Session 1: Welcome and introduction

Speaking assessment: Evolving and adapting to a changing world

Evelina D Galaczi, Cambridge English Language Assessment

The evolution of the assessment of speaking ability throughout the last century has been shaped by two competing forces – one expansive and the other reductionist. Theoretical developments in the fields of second/foreign (L2) language pedagogy and assessment, as well as external socio-political factors, have played an expansive role vis-à-vis the conceptualisation of the speaking construct. Conversely, psychometric concerns and considerations of practicality have played a reductionist role in limiting the status of speaking assessment in test batteries or in narrowing down the construct. This presentation will trace the development of speaking assessment from the late 19th to early 21st century, focusing on key tests and the pedagogic, theoretical and empirical forces which have shaped them.

Pedagogic developments at the end of the 19th century

The emergence of the Direct Method and Reform Movement in classrooms in the 1870s and 1880s brought about an interest in the use of language. Key questions and assumptions shifted and the belief espoused by the Grammar-Translation Method that speaking should not be tested was transformed into: *'How should speaking be tested?'* The Certificate of Proficiency in English exam was introduced in such a theoretical climate in 1913 by the University of Cambridge Local Examinations Syndicate. This was a milestone in the evolution of the speaking construct, since the test included an obligatory speaking component which comprised a Conversation task alongside other more traditional tasks such as dictation and reading aloud. This test signalled an engagement with direct assessment of speaking, albeit at the time narrowly defined as pronunciation accuracy.

The growing role of face-to-face speaking tests in the 20th century

The embracing of speaking assessment in the UK during the first three decades of the 20th century contrasted with scepticism in the US about the theoretical desirability of testing speaking ability, largely due to the perceived subjectivity of the assessment and the practical difficulties of administering speaking tests. The socio-economic climate after World War Two, however, highlighted the need for English speakers and marked the growing role of English as a world language. This, in turn, played an expansive role regarding the conceptualisation of speaking assessment on both sides of the Atlantic and led to the growing acceptance and use of face-to-face speaking tests with a conversational component. Key manifestations of this change,

e.g. the introduction of the Lower Certificate in English at Cambridge in 1939, the launch of the United States Army Specialised Training Programme in 1943, and the Foreign Service Interview (FSI) in 1952 developed to assess the readiness of American Foreign Service personnel to communicate in real-life assignments abroad, placed speaking proficiency at the core of L2 proficiency. They also introduced some key validity-related developments, the main being an explicit conceptualisation of the speaking construct at different proficiency levels.

As face-to-face speaking tests became more widespread, considerations of reliability and rater effects (first voiced by Edgeworth in 1888) gained more importance. In the 1940s in the UK Jack Roach, one of the key figures behind the Cambridge Proficiency and Lower Certificate exams, addressed the fundamental psychometric issues of reliability and validity through his concern with keeping standard comparable across individual examiners. The FSI team faced similar issues in the 1950s, and the FSI assessment scale signalled an important psychometric approach to defining a criterion external to rater intuition and test candidature itself.

The communicative paradigm and growing awareness of test authenticity

The communicative movement in the 1970s gave face-to-face speaking test a strong pedagogic and theoretical impetus. Influential ideas regarding the role of performance vis-à-vis competence and the role of context were offered in the 1970s by Hymes (1974), Halliday (1975) and Van Dijk (1977), and paved the way for models of Communicative Competence and Communicative Language Ability (Canale and Swain 1980, Bachman 1990) in the 1980s and 1990s. These frameworks provided a theoretical basis for the speaking construct and led to an interest in assessing the functional and communicative aspect of language. Test tasks took on a real-life purpose. Theoretically and empirically, test authenticity took centre stage, aided by the introduction of qualitative research methodologies in speaking test research in the 1990s. Following the now classic appeal by Van Lier (1989) to look ”inside” the language proficiency interview, i.e. to analyse the discourse produced in speaking tests, a key question emerged: *Is the interaction elicited in speaking tests sufficient to assess communicative competence?*

Paired and group tests: expanding the speaking test construct

The introduction of paired and group speaking tests in the 1990s signalled a further expansion in construct conceptualisation. The paired/group format presented a range of interaction possibilities, including peer-peer interaction tasks, where the conversational rights and responsibilities of the participants were more balanced and a wider spectrum of functional competence could be sampled. Co-construction of interaction and the dynamic two-way influence of the interlocutors, therefore, vastly broadened the construct underlying paired and group speaking tests. The speaking test construct now drew not just on communicative theoretical frameworks, but also on frameworks of interactional competence (Kramsch 1986) and pushed construct definition of speaking tests into a new and broader conceptual terrain. The conceptualisation moved beyond a view of language competence as residing within

an individual to a more social view where communicative language ability and the resulting performance reside within a social and jointly-constructed context. At the same time, empirical findings indicated that individual interviewing techniques and the background variables which interlocutors (both test takers and examiners) bring to the speaking test could affect a candidate's performance. A key construct-related question emerged: *Is the variability inherent in interaction construct-irrelevant variance (and therefore to be avoided) or is it part of the speaking construct?*

The influence of technology

The 1970s saw the growing use of computer-delivered tests, which a few decades later were followed by computer-scored tests where the assessment of speaking process was completely automated. The advent of computer-based language testing highlighted a key question: *How does the delivery medium change the nature of the construct being evaluated?* On the one hand, computer-based speaking tests were seen as addressing a variety of caveats associated with the human factor in direct assessment of speaking, such as providing uniformity of administration, as well as the practical benefit of making the costs manageable. On the other hand, they also narrowed down the conceptualisation of the speaking test construct from a socio-cognitive construct definition in face-to-face tests, where speaking is viewed both as a cognitive trait and a social interactional one (Taylor 2011), to a psycholinguistic definition which places emphasis on the cognitive dimension of speaking (Van Moere 2012).

A glimpse into the future

Throughout the last century language testers have had at their disposal a range of speaking test formats which have been useful in eliciting and assessing speaking skills. All of these formats bring their strengths and caveats and have varied applicability for different contexts. No test format is inherently superior, and the issue of *fitness for purpose* has emerged as fundamental in driving discussions and decisions about underlying test constructs. The key question informing current debates about speaking assessment, therefore, is not whether to use a certain format or not, but: *In what contexts is a speaking test fit for purpose?* This question will continue to inform future debates as new developments in technology and natural language processing research expand the existing array of speaking tests and lead to a symbiotic relationship between technology and humans in speaking assessment.

References

Bachman, L (1990) *Fundamental considerations in language testing,* Oxford: Oxford University Press.

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

Edgeworth, F Y (1888) The statistics of examinations, *Journal of the Royal Statistical Society* 51**,** 599-635.

Halliday, M A K (1975) *Learning how to mean,* London: Edward Arnold.

