63

Acoustic and Temporal Analysis for Assessing Speaking

Okim Kang
Northern Arizona University, USA

Lucy Pickering
Texas A&M University, Commerce, USA

Introduction

Oral assessment in language learning has received increasing attention among second language acquisition (SLA) researchers. This growing interest is likely a product of the increased interpretability of test scores and potential validity of the scores when linked to real-world criteria (Bonk & Ockey, 2003). However, assessing speaking skill can be more challenging than assessing other skills because of the possible subjective nature of listener comprehension, the complexity of rater reliability, and the validity of the performance itself. Of these challenges, the potential variability in rater judgments has been of particular concern for language assessment as a source of measurement error (e.g., Bachman, Lynch, & Mason, 1995).

A variety of human rater biases are attested to in the perceptions of speaking proficiency, and the speaking assessment may have a limited basis in the linguistic characteristics of the speaker's oral production. Although sophisticated statistical techniques derived from Rasch scaling or generalizability theory (G-theory) can in principle equate practiced ratings which may display different degrees of rigor or leniency among raters (Lumley & McNamara, 1995), a technology-based measurement strategy that compensates for the variation in rater judgments of oral proficiency is much to be desired (Kang, Rubin, & Pickering, 2010). In fact, certain acoustical and temporal features of non-native speakers' (NNSs') pronunciation, measurable by means of instrumentation rather than by listener impressions, can now provide supplementary parameters for "degree of accentedness."

Thanks to advances in speech science, we can readily identify acoustic and temporal features of pronunciation that affect listeners' comprehensibility. That is, computer-assisted instruments can conveniently examine some elements of the physical facts of human utterances. In this chapter, the primary focus lies in

The Companion to Language Assessment, First Edition. Edited by Antony John Kunnan.

DOI: 10.1002/9781118411360.wbcla056

a discussion of instrumental measures with regard to speaking assessment in general, and addresses both temporal (voice time and duration) and acoustic (e.g., fundamental frequency, amplitude, or spectral behavior for intonation in particular) parameters used for various operational constructs of NNS speech evaluation. For this purpose, the constructs of listeners' judgments such as intelligibility, comprehensibility, and accentedness are construed in one broad sense of listeners' evaluation of NNS speech, even though they are addressed separately in the literature (Derwing & Munro, 2005).[1] This broad approach includes listeners' ratings of NNSs' oral proficiency and fluency. This chapter will also address the difference between automated scoring systems and systems such as those discussed thus far that rely on both instrumental and auditory analyses.

One caveat to note initially is that the instrumental analysis can be indeed dependent upon perceptual subjectivity itself to some extent, although it is known to objectively describe and evaluate speech data. (See "Challenges for objective measures of the speech signal in oral assessment" below.) Thus, this chapter posits that the instrumental analysis alone should not be the sole basis for objective interpretation of candidates' scores in speaking assessment, but instead a useful methodology to identify information about a candidate's speech that would contribute to scoring decision making or to assessment rubric development.

Background to Acoustic and Temporal Measures

Perceptual ratings in oral assessment, such as measurement of the percentage of correctly identified words or rating scales using 5-, 7-, or 9-point scales, may suffer from measurement errors due to their dependency on raters' backgrounds, subjectivity, and other social issues (Kang & Rubin, 2009). In our social contexts, up to a quarter of the variance in listener judgment is attributed to factors such as listeners' expectations, attitudes, and stereotypes as opposed to the nature the speech itself (Derwing, Frazer, Kang, & Thompson, forthcoming). An alternative approach to supplement this human rater variability is the application of instrumental analysis which can objectively evaluate candidates' speech. Since computers began to become available to speech researchers in the 1960s, speech analysis research has evolved substantially (Mattingly, 2011). For example, computer-assisted speech analysis (e.g., use of a KayPentax Computerized Speech Laboratory [CSL], http://www.kayelemetrics.com, or freeware such as Praat, http://www.praat.org; see also http://www.fon.hum.uva.nl/praat and http://www.tc.umn.edu/~parke120/praatwebfiles) is becoming more commonplace in the assessment of speech patterns (e.g., Pickering, 2004; Kang et al., 2010).

The instrumental analysis can examine the production of NNS speech at both segmental and suprasegmental levels. While the segmental analysis often focuses on the "accuracy" of NNSs' consonant and vowel formation, the suprasegmental analysis takes account of the role that differences in speaking rates, intonation patterns, and other prosodic features may play in listeners' comprehension. This methodology often incorporates discourse analysis to supplement the instrumental analysis, wherein an analyst identifies a pragmatic context in which a particular intonational contour would be expected (Pickering, 2001). Following the discourse

analysis, computer-based analysis is used to confirm (or disconfirm) that the expected contour does indeed appear at that site in the speech stream. This is especially one of the big methodological differences between computer-programmed automated scoring systems, which are described in the following paragraph, and auditory–instrumental combined analysis. Studies have suggested that features (e.g., pitch range) identified via the combined acoustic analysis explain variance in listeners' judgments of NNS speech (e.g., Kang et al., 2010).

