Automatic Assessment of Spoken Modern Standard Arabic

Jian Cheng, Jared Bernstein, Ulrike Pado, Masanori Suzuki
Pearson Knowledge Technologies
299 California Ave, Palo Alto, CA 94306
jian.cheng@pearson.com

Abstract

Proficiency testing is an important ingredient in successful language teaching. However, repeated testing for course placement, over the course of instruction or for certification can be time-consuming and costly. We present the design and validation of the Versant Arabic Test, a fully automated test of spoken Modern Standard Arabic, that evaluates test-takers' facility in listening and speaking. Experimental data shows the test to be highly reliable (test-retest r=0.97) and to strongly predict performance on the ILR OPI (r=0.87), a standard interview test that assesses oral proficiency.

1 Introduction

Traditional high-stakes testing of spoken proficiency often evaluates the test-taker's ability to accomplish communicative tasks in a conversational setting. For example, learners may introduce themselves, respond to requests for information, or accomplish daily tasks in a role-play.

Testing oral proficiency in this way can be time-consuming and costly, since at least one trained interviewer is needed for each student. For example, the standard oral proficiency test used by the United States government agencies (the Interagency Language Roundtable Oral Proficiency Interview or ILR OPI) is usually administered by two certified interviewers for approximately 30-45 minutes per candidate.

The great effort involved in oral proficiency interview (OPI) testing makes automated testing an attractive alternative. Work has been reported on fully automated scoring of speaking ability (e.g., Bernstein & Barbier, 2001; Zechner et al., 2007, for English; Balogh & Bernstein, 2007, for English and Spanish). Automated testing systems do not aim to simulate a conversation with the test-taker and therefore do not directly observe interactive human communication. Bernstein and Barbier (2001) describe a system that might be used in qualifying simultaneous interpreters; Zechner et al. (2007) describe an automated scoring system that assesses performance according to the TOEFL iBT speaking rubrics. Balogh and Bernstein (2007) focus on evaluating *facility* in a spoken language, a separate test construct that relates to oral proficiency.

“Facility in a spoken language” is defined as “the ability to understand a spoken language on everyday topics and to respond appropriately and intelligibly at a native-like conversational pace” (Balogh & Bernstein, 2007, p. 272). This ability is assumed to underlie high performance in communicative settings, since learners have to understand their interlocutors correctly and efficiently in real time to be able to respond. Equally, learners have to be able to formulate and articulate a comprehensible answer without undue delay. Testing for *oral proficiency*, on the other hand, conventionally includes additional aspects such as correct interpretation of the pragmatics of the conversation, socially and culturally appropriate wording and content and knowledge of the subject matter under discussion.

In this paper, we describe the design and validation of the Versant Arabic Test (VAT), a fully automated test of facility with spoken Modern Standard Arabic (MSA). Focusing on facility rather than communication-based oral proficiency enables the creation of an efficient yet informative automated test of listening and speaking ability. The automated test can be administered over the telephone or on a computer in approximately 17 minutes. Despite its much shorter format and constrained tasks, test-taker scores on the VAT

Proceedings of the NAACL HLT Workshop on Innovative Use of NLP for Building Educational Applications, pages 1–9,
Boulder, Colorado, June 2009. ©2009 Association for Computational Linguistics

strongly correspond to their scores from an ILR Oral Proficiency Interview.

The paper is structured as follows: After reviewing related work, we describe Modern Standard Arabic and introduce the test construct (i.e., what the test is intended to measure) in detail (Section 3). We then describe the structure and development of the VAT in Section 4 and present evidence for its reliability and validity in Section 5.

2 Related Work

The use of automatic speech recognition appeared earliest in pronunciation tutoring systems in the field of language learning. Examples include SRI's AUTOGRADER (Bernstein et al., 1990), the CMU FLUENCY system (Eskenazi, 1996; Eskenazi & Hansma, 1998) and SRI's commercial EduSpeak system (Franco et al., 2000). In such systems, learner speech is typically evaluated by comparing features like phone duration, spectral characteristics of phones and rate-of-speech to a model of native speaker performances. Systems evaluate learners' pronunciation and give some feedback.

