TECHNOLOGY FOR ORAL COMMUNICATION

ACKNOWLEDGMENTS

The papers in this volume were presented at the Technology in Second Language Learning (TSLL) conference held at Iowa State University on September 17–19, 2009.

We thank the invited speaker Robert Godwin-Jones. We also thank the Departments of English and World Languages and Culture as well as the Program in Linguistics at Iowa State University for their financial support.

EDITORS
John M. Levis & Kimberly LeVelle
TESL/Applied Linguistics
Iowa State University

TABLE OF CONTENTS

Introduction

Plenary Address

Selected Papers

LeVelle, K. R. & Levis, J. M. (2011) Using technology for teaching and learning oral communication**.** In J. M. Levis & K. R. LeVelle (Eds.). *Technology for oral communication* (pp. 1-3). Ames, IA: Iowa State University.

USING TECHNOLOGY FOR TEACHING AND LEARNING ORAL COMMUNICATION

Kimberly LeVelle

John Levis

Iowa State University

The papers included here are selections from the seventh annual conference of Technology in Second Language Learning Conference, held in conjunction with the first conference of Pronunciation in Second Language Learning and Teaching. The two conferences together brought in approximately 90 participants from 10 US states and 5 foreign countries. We were excited by the number of papers and posters at the conference as well as the lively and spirited discussions by the attendees. In fact, we frequently overheard conversations carrying over into breaks, lunches, and into the evening. We were pleased to even watch as presenters referred to papers from the previous day and at least one presenter adapted his presentation dramatically to connect it more fully with a conversation that had developed from earlier papers.

The theme for this year's TSLL conference was Technology for Oral Communication. This theme was interpreted widely by the presenters at the conference with presentations and posters about research into different skills (e.g. listening, speaking), languages (e.g. English, Chinese) and purposes (e.g. assessment, teaching methodologies). While much of the research connects technology advances with teaching, there were also posters and papers that used technology to understand more about different varieties of language or to make our analyzing of language faster and more systematic. These suggest interesting possibilities in speech language pathology, forensic linguistics, as well as language teaching.

Technology is becoming an increasingly integral part of teaching in general, and particularly for teaching language. However, using technology (other than tapes and CDs) in the teaching and learning of oral skills (listening, speaking, pronunciation) is still less common in foreign and second language classrooms. With improvements in computer software, connectivity, and hardware, the possibilities for using technology with listening and speaking have been increasing available to all teachers, not just the technologically savvy. The papers in this section provide a glimpse into the diversity of new and exciting research into the use of technology for teaching and learning languages.

In the plenary talk for the conference, Robert Godwin-Jones of Virginia Commonwealth University looks back at the historical development of computerized approaches to language learning in *Trends in Speech Technologies for Language Learning*. Starting with a similar talk given in the mid-1990s at Iowa State, Godwin-Jones describes the evolution of computer-assisted language learning applications as they have moved from being slow and largely bound to desktop application platforms to providing much greater flexibility, greater use of communicative options and being flexibly available anywhere through web applications. Godwin-Jones especially discusses the development of speech recognition tools which promise to be the next frontier in applications for language learning. His paper includes links to valuable resources for anyone interested in computer-assisted language learning.

In *The development and validation of a computerized task-based Chinese oral performance assessment tool,* Tai-Heng Shih, Hsin-Yu Chang, and Yi-Jen Huang of Michigan State University describe the development of a computerized oral performance test designed to replace a test that was both intensive in its use of teacher time and was felt to be insufficient in testing the students' oral skills. Various spoken language tasks meant to approximate authentic speaking activities were included in the computerized test, most of which were found to be adequate for testing oral performance. Interestingly, students seemed to prefer the old, one-on-one interview format which allowed them to memorize answers. In addition, they preferred talking to a person rather than recording on a computer. Overall, the new test appeared to better assess the oral abilities of the students, but the authors suggest that fully adopting the new test format's washback on teaching and learning is unclear.

Ghinwa Alameen of Iowa State University examines listening comprehension in *The Role of Video Subtitling in Listening Comprehension*. She examines how three types of subtitles affect English language learners' ability to understand video input. Her participants had one of three types of subtitling options: full-text subtitles, keyword subtitles, and summary subtitles, a middle ground between the other two in terms of reading and interpretive load. The three types of subtitles showed no differences in the comprehension scores achieved by the subjects, and there were varied attitudes toward the different types of subtitling, suggesting that both learning styles and reading abilities affect choices. Full-text subtitles seemed the most problematic because the reading load was so much greater. However, some students seemed to prefer them. The paper suggests that there is not one-size-fits-all approach to subtitling, although there was no difference in comprehension with the three subtitling options.