Finally and most recently, instrumentally identified measures are used to help understand the process of automated scoring. This is the latest development in language assessment and testing due to advances in speech recognition and processing technologies (see, e.g., Xi, 2010b). Currently, tests are in some use in the English as a second language (ESL) field: for example, Versant, also known as PhonePass, produced by Ordinate Corporation; and Speech Rater, developed by Educational Testing Services alongside their Internet-based (iBT) Test of English as a Foreign Language (TOEFL) (http://www.ets.org/Media/Tests/TOEFL/pdf/Speaking_Rubrics.pdf). For instance, subscores of Versant tests for reading fluency and repeat fluency are measures of suprasegmental features (timing, pause, or rhythm). However, as Chapelle and Chung (2010) note, the mechanisms that underlie these tests remain largely opaque and unknown to most professionals in second language (L2) assessment, not least because these are commercial not academic ventures. In addition, adopting these automated speech scoring systems still faces various challenges in terms of establishing validity for test score use and decisions made on the basis of automated test scores, or accurately evaluating communicative functions. The lack of adequacy in testing the communicative competence of candidates is an ongoing concern for those who seek a valid means to automatically test and score learner speech.

Acoustic and Temporal Parameters Measured in Assessing Speaking

Various aspects of NNS pronunciation can be considered in listeners' assessments of speaker proficiency. Studies have investigated the impact of acoustic and temporal features on listeners' judgments of NNSs' oral performance (e.g., Kang et al., 2010) or the correlations between objective measures of speech rates and listeners' rating scores (e.g., Munro & Derwing, 2001; Cucchiarini, Strik, & Boves, 2002). In the early 1980s and 1990s, acoustic studies largely compared NNSs' speech production with the patterns of native speakers' (NSs') speech. Gradually, however, studies began to use acoustic and temporal parameters as indicators of listeners' perceptions.

Segmental acoustic parameters include features of accent such as consonants, the Voice Onset Times (VOTs), or vowel formants. VOT refers to the duration of the period of time between the release of a stop consonant and the beginning of voicing. An easy way to visualize VOT is by reference to the waveform of a sound. Figure 63.1 shows a waveform of the word *tie* spoken by the author, an advanced Korean speaker of English. The left vertical line indicates the moment of release of the stop consonant /t/ pronounced as [t^haɪ]. The VOT is about .08 milliseconds

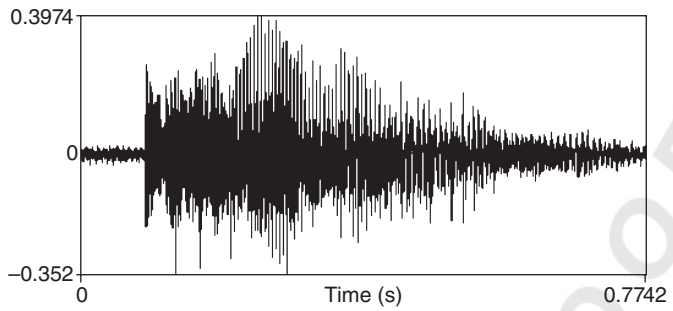

Figure 63.1 The waveform of the word *tie* spoken by an advanced Korean speaker of English

from the spike indicating the release of the stop consonant to the start of the oscillating line indicating the vibration of the vocal folds in the vowel of [aɪ]. An example of the VOT study using NNSs' speech is Flege and Eefting's (1987) research, which compared VOT differences of English stop consonants (e.g., /p/, /t/, /k/) produced by NSs and Spanish L2 speakers. Spanish speakers of English produced shorter VOTs in English initial voiceless stops than did NSs.

In acoustic phonetics, vowels are classified according to particular values called formants, which are a concentration of acoustic energy, that is, a group of overtones corresponding to a resonating frequency of the air in the vocal tract (Ladefoged, 2001). (Examples of the formants are shown as dark voice bars in Figure 63.2.) Accordingly, English vowels are characterized by three formants (F1, F2, and F3) which are used to describe vowel structures. For example, in Wilson, Fujinuma, Horiguchi, and Kazuaki's (2009) study analyzing the speech of low intermediate Japanese speakers, when a consonant /s/ occurs before a high front vowel /i/, it becomes palatalized as in /ʃ/. (i.e., *sea* and *sit* are pronounced as "she" and "shit"). As for the vowel formants, Japanese speakers' F1 value of the low back vowel as in /ɑ/ is way lower than that of NSs. (In Figure 63.2, examples of the F1formant are illustrated as the lowest voice bars.) Overall, using speech analysis programs we can identify the characteristics of individual phonemes, the location of formants, or the presence of voicing.

Numerous studies have investigated the relationships between temporal measures and listeners' judgments of NNS speech (e.g., Trofimovich & Baker, 2006; Isaacs, 2008; Kang, 2010). After Munro and Derwing's (2001) finding, a common belief is that there is a curvilinear relationship between speaking rates and listeners' judgments of L2 comprehensibility and accent. That is, NNS utterances should be somewhat slower than the typical rate for an NS utterance but faster than what L2 learners often produce. Parameters of speaking rates are measured via syllables per second, articulation rate (mean number of syllables per second excluding pauses), phonation–time ratio (percentage of time producing audible speech), and mean length of run (an average number of syllables between pauses). Some or all of these temporal variables often strongly predict L2 performance judgments.

Pauses are an especially important element with regard to speaking rate, and relationships between pausing and speaking assessment have been also widely