Automated measurement of more comprehensive speaking and listening ability was first reported by Townshend et al. (1998), describing the early PhonePass test development at Ordinate. The PhonePass tests returned five diagnostic scores, including reading fluency, repeat fluency and listening vocabulary. Ordinate's Spoken Spanish Test also included automatically scored passage retellings that used an adapted form of latent semantic analysis to estimate vocabulary scores.

More recently at ETS, Zechner et al. (2007) describe experiments in automatic scoring of test-taker responses in a TOEFL iBT practice environment, focusing mostly on fluency features. Zechner and Xi (2008) report work on similar algorithms to score item types with varying degrees of response predictability, including items with a very restricted range of possible answers (e.g., reading aloud) as well as item types with progressively less restricted answers (e.g., describing a picture – relatively predictable, or stating an opinion – less predictable). The scoring mechanism in Zechner and Xi (2008) employs features such as the average number of word types or silences for fluency estimation, the ASR HMM log-likelihood for pronunciation or a vector-based similarity measure to assess vocabulary and content. Zechner and Xi present correlations of machine scores with human scores for two tasks: r=0.50 for an opinion task and r=0.69 for picture description, which are comparable to the modest human rater agreement figures in this data.

Balogh and Bernstein (2007) describe operational automated tests of spoken Spanish and English that return an overall ability score and four diagnostic subscores (sentence mastery, vocabulary, fluency, pronunciation). The tests measure a learner's facility in listening to and speaking a foreign language. The facility construct can be tested by observing performance on many kinds of tasks that elicit responses in real time with varying, but generally high, predictability. More predictable items have two important advantages: As with domain restricted speech recognition tasks in general, the recognition of response content is more accurate, but a higher precision scoring system is also possible as an independent effect beyond the greater recognition accuracy. Scoring is based on features like word stress, segmental form, latency or rate of speaking for the fluency and pronunciation subscores, and on response fidelity with expected responses for the two content subscores. Balogh and Bernstein report that their tests are highly reliable (r>0.95 for both English and Spanish) and that test scores strongly predict human ratings of oral proficiency based on Common European Framework of Reference language ability descriptors (r=0.88 English, r=0.90 Spanish).

3 Versant Arabic Test: Facility in Modern Standard Arabic

We describe a fully operational test of spoken MSA that follows the tests described in Balogh and Bernstein (2007) in structure and method, and in using the facility construct. There are two important dimensions to the test's construct: One is the definition of what comprises MSA, and the other the definition of facility.

3.1 Target Language: Modern Standard Arabic

Modern Standard Arabic is a non-colloquial language used throughout the Arabic-speaking world for writing and in spoken communication within public, literary, and educational settings. It differs from the colloquial dialects of Arabic that are spoken in the countries of North Africa and the Middle East in lexicon and in syntax, for example in the use of explicit case and mood marking.

Written MSA can be identified by its specific syntactic style and lexical forms. However, since all short vowels are omitted in normal printed material, the word-final short vowels indicating case and mood are provided by the speaker, even when reading MSA aloud. This means that a text that is syntactically and lexically MSA can be read in a way that exhibits features of the regional dialect of the speaker if case and mood vowels are omitted or phonemes are realized in regional pronunciations. Also, a speaker's dialectal and educational background may influence the choice of lexical items and syntactic structures in spontaneous speech. The MSA spoken on radio and television in the Arab world therefore shows a significant variation of syntax, phonology, and lexicon.

3.2 Facility

We define *facility* in spoken MSA as the ability to understand and speak contemporary MSA as it is used in international communication for broadcast, for commerce, and for professional collaboration. Listening and speaking skills are assessed by observing test-taker performance on spoken tasks that demand understanding a spoken prompt, and formulating and articulating a response in real time.

Success on the real-time language tasks depends on whether the test-taker can process spoken material efficiently. Automaticity is an important underlying factor in such efficient language processing (Cutler, 2003). Automaticity is the ability to access and retrieve lexical items, to build phrases and clause structures, and to articulate responses without conscious attention to the linguistic code (Cutler, 2003; Jescheniak et al., 2003; Levelt, 2001). If processing is automatic, the listener/speaker can focus on the communicative content rather than on how the language code is structured. Latency and pace of the spoken response can be seen as partial manifestation of the test-taker’s automaticity.