In *Determining L1 & L1 Degree of Accent From Phonetic Transcription,* Paul Rodrigues (Indiana University and University of Maryland Center for Advanced Study of Language) describes an automatic grading and accent classification program using the IPA transcription of English speech by English speakers and by nonnative speakers of English from four different L1 backgrounds (Spanish, Portuguese, Arabic, and Russian). The classification system performed successfully for English and three of the four non-English L1s (Portuguese being the exception). The goal of the classification system is to use IPA language transcription as a way to automatically classify accents. The results suggest that certain segmental features (such as consonant clusters) may be more reliable markers of accents.

This sampling of papers gives a small hint of the topics discussed during the conference. The use of technology for promoting oral communication will continue to be an important topic for years to come. Assessment of oral proficiency, individualized instruction, connections to speaking, listening and pronunciation practice, and the development of new technologies will all continue to impact the field of second language learning.

Godwin-Jones, R. (2011) Trends in speech technologies for language learning. In J. M. Levis & K. R. LeVelle (Eds.). *Technology for oral communication* (pp. 4-7). Ames, IA: Iowa State University.

TRENDS IN SPEECH TECHNOLOGIES

FOR LANGUAGE LEARNING

Robert Godwin-Jones, Virginia Commonwealth University

Today many of the speech tools for language learning are Web delivered. The shift away from desktop applications is by no means complete, but it seems to be inevitable, as does the ever greater use of mobile devices. When I last participated in a linguistics conference at Iowa State University, this development was just beginning. This was the Computers in Applied Linguistics Conference, in 1994. The examples I presented and discussed at that conference were mostly desktop applications, such as a multimedia enriched German children's story, created in HyperCard. However, I also showed a similar application, which used a new authoring and delivery system, the World Wide Web. I discussed the changes this new option presented, with its cross-platform, inexpensive authoring, providing any time, anywhere access. The main selling point of the Web, however, was then and remains today the collaborative possibilities it offered. I demoed how students were able to add their own text annotations to a story, which were then available for all students to read.

PLENARY TALK

Today, the Web offers us far richer collaborative and communicative options. Audio and video have become much more easily accessible, including in longer segments, through the use of Web streaming. JavaScript has become much more powerful than in 1994 and, through the Document Object Model (DOM), is able to access and manipulate virtually all the elements of a page, not just form fields and images, as was the case in the early days. The technique dubbed AJAX, for asynchronous JavaScript and XML, allows JavaScript to update elements of the page seamlessly by retrieving data in the background from a server (formatted in XML). This allows for data to be fetched from a database, based on a user's actions and used to update an exercise, offer remedial work, or move on to an appropriate next step. The technology is available now to create quite sophisticated Web programs, including oral language tutoring applications. Of course, most language instructors are not likely to invest the time and effort to create such a program. However, they are in large numbers using the many audio options available on the Internet today such as Skype or Google Voice. Many are creating podcasts, using free audio tools such as Audacity. Some are also taking advantage of commercial audio collaborative tools such as Wimba voice tools.

Of course, using computers to work with speech is nothing new in itself. Speech recognition and analysis goes back at least to the 1970's, when it was being used to help the hearing or speaking impaired. It was soon used as well in computer-aided language learning (CALL). When it comes to helping learners, the computer offers clear advantages over the classroom. It provides individualized, self-paced training in a non-threatening environment, allowing the user to practice as often as needed. Through electronic means, it is possible to supply a rich variety of native speaker voices, for help in modeling speech and improving aural comprehension. Computer programs are as well customizable to individual learners, and can be set up to do regular reviews at set intervals as well as to monitor and guide learner progress.

One of the ways a computer program can help students improve their pronunciation is to display a visual representation of speech. There have been programs since the early days of CALL that represent speech patterns as graphical displays. Programs such as Visi-Pitch display a waveform or a pitch contour which allow a user to compare visually learner and model speaker utterances. Unfortunately, spectrograms and other visual representations of human speech are not always helpful to students, who have difficulty in understanding the significance of the displays and in using the information to improve their speech. Offering students training in working with the displays has been shown to help, as does the addition of audio feedback (Hardison). Some programs additionally show an image of the mouth, showing physically how a sound is produced.