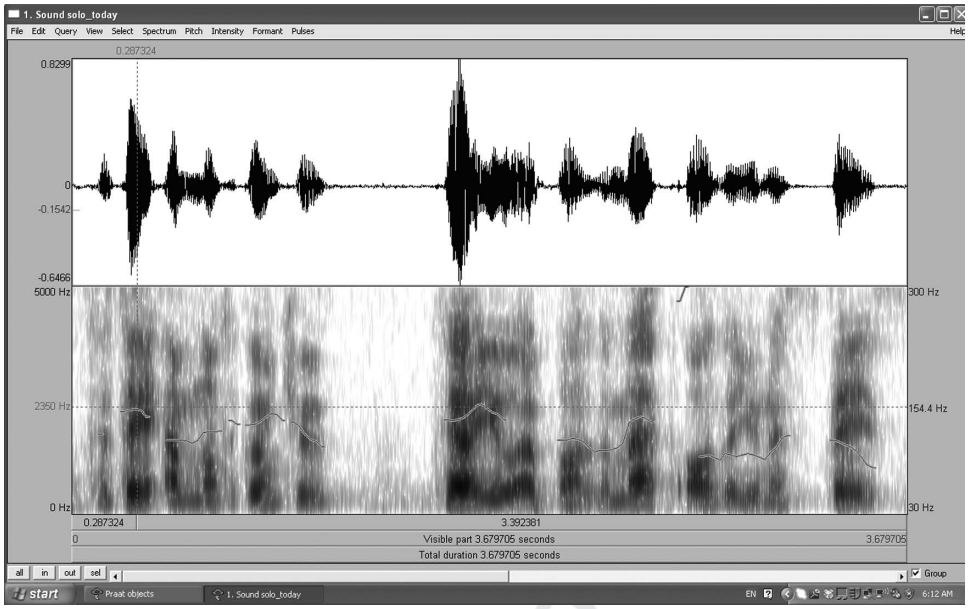

Figure 63.2 A spectrogram showing the waveform (top) and the fundamental frequency (bottom), using speech analysis software Praat, for *Today I'm not going to tell you about map of the United States* spoken by an advanced Chinese speaker of English

investigated. Pauses are measured by variables such as the number, the length, and the location of silent and filled (e.g., *eh* or *um*) pauses. Thus far, pause studies (e.g., Anderson-Hsieh & Venkatagiri, 1994; Kormos & Dénes, 2004) have demonstrated that low proficiency speakers tend to pause frequently and inappropriately, and their pause durations are longer, whereas higher proficiency learners speak faster, with less pausing and fewer unfilled pauses. Methodologically speaking, there has been still an ongoing debate among researchers about the cutoff point of pause length. That is, the cutoff point of silent pauses can vary, as 0.1 second (Anderson-Hsieh & Venkatagiri, 1994), 0.2 (Zeches & Yorkston, 1995), or 0.25 (Towell, Hawkins, & Bazergui, 1996). Terminology-wise, the terms "pauses" and "silences" are often used synonymously in automated scoring systems (e.g., Zechner, Higgins, Xi, & Williamson, 2009). They use "disfluency" as a substitute for the term "filled pause."

Speaking rate and pause measures are often preferred by automatic speech recognition (ASR) systems as objectively measurable parameters which show a high correlation with L2 fluency judgments (Zechner et al., 2009). De Jong and Wempe (2007) provide an example of the relationship between machine-based and human-based coding of temporal measures. In this study, Praat was used to automatically calculate the number of syllables in the utterance based on intensity (the amount of acoustic energy) and pitch peaks. The correlation between the human and automatic speech rate calculations was .71 (see more detail in De Jong & Wempe, 2007). Ginther, Dimova, and Yang (2010) also report robust correlations between temporal variables and other rated measures of oral proficiency; however,

these measures alone did not distinguish adjacent levels in the same way that human raters were able to. The authors add that automated rating systems are thus only able to measure a "narrow sense of fluency" (p. 394).

Prosodic features such as stress and intonation patterns also have a crucial role to play in L2 speaking assessment. First, stress features have been emphasized, as nonstandard word stress has been shown to undermine comprehensibility (Field, 2005). Misplacement of stress in disyllabic words has detrimental effects in speech processing (Cutler & Clifton, 1984). Stress patterns can be obscured in NNS speech production. Low proficiency NNSs often misuse primary stress, placing equal stress on every content word in the unit (Wennerstrom, 2000). In terms of fluency and oral proficiency judgments, advanced L2 learners used stressed words more appropriately than low–intermediate students (Kang, 2008). Acoustic parameters used for these analyses are numbers of stressed words per minute and proportion of stressed words, or the duration of stressed and unstressed syllables.

Non-native intonation patterns, particularly tone choices, have been studied in native listeners' perception of L2 English learners' speech (e.g., Kang et al., 2010). The intonation characteristics of many East Asian speakers may cause US listeners to lose concentration or to misunderstand the speaker's intent (Pickering, 2001). In particular, the choice of a rising, falling, or level pitch on the focused word of a tone unit can affect both perceived information structure and social cues in L2 discourse. A tone unit is a basic unit of intonation known also as tone group, which is a means of breaking up stretches of spoken discourse (Brazil, 1997). Another intonation feature that affects NSs' comprehension of NNSs' speech is pitch range variation. Low proficient/NNSs tend to show a compressed pitch range and a lack of variety in pitch level choices (Wennerstrom, 2000). This contraction of pitch range particularly affects NNSs' ability to indicate the beginning or the end of their discourse. Not surprisingly, this narrow pitch range factor exerts a significant negative effect on proficiency and comprehensibility ratings (Kang et al., 2010).

The intonation-related variables investigated as part of acoustic measures have included tone choices (high rising, high level, high falling, mid rising, mid level, mid falling, low rising, low level, and low falling), pitch-prominent syllables, pitch-nonprominent syllables, and other spoken discourse-related measures. (See "Applications of acoustic analysis and sample analyses" below for a fuller discussion of prominence.) In a study that distinguished these variables, Kang et al. (2010) reported that mid rising and high rising tone choices and pitch range variables were the strongest predictors for NNSs' oral proficiency and comprehensibility ratings.

