw-Hill Book Company.

Larsen-Freeman, D. (2001). Teaching grammar. In M. Celce-Murcia (Ed.), *Teaching English as a second or foreign language* (pp. 251-266), Boston, MA: Heinle & Heinle.

Larsen-Freeman, D., & Cameron, L. (2008). *Complex systems and applied linguistics.* New York: OUP.

Lazaraton, A. (1996). Interlocutor support in oral proficiency interviews: the case of CASE. *Language Testing, 13*, 151-172.

Lazaraton, A. (2002). *A qualitative approach to the validation of oral language tests.* New York: CUP.

Leather, J., & James, A. (1991). The acquisition of second language speech. *Studies in Second Language Acquisition, 13,* 305-341.

Levelt, W. J. M. (1989). *Speaking: From intention to articulation.* Cambridge, MA: MIT Press.

Levinson, S. C. (1983). *Pragmatics*. New York: CUP.

Luoma, S. (2004). *Assessing speaking*. Cambridge: Cambridge University Press.

Major, R. C. (1987). Measuring pronunciation accuracy using computerized techniques. *Language Testing, 4,* 155-169.

Major, R. C. (1998). Interlanguage phonetics and phonology. *Studies in Second Language Acquisition, 20,* 131-137.

McNamara, T. (1996). *Measuring second language performance.* Essex, England: Longman.

McNamara, T. F. (1997). "Interaction" in second language performance assessment: whose performance? *Applied Linguistics, 18,* 446-466.

Meyer, G., & Morse, R. (2003). The intelligibility of consonants in noisy vowel-consonant-vowel sequences when the vowels are selectively enhanced. *Speech communication, 41,* 429-440.

Morley, J. (1991). The pronunciation component in teaching English to speakers of other languages. *TESOL Quarterly, 25,* 481-520.

Munro, M. J. & Derwing, T. M. (1995). Foreign accent, comprehensibility, and intelligibility in the speech of second language learners. *Language Learning, 45,* 73-97.

Munro, M. J. & Derwing, T. M. (2001). Modeling perceptions of the accentedness and comprehensibility of L2 speech: the role of speaking rate. *Studies in Second Language Acquisition, 23,* 451-468.

Munro, M. J., Derwing, T. M., & Morton, S. L. (2006). The mutual intelligibility of L2 speech. *Studies in Second Language Acquisition, 28,* 111-131.

NativeAccent® [CD-ROM] (2010). Pittsburgh, PA: Carnegie Speech Company.

Ockey, G. J., & French, R. (2014). From one to multiple accents on a test of L2 listening comprehension. *Applied Linguistics*. Advance access. Downloaded from *http://applij.oxfordjournals.org*.

O'Loughlin, K. (2001). *The equivalence of direct and semi-direct speaking tests.* New York: CUP

Ogden, R. (2006). Phonetics and social action in agreements and disagreements. *Journal of Pragmatics, 38,* 1572-1775.

Pearson (2008). Versant English Test: Test Description and Validation Summary. Retrieved from *http://pearsonpte.com/research/automatedscoring*

Pennington, M. C., & Richards, J. C. (1986). Pronunciation revisited. *TESOL Quarterly, 20,* 207-225.

Perren, G. E. (1968). Testing spoken language: some unsolved problems. In A. Davies (Ed.), *Language Testing Symposium: a psycholinguistic Approach.* New York: OUP.

Pickering, L. (2001). The role of tone choice in improving ITA communication in the classroom. *TESOL Quarterly, 35,* 233-255.

Pickering, L. (2004). The structure and function of intonational paragraphs in native and nonnative speaker instructional discourse. *English for Specific Purposes 23,* 19-43.

Pierrehumbert, J. (1980). *The phonology and phonetics of English intonation*, Doctoral dissertation, MIT.

Pierrehumbert, J., & Hirschberg, J. (1990). *The meaning of intonational contours in the interpretation of discourse.* In P. R. Cohen, J. Morgan, and M. E. Pollack (Eds.), *Intentions in communication* (pp. 271-311). Cambridge, MA: MIT Press.

Pike, K. (1945). *The intonation of American English.* Ann Arbor, MI: University of Michigan.

Prince, E. (1981). Toward a taxonomy of given-new information. In P. Cole (Ed.), *Radical Pragmatics* (pp. 223-256). New York: Academic Press.

Purpura, J. E. (2004). *Assessing Grammar.* New York: CUP

Rajadurai, J. (2007). Intelligibility studies: a consideration of empirical and ideological issues. *World Englishes, 26,* 87-98.