There are many tools available for speech analysis and display, both free and commercial. In addition to Visi-Pitch, KayPentax has an extensive set of voice tools that it markets. One of the most popular free tools is Praat. Praat is a powerful and versatile speech analysis tool and is used today in a large number of open source speech projects. These include such tools on the high end as the powerful video annotation program Anvil and more narrowly focused projects such as the Pinyin Tutor, for helping with the pronunciation of Mandarin. Unfortunately, the learning curve for working with a tool like Praat is rather high, thus discouraging non-technical language teachers from using it in instruction. In fact, the need for technical expertise has limited the number of available speech-focused programs for language learning. Many promising projects never make it out of the prototype stage, due to limited funding.

One of the more interesting developments in recent years has been the emergence of alternative display options for visualizing speech. Applications have been created which use game-like interfaces, in order to try to more fully engage and keep the interest of learners. These range from a bowling game in which pronunciation accuracy determines how many pins are knocked down to a driving game in which adherence to the road is determined by the performance on oral drills. Such games use the voice in place of a joystick to determine what happens in the game. Some speech tools such as Visi-Pitch 4 and Computer Speech Labs include games of several different types. Such an approach is likely to be effective particularly with younger users. It would be worthwhile to have more experimentation in this area, given the immense popularity of on-line games.

Speech analysis in computers goes beyond the ability to display graphic representations of speech or to interact through speech patterns. Automatic speech recognition (ASR) allows a program to analyze speech semantically and interact with a user as a kind of conversation partner. ASR systems are ubiquitous today, familiar to consumers through help desk interactions, automated phone systems, and dictation software. Early implementations of ASR used pattern matching through a template-based system, but today most use a highly sophisticated, probability-based approach called the Hidden Markov Model. These systems, although they have improved significantly in recent years, are still far from perfect. If ASR is not totally reliable for native speakers, one can imagine the problems which are bound to arise with non-native speakers. ASR is designed for predictable language use i.e. speech patterns resembling closely those of a native speaker with native-level grammatical and syntactical accuracy. More accurate ASR systems improve performance by limiting the nature of the voice input, either by restricting the user to short utterances or constraining the conversation to a tightly controlled lexical selection.

The restrictions necessary to boost ASR accuracy to acceptable levels make it less useful for many language-learning purposes. Language teachers want students to be interacting with flowing, natural speech, not with artificially shortened utterances. And, of course, they need the systems to be able to recognize und respond appropriately to non-standard speech. Thus, one of the special needs for ASR in language learning is the availability of an extensive learner speech database which can be used as a reference for frequently made errors as well as a resource for interpreting speech deviating from native speaker models. Collecting this kind of data can be difficult and expensive, in part due to privacy issues. Nevertheless, there is a great deal of interest within the CALL community in the use of ASR in language learning implementations. The immense promise of providing speaking practice with a disembodied partner of limitless patience is too tempting to resist, despite the obstacles. In fact, advances in computing power, database collection, and computational linguistics are allowing the development of ever more powerful and accurate ASR components, as one can see in such commercial products as Tell Me More and Rosetta Stone.

Yet such programs barely tap the enormous potential of ASR to customize language learning to individual learners. Well-constructed systems could offer benefits such as 1) custom feedback, i.e. extra practice for persistent problems, as indentified in the learner history, 2) algorithm-based interval refreshers, and 3) focus on particular needs/interests of an individual, who may need domain-specific language training. Current ASR-based software is not close to filling the bill. The shortcomings often start with a lack of clarity about the user's end goal, whether it be near-native fluency and accuracy, or comprehensibility/intelligibility. Although the implicit goal is normally the latter, this is frequently not made clear in the program or to the learner. This has implications in terms of the kind of feedback provided. It is important in such applications, which can be frustrating to learners, to provide as much positive reinforcement as possible. Expecting students to imitate accurately native-level speech is hardly a reasonable expectation. Feedback on ASR systems is not usually customizable. Given the potential for inaccurate analysis of learner speech, initially providing minimal feedback seems the best approach, as long as optional, more complete feedback is also available. Nothing can be more discouraging to learners than to receive a lecture on a language issue based on a misinterpreted language cue. Providing a rich array of feedback options mitigates program errors and speech anomalies while accommodating different learning styles and student needs. There are too few independent reviews and studies of ASR programs. Many of the articles on such applications are written by the developers themselves. It would be highly useful to have comparative studies as well as analyses of software used in different contexts, for examples, as an adjunct in a classroom taught course or as part of a self-study program. Engwall and Bälter's study does a nice job of highlighting some of the advantages of computer programs over classroom pronunciation practice, but more detailed studies would be welcome.