Unlike the oral proficiency construct that coordinates with the structure and scoring of OPI tests, the facility construct does not extend to social skills, higher cognitive functions (e.g., persuasion), or world knowledge. However, we show below that test scores for language facility predict almost all of the reliable variance in test scores for an interview-based test of language and communication.

4 Versant Arabic Test

The VAT consists of five tasks with a total of 69 items. Four diagnostic subscores as well as an overall score are returned. Test administration and scoring is fully automated and utilizes speech processing technology to estimate features of the speech signal and extract response content.

4.1 Test Design

The VAT items were designed to represent core syntactic constructions of MSA and probe a wide range of ability levels. To make sure that the VAT items used realistic language structures, texts were adapted from spontaneous spoken utterances found in international televised broadcasts with the vocabulary altered to contain common words that a learner of Arabic may have encountered.

Four educated native Arabic speakers wrote the items and five dialectically distinct native Arabic speakers (Arabic linguist/teachers) independently reviewed the items for correctness and appropriateness of content. Finally, fifteen educated native Arabic speakers (eight men and seven women) from seven different countries recorded the vetted items at a conversational pace, providing a range of native accents and MSA speaking styles in the item prompts.

4.2 Test Tasks and Structure

The VAT has five task types that are arranged in six sections (Parts A through F): Readings, Repeats (presented in two sections), Short Answer Questions, Sentence Builds, and Passage Retellings. These item types provide multiple, fully independent measures that underlie facility with spoken MSA, including phonological fluency, sentence construction and comprehension, passive and active vocabulary use, and pronunciation of rhythmic and segmental units.

Part A: Reading (6 items) In this task, test-takers read six (out of eight) printed sentences, one at a time, in the order requested by the examiner voice. Reading items are printed in Arabic script with short vowels indicated as they would be in a basal school reader. Test-takers have the opportunity to familiarize themselves with the reading items before the test begins. The sentences are relatively simple in structure and vocabulary, so they can be read easily and fluently by people educated in MSA. For test-takers with little facility in spoken Arabic but with some reading skills, this task provides samples of pronunciation and oral reading fluency.

Parts B and E: Repeats (2x15 items) Test-takers hear sentences and are asked to repeat them verbatim. The sentences were recorded by native speakers of Arabic at a conversational pace. Sentences range in length from three words to at most twelve words, although few items are longer than nine words. To repeat a sentence longer than about seven syllables, the test-taker has to recognize the words as produced in a continuous stream of speech (Miller & Isard, 1963). Generally, the ability to repeat material is constrained by the size of the linguistic unit that a person can process in an automatic or nearly automatic fashion. The ability to repeat longer and longer items indicates more and more advanced language skills – particularly automaticity with phrase and clause structures.

Part C: Short Answer Questions (20 items) Test-takers listen to spoken questions in MSA and answer each question with a single word or short phrase. Each question asks for basic information or requires simple inferences based on time, sequence, number, lexical content, or logic. The questions are designed not to presume any specialist knowledge of specific facts of Arabic culture or other subject matter. An English example[1] of a Short Answer Question would be "*Do you get milk from a bottle or a newspaper?*" To answer the questions, the test-taker needs to identify the words in phonological and syntactic context, infer the demand proposition and formulate the answer.

Part D: Sentence Building (10 items) Test-takers are presented with three short phrases. The phrases are presented in a random order (excluding the original, naturally occurring phrase order), and the test-taker is asked to respond with a reasonable sentence that comprises exactly the three given phrases. An English example would be a prompt of "*was reading - my mother - her favorite magazine*", with the correct response: "*My mother was reading her favorite magazine.*" In this task, the test-taker has to understand the possible meanings of each phrase and know how the phrases might be combined with the other phrasal material, both with regard to syntax and semantics. The length and complexity of the sentence that can be built is constrained by the size of the linguistic units with which the test-taker represents the prompt phrases in verbal working memory (e.g., a syllable, a word or a multi-word phrase).