Hymes, D (1974) *Foundations in sociolinguistics: An ethnographic approach*.
Kramsch, C (1986) From language proficiency to interactional competence, *The Modern Language Journal* 70 (4), 366-372.
Taylor, L (2011) Introduction, in Taylor, L (Ed.) *Examining speaking*, Cambridge: University of Cambridge ESOL Examinations Examinations/Cambridge University Press.
Van Dijk, T A (1977) *Text and context: Explorations in the semantics and pragmatics of discourse,* London: Longman.
Van Lier, L (1989) Reeling, writhing, drawling, stretching and fainting in coils: Oral proficiency interviews as conversations, *TESOL Quarterly* 23, 480-508.
Van Moere, A (2012) A psycholinguistic approach to oral language assessment, *Language Testing* 29 (3), 325-344.

Session 2: State of the art (I)

Corpus evidence and the lexicogrammar of speaking

Ute Römer, Georgia State University

The past few decades have witnessed a massive increase in corpus research activity in a range of linguistic subfields, including strands within Applied Linguistics. Corpora are increasingly accepted as powerful tools that help us gain insights into language structure and use, and help inform language teaching and testing practice (see, for instance, Flowerdew 2012, Hawkins and Filipović 2012, Reppen 2010, and Römer 2011). This paper discusses the importance of considering corpus evidence in highlighting central aspects of spoken language and addresses the question "How can corpus tools and techniques help us shed light on the concept of speaking?"

Since spoken language is not a uniform phenomenon but varies considerably depending on the context of use, the paper does not attempt to describe speech 'in general'. Instead, it focuses on one particular, more specialized type of language: spoken English produced in a US research university setting. This type of language is captured in MICASE, the Michigan Corpus of Academic Spoken English (Simpson, Briggs, Ovens and Swales 2002), a collection of 152 transcripts and 1.8 million words, based on 200 hours of recordings of speech events from across the University of Michigan in Ann Arbor.

The paper starts out with a brief analysis of frequency word and keyword lists of academic speaking (compared to academic writing), including observations on Zipfian profiles (Zipf 1935), and then focuses on phraseological items (variably referred to as n-grams, formulaic sequences, lexical bundles, clusters, etc.) that are particularly common in speaking and carry important discourse functions. Software packages for corpus access and analysis are used to extract lists of contiguous word sequences (n-grams, e.g. *you know*, *a lot of*) and non-contiguous word sequences (phrase-frames, e.g. *a * of*, *I don't * so*) of different lengths from MICASE. The resulting lists are filtered manually for items that play a central role in academic speech and appear to have a particularly high communicative value.

The final section of the paper reviews rating scales of a selection of high-stakes speaking tests and discusses in how far these rating scales capture central aspects of spoken language as highlighted by corpus analysis. It then discusses implications of our MICASE-based findings for (academic) speaking assessment. In the light of corpus findings, the paper challenges the prevalent separation of vocabulary and syntax in assessment criteria. It questions whether scoring criteria such as "Grammatical Resource" and "Lexical Resource" (UCLES 2012: 64) can and should actually be kept separate in assessing speaking proficiency. Overall, the paper provides evidence for the interrelatedness of vocabulary and grammar in academic speech and stresses the importance of phraseology as a core, rather than a peripheral aspect of language (see Ellis 2008), adding to a growing body of existing work in corpus research on phraseology (see e.g. Biber 2009, Hoey 2005, O'Donnell, Römer and Ellis 2013, Römer 2009, 2010, Sinclair 2008). It demonstrates how corpus analysis can contribute to a better understanding of the real-world

spoken lexicogrammar and how it helps us uncover the patterned nature of speaking.

References

Biber, D (2009) A corpus-driven approach to formulaic language in English: Multi-word patterns in speech and writing, *International Journal of Corpus Linguistics* 14(3), 275–311.

Ellis, N C (2008) Phraseology: The periphery and the heart of language, in Meunier, F and Granger, S (Eds.), *Phraseology in Language Learning and Teaching* (pp. 1-13), Amsterdam: John Benjamins.

Flowerdew, L (2012) *Corpora and language education,* London: Palgrave Macmillan.

Hawkins, J A and Filipovic, L (2012) *Criterial features in L2 English,* Cambridge: Cambridge University Press.

Hoey, M (2005) *Lexical priming: A new theory of words and language,* London: Routledge.

O'Donnell, M B, Römer, U and Ellis, N C (2013) The development of formulaic sequences in first and second language writing: Investigating effects of frequency, association, and native norm, *International Journal of Corpus Linguistics* 18(1), 83-108.

Reppen, R (2010) *Using corpora in the language classroom,* Cambridge: Cambridge University Press.

Römer, U (2009) The inseparability of lexis and grammar: Corpus linguistic perspectives, *Annual Review of Cognitive Linguistics* 7, 140-162.

Römer, U (2010) Establishing the phraseological profile of a text type: The construction of meaning in academic book reviews, *English Text Construction* 3(1): 95-119. [Reprinted in Biber, D & Reppen, R (Eds.) (2012) *Corpus linguistics. Volume I: Lexical studies,* London: SAGE Publications.]

Römer, U (2011) Corpus research applications in second language teaching, *Annual Review of Applied Linguistics* 31, 205-225.

Sinclair, J M (2008) The phrase, the whole phrase, and nothing but the phrase, in Granger, S and Meunier, F (Eds.), *Phraseology: An interdisciplinary perspective* (pp. 407-410), Amsterdam: John Benjamins.

Simpson, R C, Briggs, S L, Ovens, J and Swales, J M (2002) *The Michigan Corpus of Academic Spoken English,* Ann Arbor, MI: The Regents of the University of Michigan.

UCLES (2012) *Cambridge English: Advanced. Certificate in Advanced English (CAE). Handbook for teachers,* Cambridge: University of Cambridge ESOL Examinations.

Zipf, G K (1935) *The Psycho-biology of language: An introduction to dynamic philology*. Cambridge, MA: The M.I.T. Press.

Pronunciation

Talia Isaacs, University of Bristol

From a historical perspective, it can be argued that pronunciation, more than any other component within the broad construct of second language (L2) speaking ability, has been subject to the whims of the time and the fashions of the day. That is, pronunciation, once dubbed "the Cinderella of language teaching" to depict its potentially glamorous yet marginalised existence (Kelly 1969:87), experienced a fall from grace after being a focal point of L2 instruction and teacher literacy training during its heyday-a prime example of a pendulum swing in L2 teaching methodologies and practices (Gass 1996) that have affected substantive coverage for learners in L2 classrooms, often with detrimental effects for stakeholders (e.g. Morley 1991). Naturally, the aspects of L2 speech (pronunciation) that are ascribed pedagogical value in the minds of teachers and researchers have shifted over time (Munro and Derwing 2011). However, an aerial view of developments over the past century reveals the polarised nature of researchers' and practitioners' beliefs on the relative importance of pronunciation in L2 aural/oral instruction and assessment. Although there are signs that pronunciation is increasingly acknowledged as an important component of L2 speaking ability by the applied linguistics community (Derwing and Munro 2009), is gradually being reintegrated into the L2 classroom (Isaacs 2009), and is included as an assessment criteria in the scoring rubrics of several prominent L2 speaking scales, often encapsulated by the term "intelligibility" or "comprehensibility" (Isaacs and Trofimovich 2012), such debates continue today (e.g. Brinton and Butler 2012).