As mentioned at the outset, the Web is rapidly becoming the medium for development and deployment of speech technologies. Fortunately, the W3C has been developing standards for work in this area, namely the Speech Recognition Grammar Specification (SRGS) and the Semantic Interpretation for Speech Recognition (SISR). Standards are needed so as to allow more sharing of resources. This would be welcome in the mobile market where currently different development approaches preclude creation of programs which will run on all or even a majority of devices. The immense effort involved in creating ASR calls for cooperation and openness whenever possible. Projects such as CMU SPICE are experimenting with an alternative method of creating a database collection for less commonly taught languages not likely to interest commercial developers. They have created a Web recording interface which allows speakers of the language to contribute to the database. This kind of crowd sourcing of database creation, as well as experiments underway to dynamically generate learner grammars based on actual use of a speech system by learners are among the promising developments in this field (Sagawa et all).

REFERENCES

Engwall, O., & Bälter, O. (2007). Pronunciation feedback from real and virtual language teachers. *Journal of Computer Assisted Language Learning*, 20(3), 235-262.

Hardison, D. M. (2004). Generalization of computer-assisted prosody training: Quantitative and qualitative findings. *Language Learning & Technology*, 8, 34-52.

Sagawa, H., Mitamura, T. & Nyberg E. (2004). Correction Grammars for Error Handling in a Speech Dialog System. In Dumais S., Marcu D. & Roukos S., *HLT-NAACL 2004: Short Papers* (61-64), Boston: Association for Computational Linguistics.

Shih, T.H., Chang, H.Y., & Huang, Y. J. (2010). The development and validation of a computerized task-based Chinese oral performance assessment tool. In J. M. Levis & K. R. LeVelle (Eds.). *Technology for oral communication* (pp. 8-27). Ames, IA: Iowa State University.

THE DEVELOPMENT AND VALIDATION OF A COMPUTERIZED TASK-BASED CHINESE ORAL PERFORMANCE ASSESSMENT TOOL

Tai-Heng Shih

Hsin-Yu Chang

Yi-Jen Huang

Michigan State University

Task-based language performance assessment utilizes different types of tasks that simulate real-life activities to evaluate the test-takers' communicative ability (Bachman, 2002). This study reports the development of a computerized Chinese oral performance assessment tool and the potential inferences to be drawn about the state of the test-takers' communicative language ability in real-life situations. The assessment tool developed in this project was a classroom-based achievement test. Task types included "description," "giving instructions," "reacting in situations," and "structured speaking" (Luoma, 2004). The test was computerized, and incorporated both visual and auditory input to enhance task authenticity. Participants' oral responses were recorded online and evaluated by two Chinese speaking teachers. A five-point analytic rubric was used to score the test-takers' oral performance and included different criteria in 4 components: "General description," "Fluency," "Language Use," and "Tonal production" to reflect the construct being measured. Results showed that the Chinese task-based test had appropriate item facility (0.4 to 1.0), promising item discrimination where 62% of the test items exceeded 0.40, and significant inter-rater reliability (*r*=0.854, *p*<.01). This newly designed test has great potential for drawing inferences about the learners' general speaking ability. Participants' preference for this assessment tool and pedagogical implications are discussed.

Keywords: task-based approach, computerized assessment, speaking test

INTRODUCTION

Task-based language performance assessment utilizes different types of tasks that simulate real-life activities to evaluate test-takers' communicative ability (Bachman, 2002). This study reports the development of a computerized Chinese oral performance assessment tool and the potential inferences to be drawn about the state of the test-takers' communicative language ability in real-life situations.