Part F: Passage Retelling (3 items) In this final task, test-takers listen to a spoken passage (usually a story) and then are asked to retell the passage in their own words. Test-takers are encouraged to retell as much of the passage as they can, including the situation, characters, actions and ending. The passages are from 19 to 50 words long. Passage Retellings require listening comprehension of extended speech and also provide additional samples of spontaneous speech. Currently, this task is not automatically scored in this test.

4.3 Test Administration

Administration of the test takes about 17 minutes and the test can be taken over the phone or via a computer. A single examiner voice presents all the spoken instructions in either English or Arabic and all the spoken instructions are also printed verbatim on a test paper or displayed on the computer screen. Test items are presented in Arabic by native speaker voices that are distinct from the examiner voice. Each test administration contains 69 items selected by a stratified random draw from a large item pool. Scores are available online within a few minutes after the test is completed.

4.4 Scoring Dimensions

The VAT provides four diagnostic subscores that indicate the test-taker's ability profile over various dimensions of facility with spoken MSA. The subscores are

- *Sentence Mastery*: Understanding, recalling, and producing MSA phrases and clauses in complete sentences.
- *Vocabulary*: Understanding common words spoken in continuous sentence context and producing such words as needed.
- *Fluency*: Appropriate rhythm, phrasing and timing when constructing, reading and repeating sentences.
- *Pronunciation*: Producing consonants, vowels, and lexical stress in a native-like manner in sentence context.

[1] See Pearson (2009) for Arabic example items.

The VAT also reports an Overall score, which is a weighted average of the four subscores (Sentence Mastery contributes 30%, Vocabulary 20%, Fluency 30%, and Pronunciation 20%).

4.5 Automated Scoring

The VAT's automated scoring system was trained on native and non-native responses to the test items as well as human ability judgments.

Data Collection For the development of the VAT, a total of 246 hours of speech in response to the test items was collected from natives and learners and was transcribed by educated native speakers of Arabic. Subsets of the response data were also rated for proficiency. Three trained native speakers produced about 7,500 judgments for each of the Fluency and the Pronunciation subscores (on a scale from 1-6, with 0 indicating missing data). The raters agreed well with one another at r≈0.8 (r=0.79 for Pronunciation, r=0.83 for Fluency). All test administrations included in the concurrent validation study (cf. Section 5 below) were excluded from the training of the scoring system.

Automatic Speech Recognition Recognition is performed by an HMM-based recognizer built using the HTK toolkit (Young et al., 2000). Three-state triphone acoustic models were trained on 130 hours of non-native and 116 hours of native MSA speech. The expected response networks for each item were induced from the transcriptions of native and non-native responses.

Since standard written Arabic does not mark short vowels, the pronunciation and meaning of written words is often ambiguous and words do not show case and mood markings. This is a challenge to Arabic ASR, since it complicates the creation of pronunciation dictionaries that link a word's sound to its written form. Words were represented with their fully voweled pronunciation (cf., Vergyri et al., 2008; Soltau et al., 2007). We relied on hand-corrected automatic diacritization of the standard written transcriptions to create fully-voweled words from which phonemic representations were automatically created.

The orthographic transcript of a test-taker utterance in standard, unvoweled form is still ambiguous with regard to the actual words uttered, since the same consonant string can have different meanings depending on the vowels that are inserted. Moreover, the different words written in this way are usually semantically related, making them potentially confusable for language learners. Therefore, for system development, we transcribed words with full vowel marks whenever a vowel change would cause a change of meaning. This partial voweling procedure deviates from the standard way of writing, but it facilitated system-internal comparison of target answers with observed test-taker utterances since the target pronunciation was made explicit.

Scoring Methods The Sentence Mastery and Vocabulary scores are derived from the accuracy of the test-taker's response (in terms of number of words inserted, deleted, or substituted by the candidate), and the presence or absence of expected words in correct sequences, respectively.

The Fluency and Pronunciation subscores are calculated by measuring the latency of the response, the rate of speaking, the position and length of pauses, the stress and segmental forms of the words, and the pronunciation of the segments in the words within their lexical and phrasal context. The final subscores are based on a non-linear combination of these features. The non-linear model is trained on feature values and human judgments for native and non-native speech.