The physical features listed above along with suggestions from the literature (e.g., Cucchiarini et al., 2002) are used as bases for automated scoring systems. Indeed, the knowledge of acoustic and temporal properties of sound can be helpful for understanding how speech recognition works. Acoustic models exclusively trained on NNS speech can extract these temporal and acoustic features, which are scaled and transformed into fluency and pronunciation scores in the system (Bernstein, van Moere, & Cheng, 2010). For example, the TOEFL Practice Online (TPO) has a set of 11 features for use in the scoring model, whose focus is mainly on fluency, with pronunciation, vocabulary diversity, and grammatical

accuracy added to the mix (Zechner et al., 2009). Among the 11 selected features, 8 deal with fluency aspects (e.g., articulation rate or duration of silence per word) with 1 pronunciation and 2 other language-use features. One of the rationales for choosing these features is high correlations between these and human rating scores, as they are known to represent the overall quality of speech. Nevertheless, the intonation aspect and its interpretation in the pragmatic context are yet to be applied in these automated systems.

Applications of Acoustic Analysis and Sample Analyses

In this section some sample analyses of spoken discourse assessment are presented, using a combination of auditory and instrumental measures (Kang, 2010; Kang et al., 2010). In other words, the subjective auditory perceptions of a human analyst have been combined with the objective instrumental measurements of the speech signal. As noted above, although temporal measures can be fairly successfully scored automatically, crucial prosodic features such as intonation and stress are less easily scored. This is particularly the case when dealing with discourse as opposed to more constrained language samples. Combinations of auditory and instrumental analysis of acoustic features tend to use hardware and software programs such as CSL or Praat for pitch-related measures. As for sole temporal measures, sound editing programs such as Audacity or Soundforge can be employed.

Speech samples are recorded in digital .wav format and transcribed orthographically and prosodically (see Excerpt 1 below). As acoustic parameters are gradient in nature, a range of baseline NS realizations of the features is also measured. As described in Ladefoged's (2001) *A Course in Phonetics*, sound consists of small variations in air pressure that occur rapidly one after another. Actions of the speakers' vocal organs cause these variations, which move through the air somewhat as ripples move on a pond. When these variations reach the ear of a listener, they cause the eardrum to vibrate, which creates sound waves. These waveforms of speech sounds can be readily observed on a computer program such as CSL or Praat.

For analysis, three acoustic indicators are generated: (1) spectrograms, (2) frequency or pitch of fundamental formant (F_0), and (3) intensity (volume of vocalization). A spectrogram is a "graphic representation of sounds in terms of their component frequencies, in which time is shown on the horizontal axis, frequency on the vertical axis, and the intensity of each frequency at each moment in time by the darkness of the mark" (Ladefoged, 2001, p. 276). "Frequency" is a technical term for an acoustic property of a sound. It refers to the number of complete cycles of variation in air pressure occurring in a second. The unit of this frequency measurement is the Hertz (Hz). Figure 63.2 shows a spectrogram of an advanced Chinese speaker's speech, *Today I'm not going to tell you about map of the United States*, using Praat. The upper part of the figure shows the waveform. The fundamental frequency (pitch) is illustrated below. Time is shown on the horizontal axis, and frequency (from 0 to 5,000 Hz on the left and from 30 to 300 Hz on the right) on the vertical axis.

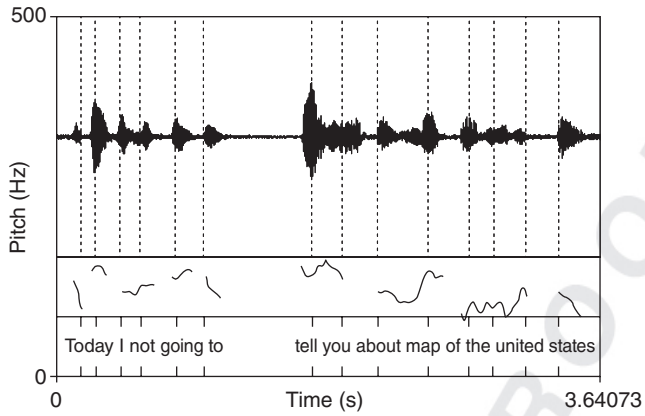

Figure 63.3 An example of the transcription shown for pitch ranges in Praat (Kang, 2010, p. 306). © 2010 with permission from Elsevier, http://www.sciencedirect.com/science/journal/0346251X
Due to the contraction of the spectrogram itself to fit the limited space, the pitch contour and phonological segments may not appear to be exactly parallel.

From the three indicators listed above, plotted against the transcripts of the speech samples, the variables of interest are derived. Figure 63.3 exemplifies a picture of the pitch analysis matched with a script via the Praat freeware program, using the Chinese speaker's speech in Figure 63.2. Note that the pitch of a sound depends on the rate of vibration of the vocal folds. A high pitch sound involves a higher frequency of vibration than a low pitch sound. Different sounds mean that there are differences in pitch, loudness, and quality. Especially, the higher pitch and louder volume (the darkness of the waveform) are represented as prominence (a peak of intonation) syllables.

In Figure 63.3, words such as "toDAy, GOing, TEll, mAp, UNIted, StATes" appear to have received prominence; therefore, they have been transcribed prosodically in capitalized letters. Note that in the final decision on these prominent syllables, the auditory judgments need to be combined with this instrumental analysis. For example, we can calculate the proportion of these prominent words relative to the total number of words. For the pitch range measure, we look at the midpoint of the vowel in the prominent syllable, read F_0 values, and calculate the range of the sample by subtracting the minimum F_0 from the maximum F_0 across the speech sample. In Figure 63.2, the dotted line points at the word "toDAy," of which the F_0 value is 154.4 Hz, shown on the right-hand axis. More examples of variables measures for suprasegemental features are presented in Table 63.1.