This computerized task-based Chinese oral test was created to replace a Chinese oral exam which was given at the end of every second semester in the students' first year of Chinese class. The old oral exam was administered in a one-on-one interview format. A week before the interview, the students received a preparation sheet entitled 'Final oral exam questions' with thirty questions. Five of these questions would be randomly and orally given to each test-taker during the old oral exam. Since the students knew the test questions already, they could prepare and memorize their answers in advance. During the interview, questions were asked without context and students recited their prepared answers. However, memorizing sentences does not mean students learned how to use the language in real life. Examples of the questions in the old interview oral exam are described in (1):

(1) a. nǐ xǐ huān yì biān chī fàn, yì biān tīng lù yīn mā ?
Do you like having meals and listening to the recordings at the same time?
b. nǐ de xié hé nǐ péng yǒu de xié yí yang dà mā ?
Are your shoes as large as your friend's shoes?
c. wǒ yào dào jī chăng qù, nǐ néng gào sù wǒ zěn me zǒu mā ?
I'm going to the airport; can you tell me how to get there?

The old oral exam had several problems. The questions were not well-designed because the students could simply answer "Yes" for (1a) since this question could not elicit students' further description; (1b) seemed awkward because there was no context provided; (1c) did not present a map or reference places but asked the students to describe the route. Students could answer with a very simple sentence like "Go straight!" and finish the question without using the specific structures learned in class. Due to the lack of context and difficulties with elicitation in the old exam, the new oral exam was designed based on task completion with visual and aural aids.

In addition to the defects of the question content, a problem regarding construct validity arose in the old oral exam. Since the questions were orally given in Chinese to the students, they might produce grammatically perfect sentences which, however, were not the correct answers to the questions asked because they misheard the questions. In this case, it is controversial whether the teacher should mark down the students' speaking ability because of their low listening ability. To improve the construct validity of the oral test, this newly designed test presented the questions in simple written English to make sure that the students' speaking ability was appropriately tested. In addition, each written question was combined with its spoken Chinese translation played to test-takers in the environment of a Chinese conversation to facilitate test-takers' Chinese production.

This study tested the validation of the computerized task-based Chinese speaking test. Although a task-based approach has been well-recognized and accepted in teaching, there has been little research on the effectiveness of the task-based approach to assessment. Hence, the first research question can be stated as

> **Research Question 1:** Can a task-based approach be implemented on the development of an effective computerized oral proficiency test?

If the answer to Research Question 1 is positive, the following concern of this study is the participants' comments on and preference to the new test. The old interview oral exam was administered to the same group of students one semester prior to the new computerized test. This paper does not include the administration or the examination of the validation of the old test, but instead compares the participants' preference to these two types of oral test. Accordingly, the second research question is stated as

> **Research Question 2:** Do the participants prefer the new computerized oral test or the old interview test?

The following sections of this article contain the construct, test description, methodology, data analysis, results and discussion, and the conclusion.

Construct

The content of the new test comprised nine tasks modified from the nine lessons taught to the test-takers during the semester in which the test was administered. The topics of the lessons are listed in Appendix A**.** The textbooks used for this semester were *Level one, Part 1 Integrated Chinese simplified version* (Lesson 8 to Lesson 10) and *Level one, Part 2 Integrated Chinese simplified version* (Lesson 11 to Lesson 16).

Test Description

The assessment tool was a classroom-based achievement test. The tasks were designed to align with the syllabus of the Chinese foreign language course. Task types included "description", "giving instructions", "reacting in situations", and "structured speaking" according to Luoma's (2004) classification. The test was computerized, and incorporated both visual and auditory input to enhance task authenticity.

This newly-designed test provided context with visuals to facilitate interaction in a more authentic way, and allowed test-takers to control the time of responding to questions. To prevent students from cheating by overhearing other students' production during the test, we designed the computerized test into four sets of test questions, each of which contained five different topics (tasks) out of the nine topics taught during this semester and were presented in different orders. A total of 26 questions were designed for the four different sets of tests. To fairly design the difficulty of the four test sets, the students were asked to complete Questionnaire 1 entitled 'The difficulty of lessons' before they took the computerized test. Once the test sets had equivalent difficulty, each student was randomly assigned one of the four test sets during the oral exam (the adjacent students had different sets). In this way, all of the 26 items could be tested.

During the test session, the students had to log into the 'Audio Dropbox' online recording program created by the Center for Language Education and Research to record their oral production. The test questions were presented by Microsoft PowerPoint in a display-only format so that the students could not go back to answer previous questions if they did not finish them.