Figure 1 shows how each subscore draws on responses from the different task types to yield a stable estimate of test-taker ability. The Pronunciation score is estimated from responses to Reading, Repeat and Sentence Build items. The Fluency score uses the same set of responses as for Pronunciation, but a different set of acoustic features are extracted and combined in the score. Sentence Mastery is derived from Repeat and Sentence Building items and Vocabulary is based on responses to the Short Answer Questions.

5 Evaluation

For any test to be meaningful, two properties are crucial: *Reliability* and *validity*. Reliability represents how consistent and replicable the test scores are. Validity represents the extent to which one can justify making certain inferences or decisions on the basis of test scores. Reliability is a necessary condition for validity, since inconsistent measurements cannot support inferences that would justify real-world decision making.

To investigate the reliability and the validity of the VAT, a concurrent validation study was conducted in which a group of test-takers took both

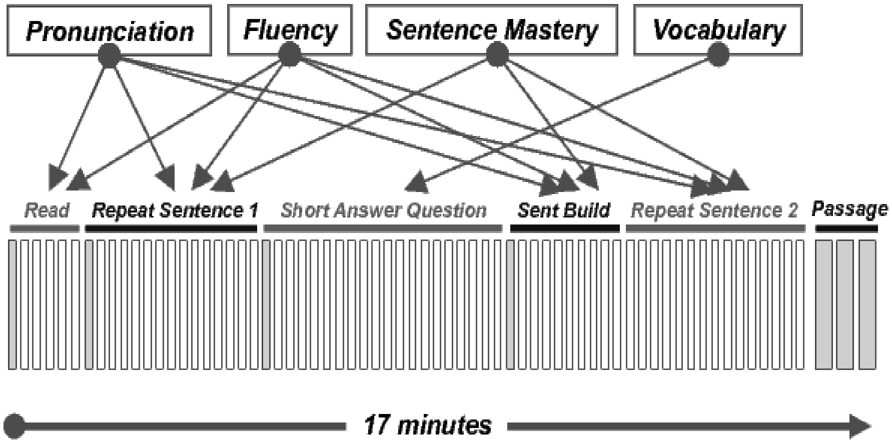

Figure 1: Relation of subscores to item types.

the VAT and the ILR OPI. If the VAT scores are comparable to scores from a reliable traditional measure of oral proficiency in MSA, this will be a piece of evidence that the VAT indeed captures important aspects of test-takers' abilities in using spoken MSA.

As additional evidence to establish the validity of the VAT, we examined the performance of the native and non-native speaker groups. Since the test claims to measure facility in understanding and speaking MSA, most educated native speakers should do quite well on the test, whereas the scores of the non-native test-takers should spread out according to their ability level. Furthermore, one would also expect that educated native speakers would perform equally well regardless of specific national dialect backgrounds and no important score differences among different national groups of educated native speakers should be observed.

5.1 Concurrent Validation Study

ILR OPIs. The ILR Oral Proficiency Interview is a well-established test of spoken language performance, and serves as the standard evaluation tool used by United States government agencies (see www.govtilr.org). The test is a structured interview that elicits spoken performances that are graded according to the ILR skill levels. These levels describe the test-taker’s ability in terms of communicative functioning in the target language. The OPI test construct is therefore different from that of the VAT, which measures facility with spoken Arabic, and not communicative ability, as such.

Concurrent Sample. A total of 118 test-takers (112 non-natives and six Arabic natives) took two VATs and two ILR OPIs. Each test-taker completed all four tests within a 15 day window. The mean age of the test-takers was 27 years old (SD = 7) and the male-to-female split was 60-to-58. Of the non-native speakers in this concurrent testing sample, at least 20 test-takers were learning Arabic at a college in the U.S., and at least 11 were graduates from the Center for Arabic Studies Abroad program. Nine test-takers were recruited at a language school in Cairo, Egypt, and the remainder were current or former students of Arabic recruited in the US.

Seven active government-certified oral proficiency interviewers conducted the ILR OPIs over the telephone. Each OPI was administered by two interviewers who submitted the performance ratings independently after each interview. The average inter-rater correlation between one rater and the average score given by the other two raters administering the same test-taker's other interview was 0.90.