Excerpt 1 below shows the prosodic transcription of the same speech sample. (Numbers in parentheses = the length of pauses produced; // = dividing run or tone unit; capital letters = prominent syllables; numbers below the stressed syllables = the F_0 reading of the vowel measured in Hz at the midpoint of the vowel.)

Table 63.1 Selected suprasegmental measures

Measures	*Submeasures*	*Descriptions*
Rate measures	Syllable per second	Mean number of syllables produced per second for the 60-second sample
	Articulation rate	Mean number of syllables produced per minute over total amount of time talking and excluding pause time
	Mean length of run	Average number of syllables produced in utterances between pauses of 0.1 second and above
	Phonation time ratio	Percentage of time spent speaking as a proportion of total time taken to produce the speech sample
Pause measures	Number of silent pauses	Number of silent pauses per 60-second task
	Mean length of silent pauses	Total length of pauses of 0.1 second or greater divided by total number of these pauses
	Number of filled pauses	Number of filled pauses (not including repetitions, restarts, or repairs) per 60-second task.
	Mean length of filled pauses	Average length of filled pauses occurring per 60-second task
Stress measures	Number of prominent syllables per run (pace)	Average number of prominent syllables per run
	Proportion of prominent words (space)	Proportion of prominent words to total number of words
	Prominence characteristics	Proportion of tone units (a run may have more than one unit) that do not contain a nuclear syllable (or final termination)
Pitch measures	Overall pitch range	Pitch range of the sample based on the point of F_0 minima and maxima appearing on prominent syllables per task
	Tone choice	The second measure of discourse-appropriate across-utterance pitch: Each complete unit is counted as comprising either a high, mid, or low termination accompanied by a rising (R), falling (P), or level (O) tone
	Average pitch difference between prominent and nonprominent syllables[a]	Calculated by measuring the F_0 of five prominent and five nonprominent syllables and calculating the average F_0 value for each category
	Average pitch difference between new and given items	Calculated by measuring the F_0 of the same lexical item presented initially as new information and thus appearing in following instances as given information (where possible, five lexical items were used to calculate the average F_0 for each category)

[a] Prominent syllables are divided into two categories based on where they appear in the tone unit. The first prominent syllable is called the *onset*, and the last is called the *tonic* syllable. It is the pitch level and pitch movement on these syllables that form the basis for the assessment of their communicative value within three systems (high, mid, and low). These systems realized on these two syllables (the onset and the tonic syllable) are *key*, realized on the onset syllable, and *termination*, realized on the tonic syllable (Brazil, 1997).

(.10) //todAY I'm not GOing to // (.47) // TEll you about the mAp of the UNIted
154.4 147.2 142. 145.5 124.48
StATes// (.22)
111.3

Combining measures used in a variety of recent studies, Kang et al. (2010) completed a detailed analysis of the speech signal comprising rate, pause, stress, and pitch measures, as shown in Table 63.1.

Challenges for Objective Measures of the Speech Signal in Oral Assessment

It is clear following decades of research that the nature of spoken language proficiency is complex. The studies reviewed here suggest that non-native temporal and intonation patterns account, at least in part, for native listeners' assessment of L2 English learners' speech. In fact, Kang et al. (2010) found that suprasegmental features alone accounted for approximately 50% of the variance in L2 speakers' proficiency ratings. Machine-based acoustic analysis suggests an additional resource to supplement human ratings in the field of language assessment. However, this objective technique still has challenges to overcome.

Acoustic analyses are indeed subject to perceptual limitations. As Crystal (2003) argues, it is important not to become too reliant on acoustic analyses because they rely on accurate calibration of measuring devices and are often open to multiple interpretations:

> Sometimes, indeed, acoustic and auditory analyses of a sound conflict—for example, in intonation studies, one may hear a speech melody as rising, whereas the acoustic facts show the fundamental frequency of the sound to be steady. In such cases, it is for phoneticians to decide which evidence they will pay more attention to; there has been a longstanding debate concerning the respective merits of physical (i.e., acoustic) as opposed to psychological (i.e., auditory) solutions to such problems, and how apparent conflicts of this kind can be resolved. (Crystal, 2003, p. 7)

Possible ways to overcome such limitations include (1) using a combination of auditory and instrumental analysis and (2) checking inter-/intra-analyst reliability to ensure the consistency of the analysis. According to Kang (2010), in suprasegmental analyses, the internal consistency reliability between two phonetic analysts was lower in stress and pitch analyses (.86 or lower), but higher in temporal measures (.95 or higher). Discrepancies between the two analysts took place either in determining the start and end of each pause or in identifying prominent syllables. Therefore, a calibrating procedure having two analysts reach consensus may be required to ensure the reliability of the analysis. What people consider "objective" still relies on the "subjective" nature of listener perception.

Another caveat involves gender difference in acoustic analysis. Due to a gender confounding factor (i.e., male speakers having lower pitch voice than female

speakers in general) especially in intonation measures, some studies tend to use a single gender (e.g., Kang et al., 2010, investigates only male speakers). It is becoming increasingly common to make gender adjustments for pitch before starting any analysis with a different gender. That is, prior to any kind of pitch comparison between male and female voices, the pitch is transformed into semitones (Couper-Kuhlen, 1996).

Differences in spoken genre can result in additional variance to the accuracy of acoustic analysis. Scholars have used various speech stimuli for their analysis: NNSs' oral presentation speech for different proficiency levels (Hincks, 2005); international teaching assistants' in-class lectures (Pickering, 2001; Kang, 2010); iBT TOEFL responses to speaking tasks (Kang et al., 2010); and read vs. spontaneous speech (Cucchiarini et al., 2002). Depending on the types of speech samples used for analysis, speech patterns may appear differently, assuming that test-taker performance varies in response to various tasks (Fulcher, 2003).