In order to take stock of trends in L2 pronunciation research, teaching, and assessment over the past century within the broader context of developments in L2 speaking assessment, this paper will report on the state-of-the-art by first overviewing the ways in which pronunciation instructional priorities and, by implication, assessment targets, have shifted over time. The reorientation of the pedagogical goal in pronunciation teaching from the traditional focus on pronunciation accuracy (accent reduction) to the more suitable goal of intelligibility in light of the target use domain in the case of the vast majority of L2 learners (test takers), will feed into a discussion on major constructs subsumed under the umbrella term of "pronunciation" or that are often cited in L2 pronunciation research. Particular emphasis will be placed on theoretical gaps, definitional quagmires, and challenges in adequately operationalising the focal construct in assessment instruments for operational testing purposes and on implementation challenges in L2 classrooms. Finally, the paper will conclude by setting out an agenda for further research in light of globalising trends, technological advances, and the need to examine pronunciation performance on more interactional task types (e.g., the paired or group speaking test format) than have traditionally been researched in the psycholinguistically-orientated L2 pronunciation literature.

Historical overview

Pronunciation (phonetics) was heralded as foundational to L2 teaching and teacher

training by proponents of the Reform Movement at the turn of the 20th century (e.g. Sweet 1899), with phonetic transcriptions emphasised as obviating the need for a native speaking teacher to model the accurate production of target language sounds. Segmental features (vowel and consonants) continued to be highlighted in instructional materials well beyond the first half of the 20th century, including in Lado's *Language Testing* (1961), which featured chapters on testing the perception and production of segments, word stress, and intonation. This work has been unparalleled in its focus on practical issues in pronunciation item design, test administration, and scoring, with the overall goal of systematically testing those features hypothesised to minimise first language influence on target language performance. Several of the challenges that Lado (1951) highlighted over six decades ago, including that "present practice in oral-aural tests shows lack of workable criteria of what is meant by pronunciation and the role it plays in speaking and listening" and that "pronunciation does not appear to have been taken into account systematically" in rating scales, still resonate today and need to be addressed from a research and practical standpoint (1951:531).

Despite the pivotal role of pronunciation in Lado's (1961) seminal book, which is often taken to represent the birth of language testing as its own academic discipline (e.g. Spolsky 1995), the focus on pronunciation in language testing appears to have been short-lived. During periods in which the teaching methods that were often closely associated with pronunciation (e.g. decontextualised drills symbolising rote-learning) ran contrary to the mainstream intellectual currents of the time, pronunciation tended to be either shunned or ignored in applied linguistics circles. Specifically, the naturalistic approaches to teaching that emerged in the late 1960s at the onset of the communicative era and continued into the 1980s, de-emphasised pronunciation in instruction, viewing it as ineffectual or even counterproductive in fostering the acquisition of the target language (e.g. Terrell 1989). Thus, pronunciation fell drastically out of vogue for several decades, the repercussions of which are evidenced in selected publications by pronunciation proponents from 1990 onwards citing the "neglect" of pronunciation in English language teaching and learning (e.g. Rogerson and Gilbert 1990). This discourse of neglect persists today but has been absent in the area of pronunciation assessment in particular, where, until recently, L2 pronunciation has had few advocates deploring its marginalisation as an assessment criteria in L2 speaking tests or drawing attention to its exclusion from the collective research agenda (Isaacs in press). A case in point is that during the first 25 years of publications in the journal, *Language Testing*, research on pronunciation assessment was practically nonexistent, with only two pronunciation-focused articles appearing during that time period (1984-2009). In addition, although the field of language testing has arguably moved beyond Lado's (1961) skills-and-components model as the dominant theoretical view (Bachman 2000), "phonology/graphology" in Bachman (1990) Bachman and Palmer's (1996) highly influential Communicative Language Ability framework seems to have been a direct carry-over from Lado and does not appear to have been reconceptualised since that time.

However, there is reason for optimism. Pronunciation has arguably

experienced a modest, if, as yet piecemeal resurgence of attention among language assessment researchers against the backdrop of support from a small but increasingly organised applied linguistics community with an interest in L2 pronunciation teaching and learning (Isaacs in press). To date, three articles, which directly centre on issues and challenges in assessing L2 pronunciation, have appeared in *Language Testing* since 2010, others have appeared in *Language Assessment Quarterly*, and pronunciation has additionally been explicitly discussed in other articles published during this period on automated assessment or the assessment of L2 speaking and listening more generally. Finally, the inclusion of pronunciation in the state-of-the-art on the speaking construct at the 2013 *Cambridge Centenary Speaking Symposium* implies that pronunciation is, indeed, viewed as an integral part of the construct of L2 speaking, and is a positive sign.

References

Bachman, L F (1990) *Fundamental considerations in language testing,* Oxford: Oxford University Press.

Bachman, L F, and Palmer, A S (1996) *Language testing in practice,* Oxford: Oxford University Press.

Bachman, L F (2000) Modern language testing at the turn of the century: Assuring that what we count counts, *Language Testing,* 17(1), 1–42.

Brinton, D M and Butler, H (2012) The ethics of pronunciation instruction, *Special Research Symposium Issue of CONTACT,* 38(2), 76-89.

Derwing, T M and Munro, M J (2009) Putting accent in its place: Rethinking obstacles to communication, *Language Teaching,* 42(3), 1–15.

Gass, S M (1996) Second language acquisition and linguistic theory: The role of language transfer, in Richie, W C and Bhatia, T K (Eds.), *Handbook of second language acquisition* (pp. 384-403), San Diego, CA: Academic Press.

Isaacs, T (2009) Integrating form and meaning in L2 pronunciation instruction, *TESL Canada Journal,* 26(2), 1–12.

Isaacs, T (in press) Assessing pronunciation, In Kunnan, A J (Ed.), *The companion to language assessment*. Hoboken, NJ: Wiley-Blackwell.

Isaacs, T and Trofimovich, P (2012) "Deconstructing" comprehensibility: Identifying the linguistic influences on listeners' L2 comprehensibility ratings, *Studies in Second Language Acquisition,* 34(3), 475–505.

Kelly, L G (1969) *25 centuries of language teaching: An inquiry into the science, art, and development of language teaching methodology, 500 B.C.-1969,* Rowley, MA: Newbury House.

Lado, R (1951) Phonemics and pronunciation tests, *The Modern Language Journal,* 35(7), 531-542.

Lado, R (1961) *Language testing: The construction and use of foreign language tests,* London: Longman.

Morley, J (1991) The pronunciation component of teaching English to speakers of other languages, *TESOL Quarterly,* 25(3), 481–520.

Munro, M J and Derwing, T M (2011) The foundations of accent and intelligibility in pronunciation research, *Language Teaching,* 44(3), 316-327.

Rogerson, P and Gilbert, J B (1990) *Speaking clearly,* Cambridge: Cambridge University Press.

Spolsky, B (1995) *Measured words: The development of objective language testing,* Oxford: Oxford University Press.

Sweet, H (1899) *The practical study of languages: A guide for teachers and learners,* London: Dent.