The test scores used in the concurrent study are the VAT Overall score, reported here in a range from 10 to 90, and the ILR OPI scores with levels {0, 0+, 1, 1+, 2, 2+, 3, 3+, 4, 4+, 5}[2].

5.2 Reliability

Since each test-taker took the VAT twice, we can estimate the VAT’s reliability using the test-retest method (e.g., Crocker & Algina, 1986: 133). The

[2] All plus ratings (e.g., 1+, 2+, etc) were converted with 0.5 (e.g, 1.5, 2.5, etc) in the analysis reported in this paper.

correlation between the scores from the first administration and the scores from the second administration was found to be at r=0.97, indicating high reliability of the VAT test. The scores from one test administration explain 0.97^2=94% of the score variance in another test administration to the same group of test-takers.

We also compute the reliability of the ILR OPI scores for each test taker by correlating the averages of the ratings for each of the two test administrations. The OPI scores are reliable at r=0.91 (thus 83% of the variance in the test scores are shared by the scores of another administration). This indicates that the OPI procedure implemented in the validation study was relatively consistent.

5.3 Validity

Evidence here for VAT score validity comes from two sources: the prediction of ILR OPI scores (assumed for now to be valid) and the performance distribution of native and non-native test takers.

Prediction of ILR OPI Test Scores. For the comparison of the VAT to the ILR OPI, a scaled average OPI score was computed for each test-taker from all the available ILR OPI ratings. The scaling was performed using a computer program, FACETS, which takes into account rater severity and test-taker ability and therefore produces a fairer estimate than a simple average (Linacre et al., 1990; Linacre, 2003).

Figure 2 is a scatterplot of the ILR OPI scores and VAT scores for the concurrent validation sample (N=118). IRT scaling of the ILR scores allows a mapping of the scaled OPI scores and the VAT scores onto the original OPI levels, which are given on the inside of the plot axes. The correlation coefficient of the two test scores is r=0.87. This is roughly in the same range as both the ILR OPI reliability and the average ILR OPI inter-rater correlation. The test scores on the VAT account for 76% of the variation in the ILR OPI scores (in contrast to 83% accounted for by another ILR OPI test administration and 81% accounted for by one other ILR OPI interviewer).

The VAT accounts for most of the variance in the interview-based test of oral proficiency in MSA. This is one form of confirming evidence that the VAT captures important aspects of MSA speaking and listening ability.

The close correspondence of the VAT scores with ILR OPI scores, despite the difference in construct, may come about because candidates easily transfer basic social and communicative skills acquired in their native language, as long as they are able to correctly and efficiently process (i.e., comprehend and produce) the second language. Also, highly proficient learners have most likely acquired their skills at least to some extent in social interaction with native speakers of their second language and therefore know how to interact appropriately.

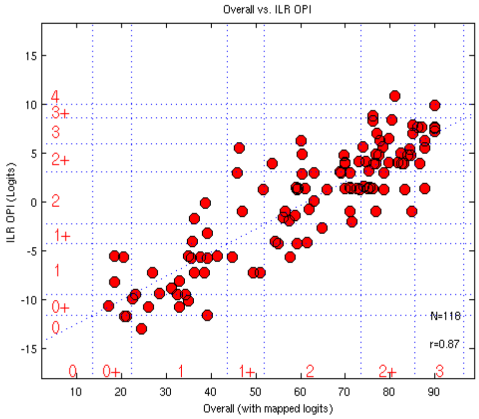

Figure 2: Test-takers' ILR OPI scores as a function of VAT scores (r=0.87; N=118).

Group Performance. Finally, we examine the score distributions for different groups of test-takers to investigate whether three basic expectations are met:

- Native speakers all perform well, while non-natives show a range of ability levels
- Non-native speakers spread widely across the scoring scale (the test can distinguish well between a range of non-native ability levels)
- Native speakers from different countries perform similarly (national origin does not predict native performance)

We compare the score distributions of test-taker groups in the training data set, which contains 1309 native and 1337 non-native tests. For each test in the data set, an Overall score is computed by the trained scoring system on the basis of the recorded responses. Figure 3 presents cumulative distribution functions of the VAT overall scores, showing for each score which percentage of test-takers performs at or below that level. This figure compares two speaker groups: Educated native speakers of Arabic and learners of Arabic. The