Each task was supposed to be finished after five minutes and the timer would start after students clicked the 'Start' button on the question slides. The timing clock would not begin until the students were finished reading the description of the task content/questions. After they practiced their answers and were ready to respond to the questions, they clicked 'Start' to proceed to the answering slide with the timing clock. The next question would pop up after the limited time which was set up according to the difficulty of the question. The timing clock was set up in this way for two reasons. First and foremost, it could reduce the degree of students' anxiety because they had sufficient time to both think about the questions and feel comfortable with their answers. Second, setting up the timer was to ensure that all of the students would have the same amount of time to answer the questions and to see whether or not the students could complete all questions within the constrained time.

METHOD

Participants

The student participants were learners of Chinese in the second semester of the first year; they were each enrolled in Chinese 102 at Michigan State University (MSU). The course is for beginning level learners and aims to establish students' listening, speaking, reading, and writing abilities. The Chinese 102 class was composed of five sections with a total of 79 enrolled students. The students consisted of English, Korean, and heritage Chinese speakers.

Raters

The raters in this study were two Chinese teachers who were also the researchers. One of the raters is a teacher for American children at both the Lansing Chinese School and the MSU Chinese School. The other rater is the participants' Chinese teacher in the Chinese language program at MSU.

Materials

Materials for this study consisted of an analytic rating scale for Chinese L2 oral performance (Appendix B); Questionnaire 1 entitled 'The difficulty of lessons' (Appendix C) and Questionnaire 2 entitled 'Comparison between the old oral interview test and the computerized task-based test' (Appendix D).

The 5-point analytic rating scale was adapted according to the TOEFL iBT Speaking rubric and the old version of evaluation guidelines for Chinese oral performance test from the Chinese program at MSU. The test-takers' oral performance was scored according to the rating scale which included different criteria in four components: "General description", "Fluency", "Language Use", and "Tonal production" to reflect the construct being measured.

The purpose of Questionnaire 1 was to understand students' perceptions about the difficulty of each lesson so that each of the four test sets could fairly contain lessons with different difficulty levels. In the questionnaire, we listed the grammar points and the task topic of each lesson to remind the students of the content they learned.

The second questionnaire was created to help teachers understand how students viewed the old oral interview exam and the computerized task-based test. It asked the test-takers what advantages and disadvantages both speaking tests had, what difficulties they encountered when taking both speaking tests, and which type of speaking test they preferred.

Data Collection

Data were collected from a pilot-test in Phase I and from the final version of the computerized task-based speaking test in Phase II.

Phase I (Pilot-test). To test the validity of this newly designed test, we administered the pilot test to eight students in April, 2009. The pilot test showed that the Chinese computerized task-based test had acceptable item facility, higher than 0.40 item discrimination, and significant inter-rater correlation (*r*=0.761, *p*<.05). Based on the results, this newly designed test had great potential for drawing inferences about learners' general speaking ability.

Some changes were made to the computerized test based on the pilot-test: (1) the time for answering questions was extended since the eight students could not finish their answers in time and (2) in order to give the students a direction for answering, a 'Hint' was added after the presence of a question.

Phase II (Revised Computer-based Speaking Test). Before taking the computerized task-based speaking test, the students were given a review sheet (Appendix E). On the sheet, students were notified of the format of the exam. Also, the sheet provided example questions for each lesson.

During the computerized test, each student was given five tasks in different orders so that they would not overhear their classmates' answering or be disturbed by their classmates' response**.** After taking the test, 72 students completed Questionnaire 2 with open-ended questions according to their personal feelings and thoughts in relation to the old oral interview test administered prior to the current study and the computerized task-based speaking test adopted in this study.

Data Analysis

The data reported here came from the two questionnaires, the analyses of item validity and inter-rater reliability. Questionnaire 1 investigated the difficulty of each lesson from the students' viewpoints and therefore each test set containing lessons with different difficulty levels could have an overall comparable difficulty level. Questionnaire 2 compared the oral interview test and the computerized task-based test. The reports of the questionnaires are described in the section of Results and Discussion.

Item validity was analyzed in terms of item facility and item discrimination. Inter-rater reliability was analyzed by using Pearson Product Moment Correlation with SPSS computer software. The findings are also reported in the section of Results and Discussion.

RESULTS AND DISCUSSION

Questionnaire 1

The results of Questionnaire 1 in Table 1 showed that the ranking of lesson difficulty from the most difficult to the easiest is: Lesson 12 (Dinning), Lesson 16 (Dating), Lesson 13 (Asking directions), Lesson 15 (Seeing a doctor), Lesson 10 (Transportation), Lesson 14 (Birthday party), Lesson 8 (Birthday party), Lesson 9 (Shopping), Lesson 11 (Weather).