Terrell, T D (1989) Teaching Spanish pronunciation in a communicative approach, in Bjarkman, P C and Hammond, R M (Eds.), *American Spanish pronunciation: Theoretical and applied perspectives* (pp. 196–214), Washington, DC: Georgetown University Press.

Session 3: State of the art (II)

Fluency

Glenn Fulcher, University of Leicester

Fluency has always been challenging to define and operationalize in speaking tests. Yet, it persists as a concept with which teachers are comfortable, and a construct that language testers find irresistible. Brumfit (1984) characterised fluency as existing in a relationship of polarity with 'accuracy', describing it as "natural language use" (1984:56). Drawing on earlier work by Fillmore (1979), he characterized fluency as (a) "filling time with talk", which implies automaticity of language processing; (b) the production of coherent sentences using the "semantic and syntactic resources of the language" appropriately; (c) selecting appropriate content for context; and (d) being creative with the language. When everything comes together with appropriate world knowledge and suitable personality, a speaker is rarely "lost for words" and does not become "tongue tied". We get the impression that speech "flows". Koponen and Riggenbach (2000) deconstruct the metaphorical understanding of fluency as "language as motion"; fluid like liquid, or flowing like a river. Speech is said to be "smooth, rapid and effortless", rather than "choppy" (Chambers 1997).

This metaphor reveals that fluency is situated partially in the production of the speaker, and partially in the perception of the listener (Freed 2000). In language testing the construct is articulated in rating scale descriptors, critiques of which became common in the 1980s (Lantolf and Frawley 1985, 1988). Based primarily on principles of internal coherence of the "more than/less than" type, qualifiers such as "*undue* hesitation" and "*excessive* pausing" also invite listener comparison with some internalised abstraction of an ideal "native speaker" (Davies 2004). In an attempt to overcome these problems Fulcher (1987, 1993, 1996) advocated the development of data-based scales to generate descriptors that were grounded in learner performance. The top and bottom descriptors were now defined by the most and least fluent performances on test tasks, and descriptors were generated from the analysis of observable performance features.

Fluency research has since broadly taken two paths: the cognitive science route and the linguistic route. Both aspire to describe the observable features of fluency in speech. After all, fluency is a construct, and must have observable elements that together define that construct for it to be measurable. Both wish to understand what raters pay attention to when making judgments about fluency. However, they differ in one critical respect. Cognitive science researchers add "L2 cognitive fluency" to the mix (Segalowitz 2010:76). It is said to be the cause of the observable features of (dys)fluency in performance. Researchers therefore wish to discover (a) which features most impact on the perceived fluency of a listener, and (b) which features can be predicted by variation in cognitive fluency. Each observable feature is therefore treated as a variable capable of objective measurement that *does not in itself need interpretation*, but is explained in terms of the effects of another construct. The practical effect is that counts of observations may be fed directly into regression models following removal of outliers and data

normalization.

The linguistic school accepts the relevance of cognitive speech models for understanding language processing and production (Levelt 1989, 1999, Field 2011). However, the explanation for all surface phenomena is not necessarily cognitive. The cognitive school pays scant attention to language as a means of communication that is largely conditioned by social interaction. From the linguistic perspective it is argued that there is no single explanation for any dysfluency feature or change in speech rate. Nor is there one-to-one mapping with components of an L2 cognitive fluency model. These features can be measured, but in some contexts they will be perceived as dysfluency, and in others as quite fluent. Speakers deliberately use silence to communicate their impressions, attitudes, emotions, and intentions (Nakane 2007, Bruneau 2008), and listeners interpret speed and pauses in these terms. Pauses are also part of our turn-management toolkit (McCarthy, 2010), and a politeness mechanism (Scollon and Scollon 1989). They are a vehicle for expressing our personality, establishing social status, and injecting suspense or comic relief into utterances (Nakane 2012). The argument from the linguistic school is that surface features must be interpreted in context, as their manifestation will vary according to task and task features (Skehan 2009) as well as speaker intention.

The choice of approach also has profound implications for how we assess fluency. If one holds that cognitive fluency can "… serve as a stand-in measure of general proficiency and L2 experience" (Segalowitz 2010:76), no interpretation of observable features is necessary. They become the indirect measure of L2 proficiency by virtue of their relation with cognitive fluency, and can be measured by current computer technology (De Jong and Wempe 2009, Little et al 2013). Tasks like sentence repetition, read-aloud, and sentence building (Van Moere 2012) provide the data necessary for the psycholinguistic/cognitive inferences that are said to allow extrapolation to claims about likely candidate performance beyond the test. Alternatively, if language performance is more than a cognitive process that involves interactive responsiveness and the use of a rich repertoire of linguistic choices to express meaning, the use of human judgment seems to be an inevitable requirement.

In this talk I consider the range of fluency/dysfluency phenomena that have been studied, and summarize key findings. I discuss whether the cognitive or linguistic path provides the most convincing interpretation of data in the light of models of speech processing. I then look at how fluency has been articulated in rating scales to show that these have been more successful than is commonly thought to be the case. I conclude with a rationale for persisting with the use of human ratings in assessing spoken fluency.

References

Brumfit, C (1984) *Communicative methodology in language teaching: The roles of fluency and accuracy*, Cambridge: Cambridge University Press.

Bruneau, T J (2008) How Americans use silence and silences to communicate, *China Media Research*, 4(2), 77-85.

Chambers, F (1997) What do we mean by fluency? *System 24*(4), 535-544.