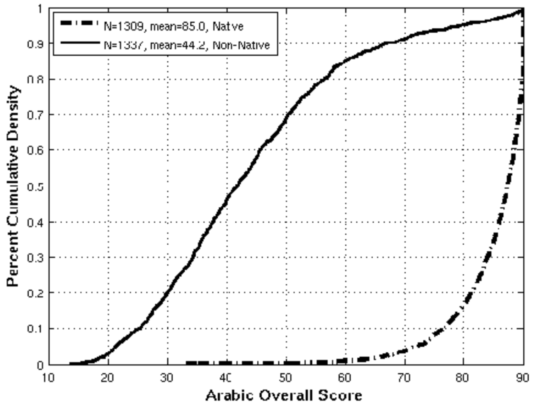

Figure 3: Score distributions for native and non-native speakers.

score distributions of the native speakers and the learner sample are clearly different. For example, fewer than 5% of the native speakers score below 70, while fewer than 10% of the learners score above 70. Further, the shape of the learner curve indicates a wide distribution of scores, suggesting that the VAT discriminates well in the range of abilities of learners of Arabic as a foreign language.

Figure 4 is also a cumulative distribution functions, but it shows score distributions for native speakers by country of origin (showing only countries with at least 40 test-takers). The curves for Egyptian, Syrian, Iraqi, Palestinian, Saudi and Yemeni speakers are indistinguishable. The Moroccan speakers are slightly separate from the other native speakers, but only a negligible number of them scores lower than 70, a score that less than 10% of learners achieve. This finding supports the notion that the VAT scores reflect a speaker's facility in spoken MSA, irrespective of the speaker's country of origin.

6 Conclusion

We have presented an automatically scored test of facility with spoken Modern Standard Arabic (MSA). The test yields an ability profile over four subscores, Fluency and Pronunciation (manner-of-speaking) as well as Sentence Mastery and Vocabulary (content), and generates a single Overall score as the weighted average of the subscores. We have presented data from a validation study with native and non-native test-takers that shows the VAT to be highly reliable (test-retest r=0.97). We also have presented validity evidence for justifying

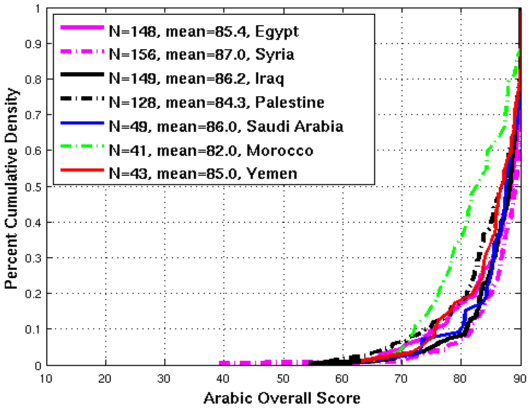

Figure 4: Score distributions for native speakers of different countries of origin.

the use of VAT scores as a measure of oral proficiency in MSA. While educated native speakers of Arabic can score high on the test regardless of their country of origin because they all possess high facility in spoken MSA, learners of Arabic score differently according to their ability levels; the VAT test scores account for most of the variance in the interview-based ILR OPI for MSA, indicating that the VAT captures a major feature of oral proficiency.

In summary, the empirical validation data suggests that the VAT can be an efficient, practical alternative to interview-based proficiency testing in many settings, and that VAT scores can be used to inform decisions in which a person's listening and speaking ability in Modern Standard Arabic should play a part.

Acknowledgments

The reported work was conducted under contract W912SU-06-P-0041 from the U.S. Dept. of the Army. The authors thank Andy Freeman for providing diacritic markings, and to Waheed Samy, Naima Bousofara Omar, Eli Andrews, Mohamed Al-Saffar, Nazir Kikhia, Rula Kikhia, and Linda Istanbulli for support with item development and data collection/transcription in Arabic.

References

Jennifer Balogh and Jared Bernstein. 2007. Workable models of standard performance in English and Spanish. In Y. Matsumoto, D. Oshima, O. Robinson, and P. Sells, editors, *Diversity in Language: Perspectives and Implications* (CSLI Lecture Notes, 176), 271-292. CSLI, Stanford, CA.