Table 1

Difficulty rankings for each lesson.

Ranking	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**
Lesson	12	16	13	15	10	14	8	9	11
Scores	3.41	3.40	3.36	3.21	3.18	3.10	3.09	2.96	2.57

Each of the four test sets was assigned one or two difficult lessons, one middle-difficult lesson, and one easy lesson. The tasks and their order in each test set are listed in Table 2.

Table 2

The order of tasks in the four test sets.

	Set A	**Set B**	**Set C**	**Set D**
First task	In the cafeteria	Weather forecast	Giving directions	Advise your patient
Second task	Dating and moving out	In the cafeteria	Weather forecast	A morning of Little Wang
Third task	Weather forecast	A morning of Little Wang	In the cafeteria	Giving directions
Fourth task	Shopping	Giving directions	Transportation	In the cafeteria
Fifth task	Giving directions	Transportation	Birthday party	Weather forecast

Questionnaire 2

Seventy-two test-takers participated in Questionnaire 2 but eight of the questionnaires were incomplete. Among the valid 64 questionnaires, 44 test-takers preferred the old one-on-one interview speaking test because they felt (1) that talking to a person was more comfortable and realistic and they enjoyed being around people (27 people), (2) that they were more prepared and the test was easy because they knew the test questions in advance (four people), and (3) they felt less nervous with unlimited time for answering (13 people).

Twenty test-takers preferred taking the computerized task-based test because they felt that (1) talking to a teacher was intense and stressful since they were put on the spot even though they had unlimited time to answer questions, (2) they were less nervous and relaxed due to plenty of time to practice the answers, (3) visual reference and contextual references were helpful, (4) they did not have to memorize the answers in advance and test their genuine speaking ability and could test what they knew, (5) all testing questions were similar to real life situations, (6) the computerized task-based test was far more organized than the interview test, and (7) the computerized task-based test provided an objective grading method by excluding the possibility of teacher subjectivity.

Item Validity

Twenty-six questions formed four different test sets (Appendix F) for the computerized oral exam (the validity of the old oral interview test was not examined in this study). The analysis of item facility (IF) for each of the 26 items showed that IF ranged from 0.4 to 1.0 (only six items had IF over 0.9), which indicated that the difficulty of the items was appropriate. Most of the items with high IF were the initial questions of the topics that were designed to increase the students' confidence and to reduce their anxiety.

The analysis of item discrimination (ID) for each item showed that 16 items had ID over 0.40 (very good), six items had ID between 0.30 and 0.39 (reasonably good), and two items had ID between 0.20 and 0.29 (marginal items). Only two items had ID below 0.19 (revised items), one of which was the initial question on the topic of weather. The other one was "What is your favorite season? Why?" The high IF and 0 ID of this question resulted from the lenient standard when raters graded this kind of open-ended questions with the students' various answers. That 62% of the test items had ID over 0.40 and the overall IF was moderate indicated that the test was well-designed and valid in terms of discriminating students with good performance from those with bad performance.

Inter-rater Reliability

The analysis of inter-rater reliability was conducted by analyzing 32 of the 75 students' total average scores and the scores from the four categories of the rating scale. The results showed high inter-rater reliability for the five elements (Table 3). Among the five elements, the inter-rater reliability of tonal production was relatively low (but still significant). The reason for this may result from the background of the two raters. One of the raters is the test-takers' instructor who rated the tonal production strictly because he had high expectations of the participants' achievement. However, the other rater did not teach Chinese to college students and had wider tolerance of errors.

Table 3

Inter-rater reliability coefficients.

	Overall correlation	**General description**	**Fluency**	**Language use**	**Tonal production**
Pearson's *r*	.854**	.893**	.836**	.920**	.632**

**p< 0.01 (2-tailed).

CONCLUSION

The present study first analyzed the validity of the computerized task-based classroom achievement Chinese Speaking test. The results showed that this computerized test could be used to draw inferences about the students' speaking ability. The IF of each item, located between 0.4 and 1.0, indicated that the difficulty of the items was appropriate, except for the six items with an individual IF over 0.9. The items with high IF were designed as warm-up questions to enhance the test-takers' confidence.