Davies, A (2004) The native speaker in applied linguistics, in Davies, A and Elder, C (Eds.), *The Handbook of Applied Linguistics*, Oxford: Blackwell, 431-450.
De Jong, N and Wempe, T (2009) Praat script to detect syllable nuclei and measure speech rate automatically, *Behavior Research Methods,* 41,385-390.
Field, J (2011) Cognitive Validity, In Taylor, L. (Ed.) *Examining speaking: Research and practice in assessing second language speaking,* Cambridge: UCLES/Cambridge University Press, 65-111.
Fillmore, L W (1979) Individual differences in second language acquisition, in Fillmore, C, Kempler, D and Wang, W Y S (Eds) Individual Differences in Language Ability and Language Behavior, New York: Academic Press.
Freed, B F (2000) Is fluency, like beauty, in the eyes (and ears) of the beholder? In Riggenbach, H (Ed.), *Perspectives on fluency*, Ann Arbor: University of Michigan Press, 243-265.
Fulcher, G (1987) Test of oral performance: The need for data-based criteria, *English Language Teaching Journal*, 41(4), 287-291.
Fulcher, G (1993) The Construction and Validation of Rating Scales for Oral Tests in English as a Foreign Language, unpublished PhD thesis, University of Lancaster, UK.
Fulcher (1996) Does thick description lead to smart tests? A rating- scale approach to language test construction, Language Testing 13 (2), 208–38.
Koponen, M and Riggenbach, H (2000) Overview: Varying perspectives on fluency, in Riggenbach, H (Ed.), *Perspectives on fluency,* Michigan: University of Michigan Press, 5-24.
Lantolf, J P and Frawley, W (1985) Oral proficiency testing: A critical analysis, *The Modern Language Journal,* 69, 337-345.
Lantolf, J P and Frawley, W (1988) Proficiency: Understanding the construct, *Studies in Second Language Acquisition*,10, 181-196.
Levelt, W J M (1989) Speaking, Cambridge, MA: MIT Press.
Levelt, W J M (1999) Language production: a blueprint of the speaker, in Brown, C and Hagoort, P (Eds) Neurocognition of Language, Oxford: Oxford University Press, 83–122.
Little, D R, Oehmen, R, Dunn, J, Hird, K and Kirsner, K (2013) Fluency profiling system: An automated system for analysing the temporal properties of speech, *Behavior Research Methods* 45(1), 191 – 202.
McCarthy, M (2010) Spoken fluency revisited, *English Profile Journal*, 1, e4.
Nakane, I (2007) *Silence in intercultural communication: Perceptions and performance*, Amsterdam: John Benjamins.
Nakane, I (2012) Silence, in Paulston, C B, Kiesling, S F, and Rangel, E S (Eds.), *Handbook of intercultural discourse and communication*, Oxford: Blackwell, 158-179.
Scollon, R and Scollon, S (1989) *Narrative, literacy and face in interethnic communication*. Norwood, New Jersey: Ablex.
Segalowitz, N (2010) *The cognitive bases of second language fluency*, New York: Routledge.
Skehan, P (2009) Modelling second language performance: Integrating complexity, accuracy, fluency and lexis, *Applied Linguistics*, 30(4), 510-532.
Van Moere, A (2012) A psycholinguistic approach to oral language assessment, *Language Testing*, 29(3), 325-344.

Interactive communication

Lynda Taylor, University of Bedfordshire and Consultant to Cambridge English Language Assessment

In his analysis of the nature of spoken language ability from a cognitive science perspective, Field (2011) noted that most forms of speaking are *reciprocal*, i.e. they involve more than one person, with participants in the spoken encounter switching regularly between the roles of speaker and listener. As they do this, the speaker needs to be attuned to their partner's contribution, while the listener assumes a responsive and dynamic role with the power to maintain the direction of the discourse, or to alter it in a way that may not have been intended by the previous speaker. The exchange between speaker/listeners is characterised by features such as repetition of lexical and syntactic patterns, topic development, and co-ordination of the participants' respective contributions to the emerging discourse through a process of *turn management* which may sometimes involve *repair*. Field (2011) emphasised the substantial cognitive demands that are imposed upon the speaker/listener in a reciprocal (or dialogic) speaking task due to the fact that such an encounter typically takes place under the pressure of real time and often under conditions that bring additional pressures, e.g. lack of topic familiarity, constrained linguistic resources, or an asymmetrical power relationship between participants.

Over the past half century, psycholinguistic accounts of speech production (e.g. Levelt 1989, 1999) together with sociolinguistic accounts of spoken interaction in both L1 and L2 (e.g. Beebe 1980, Coates 1993, Hymes 1972, Wolfson 1989) have significantly improved our understanding and description of the nature of spoken language ability for the purposes of teaching, learning and assessment. Research in sociolinguistics led to the notion of *interactional competence* (Kramsch 1986), according to which *talk-in-interaction* is co-constructed by the participants according to personal characteristics and social conventions (see also Markee 2000). From the 1990s onwards, the application of qualitative research methodologies, such as discourse and conversation analysis, permitted the close scrutiny and detailed identification of differing patterns of talk-in-interaction. This in turn enabled the construct of speaking ability to be enriched in the fields of language pedagogy and assessment (Bygate, Skehan and Swain (Eds) 2001, Lazaraton 2002, McNamara 1997).

From a language assessment perspective, the notion of co-construction in spoken interaction has important implications for the design and use of speaking tests. A range of different test formats has been developed over the years for assessing speaking proficiency but the extent to which these formats permit *reciprocal* interaction to take place can vary significantly. Compare, for example, a speaking test taken over the telephone or through a computer laboratory, with a face-to-face interview or a group oral test (see Fulcher 2003, Luoma 2004 and O'Loughlin 2001 for specific examples of speaking test formats and their features). Even within the narrower range of direct (face-to-face) speaking test formats, patterns of reciprocal interaction can vary, depending upon whether the participants are equal partners in the exchange (e.g. peer candidates in a paired discussion), or

whether they are assigned different roles and responsibilities (e.g. an interviewer who leads the interaction and a candidate who is obliged to follow that lead). It has been argued that testing students' speaking ability in pairs (or in groups) may allow them to demonstrate their interactional competence or interactive communication skills more broadly than in the traditional one-on-one oral proficiency interview format (Taylor and Wigglesworth 2009). Recent years have seen a growth in the use of paired and group speaking test formats, enabling more research to take place in this area and thus provide empirical evidence to support or refute such an assertion (e.g. Brooks 2009, Galaczi 2008, May 2009, Nakatsuhara 2013).

Direct speaking test formats permit the elicitation of a sample of talk-in-interaction so that we can evaluate this for what it tells us about a test taker's interactive communication skills, i.e. their ability to engage in the co-construction of spoken language in a purposeful manner. However, the design and use of such speaking test formats raises a number of issues and challenges for test providers. First, such formats necessarily involve listening as well as speaking skills (Field 2011) though this is not always adequately accounted for within the construct definition. Secondly, a range of interlocutor variables has been shown to impact on speaker performance and score outcomes. Research suggests that features such as interlocutor gender, age, personality, acquaintanceship and conversation style co-operate to shape the spoken exchange (Berry 2007, Brown 2003, O'Sullivan 2002), as well as the asymmetrical distribution of power between interviewer and test taker. Even in paired or group tasks between peers, interlocutor variables such as proficiency level and intelligibility have been shown to influence outcomes (Nakatsuhara 2013). A third issue relates to how raters assign scores to individual test takers on the criterion of interactional competence or interactive communication, and what such scores mean (Fulcher and Davidson 2007, Taylor and Wigglesworth 2009). According to He and Young (1998), interactional competence is not a trait that resides in an individual, nor is it independent of the interactive practice in which it is constituted. Finally, it could be argued that the traditional notion of interactive communication is largely premised upon face-to-face spoken interaction involving physical proximity (i.e. two or more people in the same room). The recent growth of internet-based communication opportunities, much of it 'face-to-face' but using Voice over Internet Protocol (VOIP) (e.g. via Skype, Facetime), may mean that current understanding of interactional competence or interactive communication will need revising to take account of the impact of technology.

This presentation will briefly review how we currently understand the construct of interactive communication as it relates to spoken language proficiency. We will consider some of the challenges and (as yet) unresolved issues over how we adequately account for it in our speaking assessment practices and we shall reflect on some possible directions for future research and test development.

References

Beebe, L M (1980) Sociolinguistic variation in style shifting in second language acquisition, Language Learning 30, 433–447.