Jared Bernstein and Isabella Barbier. 2001. Design and development parameters for a rapid automatic screening test for prospective simultaneous interpreters. *Interpreting, International Journal of Research and Practice in Interpreting*, 5(2): 221-238.

Jared Bernstein, Michael Cohen, Hy Murveit, Dmitry Rtischev, and Mitch Weintraub. 1990. Automatic evaluation and training in English pronunciation. In *Proceedings of ICSLP*, 1185-1188.

Linda Crocker and James Algina. 1986. *Introduction to Classical & Modern Test Theory*. Harcourt Brace Jovanovich, Orland, FL.

Anne Cutler. 2003. Lexical access. In L. Nadel, editor, *Encyclopedia of Cognitive Science*, volume 2, pp. 858-864. Nature Publishing Group.

Maxine Eskenazi. 1996. Detection of foreign speakers' pronunciation errors for second language training – preliminary results. In *Proceedings of ICSLP '96*.

Maxine Eskenazi and Scott Hansma. 1998. The fluency pronunciation trainer. In *Proceedings of the STiLL Workshop*.

Horacio Franco, Victor Abrash, Kristin Precoda, Harry Bratt, Raman Rao, John Butzberger, Romain Rossier, and Federico Cesar. 2000. The SRI EduSpeak system: Recognition and pronunciation scoring for language learning. In *Proceedings of InSTiLL*, 123-128.

Jörg Jescheniak, Anja Hahne, and Herbert Schriefers. 2003. Information flow in the mental lexicon during speech planning: Evidence from event-related potentials. *Cognitive Brain Research*, 15(3):858-864.

Willem Levelt. 2001. Spoken word production: A theory of lexical access. *Proceedings of the National Academy of Sciences*, 98(23):13464-13471.

John Linacre. 2003. *FACETS Rasch measurement computer program*. Winstep, Chicago, IL.

John Linacre, Benjamin Wright, and Mary Lunz. 1990. A Facets model for judgmental scoring. *Memo 61*. MESA Psychometric Laboratory. University of Chicago. Retrieved April 14, 2009, from http:// http://www.rasch.org/memo61.htm.

George Miller and Stephen Isard. 1963. Some perceptual consequences of linguistic rules. *Journal of Verbal Learning and Verbal Behavior*, 2:217-228.

Pearson. 2009. *Versant Arabic test – test description and validation summary*. Pearson. Retrieved April 14, 2009, from http://www.ordinate.com/technology/VersantArabicTestValidation.pdf.

Hagen Soltau, George Saon, Daniel Povy, Lidia Mangu, Brian Kingsbury, Jeff Kuo, Mohamed Omar, and Geoffrey Zweig. 2007. The IBM 2006 GALE Arabic ASR system. In *Proceedings of ICASSP 2007*, 349-352.

Brent Townshend, Jared Bernstein, Ognjen Todic & Eryk Warren. 1998. Estimation of Spoken Language Proficiency. In *STiLL: Speech Technology in Language Learning*, 177-180.

Dimitra Vergyri, Arindam Mandal, Wen Wang, Andreas Stolcke, Jing Zheng, Martin Graciarena, David Rybach, Christian Gollan, Ralf Schlüter, Karin Kirchhoff, Arlo Faria, and Nelson Morgan. 2008. Development of the SRI/Nightingale Arabic ASR system. In *Proceedings of Interspeech 2008*, 1437-1440.

Steve Young, Dan Kershaw, Julian Odell, Dave Ollason, Valtcho Valtchev, and Phil Woodland. 2000. *The HTK Book Version 3.0*. Cambridge University Press, Cambridge, UK.

Klaus Zechner and Xiaoming Xi. 2008. Towards automatic scoring of a test of spoken language with heterogeneous task types. In *Proceedings of the Third Workshop on Innovative Use of NLP for Building Educational Applications*, 98-106.

Klaus Zechner, Derrick Higgins, and Xiaoming Xi. 2007. SpeechRater™: A construct-driven approach to score spontaneous non-native speech. In *Proceedings of the Workshop of the ISCA SIG on Speech and Language Technology in Education*.