When considering the practicality or applicability of an acoustic approach that combines auditory and instrumental analysis, one must take into account the labor intensiveness involved. For a one-minute NNS speech sample, it takes at least 30–45 minutes to identify runs and the location or length of pauses (silent and filled). It takes approximately another 45 minutes to perform the prosodic analysis (i.e., measure fundamental frequency [F_0] for prominent syllables and analyze tone choice).

Acknowledging this labor intensity, automatic speech assessment tools have received growing attention (Franco et al., 2010). However, ASR still faces numerous problems in terms of its accuracy of the measures and feedback (Levis, 2007). Speech recognition systems, at least up until now, seem to offer more accuracy for NS than for NNS speech (Ehsani & Knodt, 1998; see also http://www.speech.sri.com). With accented NNSs' speech, the accuracy of the speech program significantly dropped (95% with NS speech in Ehsani & Knodt, 1998, but 70% in Derwing, Munro, & Carbonaro, 2000). In addition, as the speech recognition systems tend to measure prosody of speech without reference to linguistic organization, the precision problem especially arises with suprasegmental errors (Levis, 2007). For instance, when it comes to tone choice analysis, there is great difficulty in identifying a tone unit especially with the speech of a low proficiency speaker. Following Brazil's (1997) protocol, a tone unit contains one or two prominent syllables, which may coincide with syntactic and pause boundaries. However, low proficient NNSs frequently use primary stress on every word in a message unit, regardless of its function or semantic importance (Wennerstrom, 2000). Their pauses often appear randomly and irregularly. As a result, recognizing tone unit boundaries is not a clear-cut procedure in much NNS speech.

Future Directions

To the degree that conformity to NS comprehensibility constitutes a criterion for oral proficiency, acoustic and temporal parameters measured via instrumentation can help interpret candidates' scores in assessing speaking skills. The knowledge of these instrumentally analyzed properties can be also used for rubric

development or rater training in oral proficiency testing. Currently, descriptors of rubrics used in high stakes testing are still relatively general in terms of describing the pronunciation features in particular. For example, the descriptor for the Delivery dimension in the TOEFL iBT speaking rubric for Score 4 (the highest score of the holistic rating) includes this: "It may include minor lapses, or minor difficulties with pronunciation or intonation patterns" (Educational Testing Service, 2004). Raters may be confused by the term "difficulties with intonation," as it can still be ambivalent when it comes to their decision making. Acoustically identified prosodic features such as pitch range or level (flat) tones can be used as the objects of sensitization in rater training and in developing the assessment criteria those raters will employ.

In addition, the physical properties of the acoustic and temporal measures can build bases for speech recognition and processing techniques, which have increasingly drawn the attention of language testers, as these can help develop automated scoring and feedback systems. Despite some existing drawbacks as listed in the previous section, this objective analysis approach or the combined method with a human rating may also be of use in the automatic assessment of speech production. As topics on ASR effectiveness for NNS speech continue to be of interest to L2 researchers (e.g., Oh, Yoon, & Kim, 2007), the improvement of this approach to speech assessment is certainly necessary.

Acoustic research has yet to be widely applied to the field of assessment of oral performance. In fact, human raters are considered to be more able to decipher meaning from utterances in response to test questions (Godwin-Jones, 2009). Xi (2010a) notes that automatic feedback systems may only "be acceptable in low-stakes practice environments with instructor support" (p. 298). For example, as seen from the set of features used for the TPO (Zechner et al., 2009), the focus of the automatic scoring model is mainly on fluency (temporal features) with some segmental acoustic aspects. Moreover, the ASR models still fall short in that they do not examine the aspects of communicative ability on the part of the candidates. This lack of adequacy in testing the communicative competence of test takers is of ongoing concern for those who seek a valid means to automatically test and score candidates' speech (Chapelle & Chung, 2010). Incorporating more of the acoustic suprasegmental features such as intonation (e.g., tone choices or pitch ranges) into the automated scoring models could help with the issue of communicative competence to some extent, as tones are associated with particular communicative values (e.g., proclaiming with falling tones and referring with rising tones) (Brazil, 1997). Thus, proactive collaborative projects among researchers in language assessment and linguistic analysis are much needed to better develop assessment criteria and to improve assessment training.

Whereas studies have traditionally tended to examine segmentals and suprasegmentals separately, future research may investigate a constellation of acoustic features conjointly for both. This will help to answer the question of the extent to which nonprosodic features of speech contribute to ratings of oral performance, compared to suprasegmentals. In addition, these pronunciation aspects of speech identified through acoustic analysis must be interpreted in conjunction with other linguistic features. That is, further research is necessary regarding whether grammatical and lexical performance variables contribute additional variance to oral

assessment ratings, and the degree to which those other linguistic elements can compensate for dysfunctional features of pronunciation.

The main discussion of this chapter has focused on issues in large-scale assessment. Yet advances made in instrumental analysis and ASR could be used in classroom-based assessment of speech in the future (although somewhat limited at the moment). De Jong and Wempe (2007) provide good evidence of practical application by describing a method to automatically measure speech rate without the need of a transcription, using Praat. The program can quickly identify silence in speech and ultimately provide information on speech rate for learners. The Higgins, Xi, Zechner, and Williamson (2011) study has advanced the technique and built into speech recognizers a component that is able to identify speech rate. A possible scenario is that free downloadable programs such as Praat can be used for formative assessments in which teachers can casually evaluate students' oral fluency development without labor-intensive scoring procedures. How this instrumental analysis can be used in classroom-based speaking assessment is an important topic for future research.