Berry, V (2007) *Personality differences and oral test performance*, Frankfurt: Peter

Lang.
Brooks, L (2009) Interacting in pairs in a test of oral proficiency: Co-constructing a better performance, Language Testing 26 (3), 341–366.
Brown, A (2003) Interviewer variation and the co- construction of speaking proficiency, Language Testing 20 (1), 1–25.
Bygate, M, Skehan, P and Swain, M (Eds) (2001) Researching Pedagogic Tasks: Second Language Learning, Teaching and Assessment, London: Pearson.
Coates, J (1993) Women, Men and Language, London: Longman.
Field, J (2011) Cognitive validity, in Taylor, L (Ed.), *Examining speaking: Research and practice in assessing second language speaking*, Cambridge: University of Cambridge ESOL Examinations/Cambridge University Press, 65-111.
Fulcher, G (2003) Testing Second Language Speaking, Harlow: Longman/Pearson Education Ltd.
Fulcher, G and Davidson, F (2007) Language Testing and Assessment, London and New York: Routledge.
Galaczi, E D (2008) Peer- peer interaction in a speaking test: the case of the First Certificate in English examination, Language Assessment Quarterly 5 (2), 89–119.
He, A W and Young, R (1998) Language proficiency interviews: A discourse approach, in Young, R & He, A W (Eds.), *Talking and testing: Discourse approaches to the assessment of oral proficiency*, Philadelphia: John Benjamins, 1-26.
Hymes, D (1972) On communicative competence, in Pride, J and Holmes, J (Eds.), *Sociolinguistics*, Harmondsworth: Penguin, 269-293.
Kramsch, C (1986) Interactive discourse in small and large groups, in Rivers, W (Ed.), *Interactive language teaching,* Cambridge: Cambridge University Press, 17-29.
Lazaraton, A (2002) A Qualitative Approach to the Validation of Oral Language Tests, Studies in Language Testing 14, Cambridge: UCLES/Cambridge University Press.
Levelt, W J M (1989) Speaking, Cambridge, MA: MIT Press.
Levelt, W J M (1999) Language production: a blueprint of the speaker, in Brown, C and Hagoort, P (Eds) Neurocognition of Language, Oxford: Oxford University Press, 83–122.
Luoma, S (2004) *Assessing speaking*, Cambridge: Cambridge University Press.
Markee, N (2000) *Conversation analysis*, Mahwah, NJ: Lawrence Erlbaum.
May, L (2009) Co-constructed interaction in a paired speaking test: The rater's perspective, *Language Testing* 26(3), 397-422.
McNamara, T F (1997) 'Interaction' in second language performance assessment, Applied Linguistics 18 (4), 446–65.
Nakatsuhara, F (2013) *The co-construction of conversation in group oral tests*, Frankfurt: Peter Lang.
O'Loughlin, K (2001) The Equivalence of Direct and Semi- direct Speaking Tests, Studies in Language Testing 13, Cambridge: UCLES/Cambridge University Press.

O'Sullivan, B (2002) Learner acquaintanceship and oral proficiency test pair- task performance, Language Testing 19(3), 277–295.
Taylor, L and Wigglesworth, G (2009) Are two heads better than one? Pair work in L2 assessment contexts, Language Testing 26(3), 325–339.
Wolfson, N (1989) Perspectives: Sociolinguistics and TESOL, Boston MA: Heinle and Heinle.

Session 4: Technology and assessment

Computer recognition of learner speech

Helmer Strik, Radboud University

Standard ‘Automatic Speech Recognition’ (ASR) systems are generally employed to recognize words. For instance, speech-driven dictation systems convert speech spoken into a microphone to text, strings of words appearing on a screen. The ASR system itself consists of a decoder (the search algorithm) and three 'knowledge sources': the language model, the lexicon, and the acoustic models. The language model (LM) contains probabilities of words and sequences of words. Acoustic models are models of how the sounds of a language are pronounced. The lexicon is the connection between the language model and the acoustic models. It contains information on how the words are pronounced, in terms of sequences of speech sounds. Therefore, the lexicon contains two representations for every entry: an orthographic transcription representing how a word is written, and a phonological transcription representing how a word is pronounced. Since words can be pronounced in different ways, lexicons often contain more than one entry per word, i.e. the pronunciation variants, which indicate possible pronunciations of one and the same word.

ASR of native speech is already complex because of many well-known problems such as background (speech) sounds, (low) signal-to-noise ratio (SNR), end-point detection, pronunciation variation, and dysfluencies. However, ASR of learner speech is even more complex, since the grammar, the words used, and the pronunciation can deviate considerably, thus affecting all three 'knowledge sources' of the ASR system (language model, lexicon, and acoustic models, respectively). In the ASR community, it has long been known that the differences between native and non-native speech are so extensive as to degrade ASR performance considerably. Furthermore, native and non-native speech can differ in many (sometimes unexpected) ways, e.g. non-native speech often contains more broken words for (cold) reading, and many more filled pauses in spontaneous speech.

As the quality of speech technology improved, more and more researchers tried to apply it to language learning, sometimes with disappointing results. Some researchers were skeptical about the usefulness and effectiveness of ASR-based Computer Assisted Language Learning (CALL) programs: evidence gathered in different lines of research seemed to confirm that either speech technology was not mature enough, or ASR-based CALL programs were not effective in improving second language (L2) skills. For the sake of our own research, we studied this literature thoroughly and gradually acquired the impression that, while it is undeniable that speech technology still presents a number of limitations, especially when applied to non-native speech, part of this pessimism is in fact due to misconceptions about this technology and CALL in general.

For instance, in some studies unsatisfactory results were obtained when standard dictation systems were used for CALL. Such dictation systems are not

suitable for L2 training or for recognition of L2 speech, as CALL requires dedicated speech technology. Apart from the fact that the majority of dictation packages are developed for native speakers, the major problem here is that CALL and this technology have differing goals and thus require different ASR approaches. The aim of a dictation package is to convert an acoustic signal into a string of words and not to identify L2 errors, which requires a different, more complex procedure. Consequently, the negative conclusions related to the use of dictation packages should be related to those specific cases and not to ASR technology in general.

The following fragments were obtained from the website http://www.ict4lt.org/en/en_mod4-1.htm on 1 June 2013 (the website states the document was last updated on 19 April 2012):

> 1.3 Which skills can be assessed?
>
> Speaking: Very limited as yet. Automatic Speech Recognition (ASR) software is developing rapidly but it is still too unreliable to be used in accurate testing.
>
> To assess speaking skills solely by a computer, using Automatic Speech Recognition (ASR), is a very complex task and research in this area is developing rapidly. ASR can be motivating for students working independently, but computers are still not completely reliable as assessors.

So it seems that the authors acknowledge that ASR can potentially be useful, but are still skeptical about the quality.

Many of the first studies that considered employing ASR technology in the context of L2 learning focused primarily on the automatic assessment of different aspects of L2 oral proficiency, in particular L2 pronunciation. The results showed that automatic testing of certain aspects of oral proficiency was feasible: the scores obtained by means of ASR technology were strongly correlated with human judgments of oral proficiency.