Richardson, M, Bilmes, J., & Diorio, C. (2003). Hidden-articulator Markov models for speech recognition. *Speech Communication*, *41*, 511-529.

Roach, P. (2000). *English phonetics and phonology: a practical course.* New York: CUP.

Sacks, H., Schegloff, E. A., & Jefferson, G. (1974). A simplest systematics for the organization of turn-taking in conversation. *Language, 50,* 696-735.

Schegloff, E. A. (2007). *Sequence organization in interaction: A primer in conversation analysis I.* New York: CUP.

Shohamy, E. (1994). The validity of direct versus semi-direct oral tests. *Language Testing, 11*, 99-123.

Swain, M. (2001). Examining dialogue: another approach to content specification and to validating inferences drawn from test scores. *Language Testing, 18*, 275-302.

Tarone, E. (1983). On the variability of interlanguage systems. *Applied Linguistics, 4,* 142-164.

Terken, J. & Hirschberg, J. (1994). Deaccentuation of words representing "given" information: effects of persistence of grammatical function and surface position. *Language and Speech, 37,* 125-145.

van Maele, J. (2009, May 6). Re: [LTEST-L] pronunciation and CEF-based assessment. Messages posted to Language Testing Research and Practice mailing list, Archived at *http://lists.psu.edu/cgi-bin/wa?A0=LTEST-L*.

van Weeren, J., & Theunissen, T. J. J. M. (1987). Testing pronunciation: An application of generalizability theory. *Language Learning, 37,* 109-122.

Walker, M., Joshi, A., & Prince, E. F. (1998). *Centering theory in discourse.* New York: OUP

Wennerstrom, A. (1994). Intonational meaning in English discourse. *Applied Linguistics, 15,* 399-421.

Wennerstrom, A. (1998). Intonation as cohesion in academic discourse: A study of Chinese speakers of English. *Studies in second Language Acquisition, 20,* 1-25.

Wennerstrom, A., & Siegel, A. F. (2003). Keeping the floor in multiparty conversations: intonation, syntax, and pause. *Discourse Processes, 36,* 77-107.

Westwood, V. W. & Kaufmann, H. (2001). *Connected Speech.* [CD-ROM]. Victoria, Australia: Protea Textware.

Whiteson, V. (1978). Testing pronunciation in the language laboratory. *ELT Journal, 33,* 30-31.

Wichmann, A., & Blakemore, D. (2006). The prosody-pragmatics interface. *Journal of Pragmatics, 38,* 1537-1541.

Wilds, C. P. (1975). The oral interview test. In R. L. Jones & B. Spolsky (Eds.), *Testing language proficiency* (pp. 29-44). Washington, DC: Center for Applied Linguistics.

Williamson, D. (2010, April 5). Re: [LTEST-L] Any reliable essay e-rater for large scale English testing. Message posted to Language Testing Research and Practice mailing list, Archived at *http://lists.psu.edu/cgi-bin/wa?A0=LTEST-L*.

Wilson, D. & Wharton, T. (2006). Relevance and prosody. *Journal of Pragmatics, 38,* 1559-1579.

Xi, X., Higgins, D., Zechner, K., & Williamson, D. M. (2008). *Automated scoring of spontaneous speech using SpeechRater^{sm} v1.0.* (TOEFL Research Rep. No. RR-08-62). Princeton, NJ: ETS.

Xu, Y. (2005) Speech melody as articulatorily implemented communicative functions. *Speech Communcication, 46,* 220-251.

Yoshida, H. (2004). *An analytic instrument for assessing EFL pronunciation.* Doctoral dissertation, Temple University.

Young, R. (2002). Discourse approaches to oral language assessment. *Annual review of applied linguistics, 22*, 243-262.

Young, R. F. (2000, March). *Interactional competence: challenges for validity.* Paper presented at the Annual Meeting of the American Association for Applied Linguistics, Vancouver, BC, Canada.

Zechner, K., Higgins, D., Xi, X., & Williamson, D. M. (2009). Automatic scoring of non-native spontaneous speech in tests of spoken English. *Speech Communication 51,* 883-895.

Zielinski, B. W. (2006). The intelligibility cocktail: an interaction between speaker and listener ingredients. *Prospect, 3,* 20-43.

Zielinski, B. W. (2008). The listener: no longer the silent partner in reduced intelligibility. *System, 36,* 69-84.