A qualitative approach to acoustic measures may be much needed for future language assessment. Speech evaluation often falls back on quantitative methods such as using data from a large speech corpus to explore the impact of certain acoustic features on listeners' judgments. On the other hand, in-depth interviews or discussions with NNSs (e.g., why they paused at certain locations or why they emphasized certain words) can provide insights into understanding the relationship between NNSs' speech production and listeners' evaluation. This approach will not only help clarify the acoustically identified features of accented speech, but also increase the validity and reliability of the measures.

Overall, the future direction of acoustic studies involves expanding the scope of interpretation of the parameters analyzed for assessing speaking. The features measured instrumentally (i.e., particularly acoustic properties such as tone choices) should be interpreted in a more contextualized way, recognizing the social nature of oral performance through discourse and interaction analysis. Moreover, a sociolinguistic approach may help us find out whether or not the test taker is disadvantaged by his or her interlocutors' particular speech patterns. For example, if an interlocutor does not use rising tones appropriately or frequently, the other interlocutor may feel offended or less supported (Pickering, 2001). Overuse of falling tones by NNSs can give NS listeners an impression of arrogance. Much research needs to be done in this area and to expand the capacity of acoustic research itself. Finally, this chapter has not touched on important sociopolitical issues regarding NNSs' accents, such as identity and motivation, as these are not the main concern of the argument here. Another area of future research should lie in the relationship between the speech properties and physiological traits.

SEE ALSO: Chapter 8, Assessing Pronunciation; Chapter 9, Assessing Speaking; Chapter 72, The Use of Generalizability Theory in Language Assessment; Chapter 77, Multifaceted Rasch Analysis for Test Evaluation; Chapter 80, Raters and Ratings; Chapter 81, Spoken Discourse

Note

1 Unlike intelligibility, which refers to the extent to which a listener understands an utterance, comprehensibility pertains to the degree of difficulty the listener reports in attempting to understand an utterance, and accentedness represents the extent to which an L2 learner's speech is perceived to differ from native speaker norms (Derwing & Munro, 2005).

References

Anderson-Hsieh, J., & Venkatagiri, H. (1994). Syllable duration and pausing in the speech of intermediate and high proficiency Chinese ESL speakers. *TESOL Quarterly*, *28*, 807–12.

Bachman, L. F., Lynch, B. K., & Mason, M. (1995). Investigating variability in tasks and rater judgments in a performance test of foreign language speaking. *Language Testing*, *12*, 238–47.

Bernstein, J., van Moere, A., & Cheng, J. (2010). Validating automated speaking tests. *Language Testing*, *27*(3), 355–77.

Bonk, W. J., & Ockey, G. J. (2003). A many-facet Rasch analysis of the second language group oral discussion task. *Language Testing*, *20*(1), 89–110.

Brazil, D. (1997). *The communicative value of intonation in English*. Cambridge, England: Cambridge University Press.

Chapelle, C. A., & Chung, Y-R. (2010). The promise of NLP and speech processing technologies in language assessment. *Language Testing*, *27*, 301–15.

Couper-Kuhlen, E. (1996). The prosody of repetition: On quoting and mimicry. In E. Couper-Kuhlen & Margret Selting, (Eds.), *Prosody in conversation: Interactional studies* (pp. 366–405). Cambridge, England: Cambridge University Press.

Crystal, D. (2003). *A dictionary of linguistics and phonetics*. Malden, MA: Blackwell.

Cucchiarini, C., Strik, H., & Boves, L. (2002). Quantitative assessment of second language learners' fluency: Comparisons between read and spontaneous speech. *Journal of the Acoustical Society of America*, *111*(6), 2862–73.

Cutler, A., & Clifton, C. F. (1984). The use of prosodic information in word recognition. In H. Bouma & D. G. Bouwhuis (Eds.), *Attention and performance X: Control of language processes* (pp. 183–96). Hillsdale, NJ: Erlbaum.

De Jong, N. H., & Wempe, T. (2007). Automatic measurement of speech rate in spoken Dutch. *ACLC Working Papers*, *2*, 51–60.

Derwing, T., Frazer, H., Kang, O., & Thompson, R. (forthcoming). Accent and ethics: Issues that merit attention. In A. Mahboob & L. Barratt (Eds.), *Examining the CEE in TESOL: English in a multilingual context*.

Derwing, T., & Munro, M. (2005). Second language accent and pronunciation teaching: A research-based approach. *TESOL Quarterly*, *39*, 379–97.

Derwing, T., Munro, M., & Carbonaro, M. (2000). Does popular speech recognition software work with ESL speech? *TESOL Quarterly*, *34*(3), 592–603.

Educational Testing Service. (2004). *TOEFL iBT/Next Generation TOEFL Test independent speaking rubrics (scoring standards)*. Retrieved February 11, 2013 from http://www.ets.org/Media/Tests/TOEFL/pdf/Speaking_Rubrics.pdf

Ehsani, F., & Knodt, E. (1998). Speech technology in computer-aided language learning: Strengths and limitations of a new CALL paradigm. *Language Learning and Technology*, *2*(1), 54–73.

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

Flege, J. E., & Eefting, W. (1987). Cross-language switching in stop consonant perception and production by Dutch speakers of English. *Speech Communication*, *6*, 185–202.

Franco, H., Bratt, H., Rossier, R., Gadde, V. R., Shriberg, E., Abrash, V., & Precoda, K. (2010). EduSpeak: A speech recognition and pronunciation scoring toolkit for computer-aided language learning applications. *Language Testing*, *27*, 401–18.