In general, such automatic scores were calculated at a rather global level, for instance for several utterances by the same speaker, because in this way more reliable measures could be obtained. Such measures might be suitable, and in certain cases even preferable, for testing purposes, for assessing the problems of individual speakers, for providing overviews of words or phonemes that appear to be difficult and suggesting remedial exercises for the problematic cases. However, such overall measures are generally not specific enough for practice and feedback purposes.

For training, error detection is required, a procedure by which a score at a local (e.g. phoneme) level is calculated. In general, the relation between human and automatic grading improves if longer stretches of speech are used, i.e. complete utterances or a couple of utterances. Such cumulative measures can also be adopted for error detection, for instance by combining the scores of several utterances. This can be useful to assess the problems of a specific speaker, to obtain an overview and suggest remedial exercises for the problematic cases.

However, for remedial exercises immediate feedback based on local calculations is to be preferred.

To sum up, ASR of non-native speech is indeed complex. Still, if one carefully takes account of its limitations it can already be applied usefully. In several projects we have developed speech technology for language learners and have studied different aspects related to the use of such technology. For instance, we investigated the performance of the speech technology used, the way in which it could best be implemented in applications, how such applications should be designed and how feedback can best be provided to the learners, how language learners experience the use of speech technology, including its effect on their motivation, etc. In our presentation we will provide an overview.

Interactive Dialogue Systems

Diane Litman, University of Pittsburgh

The field of interactive dialogue systems uses speech and language processing to enable extended human-machine conversations. In the typical pipeline architecture of most spoken dialogue systems, a speech recognition component transcribes a spoken user utterance, a natural language understanding component extracts the transcription's syntactic structure and/or meaning, a dialogue manager determines an appropriate system response, a natural language generation component maps the response to text, and a text-to-speech component produces a spoken system utterance. There are many opportunities for assessment in the context of such a dialogue system, including utterance-level assessment of what a user says and how the user says it (e.g. to guide the real-time operation of the dialogue manager), as well as dialogue-level assessment of conversational properties such as turn-taking and dialogue structure (e.g., to evaluate a user's conversational abilities).

At the utterance-level, a dialogue manager typically uses the assessments from the speech recognition and natural language understanding components, in conjunction with an internal representation of dialogue and task state, to decide what the dialogue system should do next. For example, in a finite state dialogue manager where the dialogue states correspond to system utterances, the assessments of user utterances determine the transitions between dialogue states. Note that speech and natural language processing can be used to assess the speech files and transcriptions representing the user's utterances with respect to many linguistic dimensions. For example, in a tutorial dialogue system, syntactic analysis can been used to detect grammatical errors, while semantic analysis can be used to assess meaning with respect to an expected answer at both fine (e.g., paraphrase or entailment recognition) and coarse (e.g., on-topic or off-topic recognition) grained levels of analysis. Knowledge of pragmatics can be used to assess skills such as politeness, whereas knowledge of discourse can been used to evaluate local contextual coherence. Finally, acoustic and prosodic information particular to speech can be used to assess speaking fluency, as well as pedagogically important user states such as boredom, uncertainty, and frustration.

While utterance-level assessment in an interactive spoken dialogue system may overlap with assessment tasks in non-interactive contexts, there are often differences as well as challenges in moving to a dialogue context. For example, the goal of traditional short answer scoring is to produce a numeric score that agrees with a gold-standard human score, using statistical techniques such as lexical or semantic similarity, as well as approaches based on deeper semantic processing and inference. In contrast, the goal of short answer assessment in a dialogue system is to use similar methods to assign a label corresponding to an allowable transition from the system's current dialogue state (e.g. in a tutorial dialogue system, assessing a response to a tutor's question as correct, partially correct, or wrong, in order to reach the dialogue state corresponding to the most appropriate system feedback). Second, utterances produced during dialogue are often more spontaneous and unconstrained compared to utterances produced in non-interactive

contexts, making them less predictable and harder to assess on many dimensions. As a result, compared to text, assessment of speech proficiency has focused less on aspects such as semantics, discourse and pragmatics, and more on aspects such as pronunciation and fluency. Third, the interactive capabilities of dialogue systems suggest computing and using confidence or belief information as a method to better handle noisy utterance assessments. For example, a dialogue manager with state tracking can use methods from artificial intelligence such as Bayesian networks and discriminative models to maintain a belief distribution over dialogue states as the dialogue progresses. This is in contrast to a dialogue manager that simply uses the most likely utterance assessment to select the next dialogue state, discarding any information regarding the less-likely alternatives. Another approach to handling uncertainty is to trigger a system clarification when the best assessment of a user's utterance is of low confidence. Fourth, some types of user behaviors to be assessed only occur in interactive dialogue (e.g. turn-taking). Fifth, assessments for online dialogue management must be based on linguistic features that can be computed automatically and in real-time.

At the dialogue-level, assessment typically involves higher-level and contextual user abilities that require multiple utterances of the dialogue for analysis, and that reflect the fact that dialogue is a joint activity involving two or more conversational participants. For example, in a coherent dialogue, consecutive user utterances should not be isolated and unrelated to one another. Instead, user utterances should exhibit semantic and topical relationships with both the system's and the user's history of prior utterances. In addition, user utterances should be used to achieve appropriate conversational functions, such as providing an answer after a system question, or ending the dialogue with a closing rather than a greeting. Users should also be able to use linguistic devices such as referring expressions, discourse markers, prosody, etc. that are both consistent with the underlying relationships between utterances, and that are used at appropriate times during the conversation. With respect to turn-taking abilities, users should be able to both recognize when it is their turn in a dialogue, and use linguistic signals to convey to the system that they are maintaining or ending their turn. Users must also be able to effectively ground the system's utterances, making it clear what the user has actually heard and understood, generating confirmations to the system when necessary, and appropriately recovering from system misunderstandings.

One challenge in assessing user conversational abilities is that unlike many utterance assessment tasks, there is not usually a single best reference answer. Another challenge is that due to technology limitations, user conversations with computers often exhibit somewhat different characteristics (e.g. they are simpler and more constrained) than conversations with other humans. In addition, most research in the area of dialogue has focused on understanding human dialogue abilities in order to build better spoken dialogue systems, rather than to assess user behavior along conversational dimensions. However, there are approaches being developed to evaluate the quality of simulated (i.e. computer) users of a spoken dialogue system, with respect to features such as quantity of user activity, distribution of dialogue functions of user utterances, and overall success and efficiency of the

interaction. Evaluation measures have similarly been developed to evaluate the quality of dialogue systems with respect to optimizing user satisfaction. Such evaluation approaches could potentially be adapted to assess the dialogues abilities of human partners from their interactions with spoken dialogue systems.

Automated assessment: Moving from written text to transcribed speech

Ted Briscoe, University of Cambridge

The task of automated assessment (AA) of free text focuses in the EFL context on analysing and assessing the quality and variety of writing competence. Automated assessment systems exploit textual features in order to measure overall quality and assign a score to a text. The earliest systems used superficial features, such as word and sentence length, as proxies for examiners' judgements. More recent systems have used more sophisticated automated text processing techniques to measure grammaticality, textual coherence, pre-specified errors, and so forth.

Deployment of AA systems gives a number of advantages, such as reduced workload in marking texts, especially when applied to large-scale assessments. Additionally, automated systems guarantee the application of the same marking criteria, thus reducing inconsistency, which may arise when more than one human examiner is employed. Often, implementations include feedback with respect to the writers' writing abilities, thus facilitating self-assessment and self-tutoring.

Most work has treated AA as a supervised text classification task, where training texts are labelled with a grade and unlabelled test texts are fitted to the same grade point scale via a regression step applied to the classifier output and texts are represented in terms of manually pre-specified features. In our work with Cambridge English Language Assessment we have treated AA as a supervised discriminative machine learning problem where the task is to rank scripts on an ordinal scale and the features used to represent the text are selected and weighted automatically as part of the training process. This removes the need for the regression step better modelling the grading task, and also removes the need to manually pre-specify the aspects of the text that are assumed to be criterial for grading and recoverable using current language processing technology.

I will describe and motivate our approach to grading written text and report experimental results which suggest that our AA system's performance is essentially indistinguishable from a human examiner when applied to text similar to that seen during training. I will then go on to describe the challenges of applying similar AA techniques to the grading of speech. Primary amongst these are the issues of transcribing L2 speech produced by speakers of varying abilities and L1 backgrounds.

However, I will focus on the applicability of our language processing and AA technology to transcribed output and here I will attempt to demonstrate that the outlook is very promising.

Improving intelligent tutoring of pronunciation consonant cluster problems

Gary Pelton, Carnegie Speech

The past decade has seen improvements in the algorithmic detection of errors in non-native pronunciation (Eskenazi 1996, Neumeyer, Franco, Digalakis and Weintraub 2000). This work has led to the development of products that can be used to improve a user's pronunciation. The work of Akahane-Yamada, Kato, Adachi, Watanabe, Komaki, Kubo, Takada, and Ikuma (2004) has been converted to a complete pronunciation training system for Japanese speakers who want to speak English. It has been commercialized by ATR in Japan. NativeAccent™ by Carnegie Speech is another commercial product used to train non-native speakers to improve their pronunciation of English.

NativeAccent started as the Fluency Project (Eskenazi 1996) at Carnegie Mellon University. It has been converted to a complete pronunciation training system, and over the last 10 years thousands of people have used NativeAccent to improve their pronunciation. NativeAccent compares a user's pronunciation to a statistical model of native speakers using a technique called pinpointing. A close match is considered good pronunciation. We have these models for both Midwest American English speakers and British English from speakers whose speech is close to what used to be called BBC English. Rather than doing a detailed analysis of how NativeAccent detects pronunciation issues, this paper documents one of our explorations in how NativeAccent can be improved to give our users a better experience and to better address their specific problems. This exploration is done by examining the logs of user data looking for patterns of problems. Focusing on the student's specific problems is important in the NativeAccent intelligent tutor methodology.

The learning improvement from the use of intelligent tutoring has been shown in many domains. In particular Anderson (1993) shows reduction of one third the training time to achieve a particular level of performance in learning a programming language. The work of Koedinger, Anderson, Hadley and Mark (1997) shows an improvement of 1 standard deviation for students using an Algebra 1 intelligent tutor compared to similar students in a teacher-run Algebra 1 class. Carnegie Speech has published studies (see Eskenazi, Ke, Albornoz and Probst 1998) about how well our users do using this intelligent tutor approach. A more recent unpublished study of 120 people showed the 60 people who only participated in a teacher run pronunciation class (control group) improved 62% in seven hours, and the 60 people in the test group using NativeAccent during half of their pronunciation class time improved 104%. This difference was not quite statistically significant (r=0.07), but the time using NativeAccent was relatively small. We see substantial improvement in about 10 hours of use.

Perfect intelligent tutoring relies upon knowing what skill deficiency is the root cause of a mistake and having curriculum that helps the student learn to avoid that mistake. The graphs showing the transformation of errors into learning curves once errors can be assigned to a cause (rule) in Anderson (1993) are very compelling. The empirical results on tutors that can assign root causes supports that research.

Pinpointing provides probabilistic root cause information for substitution, affrication, deletion and other mistakes related to a single phone. Co-articulation problems and those problems associated with a phone in a context, along with errors in automatic speech analysis reduce the effectiveness of the tutoring. However, these problems aren't systematic, and pinpointing does pick up on the systematic issues the user exhibits. Carnegie Speech's experience shows that the small amount of random error doesn't affect the tutoring overmuch.

Pronunciation problems within consonant clusters are the focus of the problems investigated in this paper. Consonant clusters are where the native pronunciation of a word involves a sequence of consonants without intervening vowels. Consonant clusters are a known pronunciation problem in non-native speech (Hultzen 1993). For that reason, NativeAccent's curriculum includes consonant clusters exercises in its lessons. For example, the /t/ lesson might start with exercises on simple words like "tap", "bat" and "later" but the lesson can also include words with "t" in a consonant cluster like "string". The question asked in this paper is whether we should change our detection algorithms and/or our tutoring because our users are exhibiting systematic problems with pronouncing phones within consonant clusters that our current system could handle better. This paper uses data mining on the logs of NativeAccent users to answer both this question and to provide more insight into the various consonant cluster problems our users exhibit. The intelligent tutoring mechanism underlying NativeAccent will be described in detail so that the reader can understand how this data can improve the tutoring process.

References

Akahane-Yamada, R, Kato, H, Adachi, T, Watanabe, H, Komaki, R, Kubo, R, Takada, T and Ikuma, Y (2004) ATR CALL: A speech perception/production training system utilizing speech technology, *Proc. ICA 2004*, 2004, vol. III, 2319–2320.

Altenberg, E (2005) The judgment, perception, and production of consonant clusters in a second language, *IRAL*, 43, 53–80.

Anderson, J R (1993) *Rules of the Mind*, Hillsdale, NJ: Erlbaum.

Celce-Murcia, M, Brinton, D and Goodwin, J (1996) *Teaching Pronunciation,* Cambridge: Cambridge University Press, 1996.

Eskenazi, M (1996) Detection of foreign speakers' pronunciation errors for second language training - preliminary results, *Proc. International Conference on Spoken Language Processing*, Philadelphia.

Eskenazi, M, Ke, Y, Albornoz, J and Probst K. (1998) The Fluency Pronunciation Trainer: Update and user issues, presented at the STiLL Workshop on Speech Technology in Language Learning, Marhollmen.

Hultzen, L (1965) Consonant clusters in English, *American Speech*, 49(1), 5–19.

Koedinger, K R, Anderson, J R, Hadley, W H and Mark, M A (1997) Intelligent tutoring goes to school in the big city, *International Journal of Artificial Intelligence in Education*, 8, 30–43.

Neumeyer, L, Franco, H, Digalakis, V and Weintraub, M (2000) Automatic scoring of pronunciation quality, *Speech Communication*, 30(2), 83–93.

Session 5: Speaking assessment into the future

Concluding talk

Gad S Lim, Cambridge English Language Assessment

The contributions from the different sessions will be summarised and synthesised in this talk.