Fulcher, G. (2003). *Testing second language speaking*. London, England: Pearson.

Ginther, A., Dimova, S., & Yang, R. (2010). Conceptual and empirical relationships between temporal measures of fluency and oral English proficiency with implications for automated scoring. *Language Testing*, *27*, 379–99.

Godwin-Jones, R. (2009). Emerging technologies: Speech tools and technologies. *Language Learning and Technology*, *13*(3), 4–11.

Higgins, D., Xi, X., Zechner, K., & Williamson, D. (2011). A three-stage approach to the automated scoring of spontaneous spoken responses. *Computer Speech and Language*, *25*(2), 282–306.

Hincks, R. (2005). Measures and perceptions of liveliness in student oral presentation speech: A proposal for an automatic feedback mechanism. *System*, *33*, 575–91.

Isaacs, T. (2008). Towards defining a valid assessment criterion of pronunciation proficiency in non-native English-speaking graduate students. *Canadian Modern Language Review*, *64*, 555–80.

Kang, O. (2008). The effect of rater background characteristics on the rating of International Teaching Assistants Speaking Proficiency. *Spaan Fellow Working Papers*, *6*, 181–205.

Kang, O. (2010). Relative salience of suprasegmental features on judgments of L2 comprehensibility and accentedness. *System*, *38*, 301–15.

Kang, O., & Rubin, D. (2009). Reverse linguistic stereotyping: Measuring the effect of listener expectations on speech evaluation. *Journal of Language and Social Psychology*, *28*, 441–56.

Kang, O., Rubin, D., & Pickering, L. (2010). Suprasegmental measures of accentedness and judgments of English language learner proficiency in oral English, *Modern Language Journal*, *94*, 554–66.

Kormos, J., & Dénes, M. (2004). Exploring measures and perceptions of fluency in the speech of second language learners. *System*, *32*, 145–64.

Ladefoged, P. (2001). *A course in phonetics*. Orlando, FL: Harcourt.

Levis, J. (2007). Computer technology in teaching and researching pronunciation. *Annual Review of Applied Linguistics*, *27*, 184–202.

Lumley, T., & McNamara, T. F. (1995). Rater characteristics and rater bias: Implications for training. *Language Testing*, *12*, 54–71.

Mattingly, I. G. (2011). *A short history of acoustic phonetics in the U.S.* Retrieved February 11, 2013 from http://www.haskins.yale.edu/Reprints/HL1144.pdf

Munro, M. J., & Derwing, T. M. (2001). Modeling perceptions of the accentedness and comprehensibility of L2 speech: The role of speaking rate. *Studies of Second Language Acquisition*, *23*, 451–68.

Oh, Y. R., Yoon, J. S., & Kim, H. K. (2007). Acoustic model adaptation based on pronunciation variability analysis for non-native speech recognition. *Speech Communication*, *49*, 59–70.

Pickering, L. (2001). The role of tone choice in improving ITA communication in the classroom. *TESOL Quarterly*, *35*(2), 233–55.

Pickering, L. (2004). The structure and function of intonational paragraphs in native and nonnative speaker instructional discourse. *English for Specific Purposes*, *23*(1), 19–43.

Towell, R., Hawkins, R., & Bazergui, N. (1996). The development of fluency in advanced learners of French. *Applied Linguistics*, *17*, 84–119.

Trofimovich, P., & Baker, W. (2006). Learning second-language suprasegmentals: Effect of L2 experience on prosody and fluency characteristics of L2 speech. *Studies in Second Language Acquisition 28*, 1–30.

Wennerstrom, A. (2000). The role of intonation in second language fluency. In H. Riggenbach (Ed.), *Perspectives on fluency* (pp. 102–27). Ann Arbor, MI: University of Michigan Press.

Wilson, I., Fujinuma, J., Horiguchi, N., & Yamauchi, K. (2009, October). *Acoustic analysis of the English pronunciation of Japanese high school teachers and university students*. Presented at the 158th meeting of the Acoustical Society of America, San Antonio, TX.

Xi, X. (2010a). Automated scoring and feedback systems: Where are we and where are we heading? *Language Testing*, *27*, 291–300.

Xi, X. (Ed.). (2010b). *Language Testing*, *27*(3). (Special issue on the automated scoring of writing and speaking tests).

Zeches, J. T., & Yorkston, K. M. (1995). Pause structure in narratives of neurologically impaired and control subjects. *Clinical Aphasiology*, *23*, 155–4.

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*, 883–95.

Suggested Readings

Derwing, T. M., & Munro, M. J. (2009). Putting accent in its place: Rethinking obstacles to communication. *Language Teaching*, *42*, 1–15.

Munro, M. J. (1995). Nonsegmental factors in foreign accent: Ratings of filtered speech. *Studies in Second Language Acquisition*, *17*, 17–34.

Neumeyer, L., Franco, H., Weintraub, M., & Price, P. (1996). Automatic text-independent pronunciation scoring of foreign language student speech. *Proceedings of ICSLP 96*, 1457–60.

Swerts, M., & Geluykens, R. (1994). Prosody as a marker of information flow in spoken discourse. *Language and Speech*, *37*, 21–43.

Teixeira, C., Franco, H., Shriberg, E., & Precoda, K. (2000). Prosodic features for automatic text-independent evaluation of degree of nativeness for language learners. Retrieved February 11, 2013 from http://www.speech.sri.com/people/hef/papers/prosodic_feat_icslp2000.pdf

Wennerstrom, A. (2001). *The music of everyday speech: Prosody and discourse analysis*. Oxford, England: Oxford University Press.