According to Ebel's (1979) guidelines for ID, this test contained 16 very good items out of 26 (>0.40), six reasonably good items (0.30 to 0.39), two marginal items (0.20 and 0.29), and two poor items (<0.19). These findings made evident that most questions could distinguish good students from bad students in terms of performance. However, a few open-ended questions dealing with high IF and 0 ID impacted the rater's evaluation. A criterion should have been adopted for grading open-ended questions, since without a criterion both answers with grammatical sentences and creative answers with ungrammatical sentences were given points in this study.

The results from Questionnaire 2 answered Research Question 2. The fact that more participants preferred the old oral interview over the computerized task-based test could be credited to the fact that many students were not familiar with the procedure of the new test, or because they did not appropriately prepare for the final test. The students were given the opportunity to practice the new test beforehand but very few students took the initiative. As a result, many of the test-takers felt nervous while taking the new test. Another main reason that the students preferred taking the oral interview test is because they could memorize answers in advance. However, that kind of assessment tested a student's memorization skills instead of their speaking ability while the Chinese computerized task-based test seemed to accurately assess the students' Chinese language proficiency. Therefore, an achievement test should be designed to reflect whether or not students absorb what they learned in class, rather than give students test questions in advance.

Future Research

Questionnaire 2 indicated that 68% of the test-takers preferred the old one-on-one interview test, stating that they felt more comfortable and it was more realistic talking in person. However, since the computerized test showed some advantages, future research can incorporate the task-based approach in a new one-on-one interview test and further compare test-takers' preference for nontask-based interview tests, task-based interview tests, and task-based computerized tests and the analysis of validation of each test.

Future computerized tests may consider adapting a different recording system. This study used Audio Dropbox, which could not automatically stop recording while participants were reading the questions and practicing their answers. Hence, not only the students' final answers to the questions were recorded but also their practice sentences. However, these recordings may be valuable for second/foreign language teachers and researchers with regard to teaching speaking and investigating the process of language production.

This study suggests that the newly designed computerized speaking test was adequate for the involved Chinese class because it had high item facility and item discrimination. Also, part of the test-takers expressed in the questionnaire that they felt they had demonstrated what they had learned in class. It is worthy to conduct research on the washback effect to examine the impact of the task-based content and the computerized format on class teaching and learning.

REFERENCES

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

Ebel, R.L. (1979). *Essentials of educational measurement* (3rd ed.). Englewood Cliffs, NJ: Prentice-Hall.

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

Appendix A

Contents

第八课	学校的生活
Lesson eight	School Life
第九课	购物
Lesson nine	Shopping
第十课	交通工具
Lesson ten	Transportation
第十一课	天气
Lesson eleven	Weather
第十二课	吃
Lesson twelve	Dining
第十三课	问路
Lesson thirteen	Asking Directions

第十四课	生日舞會
Lesson fourteen	Birthday Party
第十五课	看医生
Lesson fifteen	Seeing a Doctor
第十六课	約 會
Lesson sixteen	Dating

Appendix B

Analytic Rating Scale for Chinese L2 Oral Performance

Score	**General Description**	**Fluency**	**Language Use**	**Tonal Production**
5	The response fulfills the demands of the task, with at most minor lapse in completeness It is highly intelligible and exhibits sustained, coherent discourse.	Can express spontaneously at length with a natural, smooth, and colloquial flow effortlessly. Speech is clear. It may include minor lapses, or minor difficulties with pronunciation, which do not affect overall intelligibility.	The response demonstrates good control of basic and complex grammatical structures that allow for coherent, efficient expression of relevant ideas. Contains generally effective word choice. Though some minor errors or imprecise use may be noticeable, they do not require listener effort.	Rarely mispronounces any tones in words and sentences; able to speak with correct tone sandhi.
4	The response addresses the task appropriately, but may fall short of being fully developed. It is generally intelligible and coherent, with some fluidity of	Can express fluently and spontaneously. But some conceptually difficult subjects can hinder the flow of the speech and unnatural pauses or	The response demonstrates fairly automatic and effective use of grammar and vocabulary, and fairly coherent expression of relevant ideas.	Tonal production is clear, occasionally mispronounces tones, but has mastered all tones. Inaccurate tonal productions do not interfere

	expression, though it exhibits some noticeable lapses in the expression of ideas.	errors might occur.	Response may exhibit some imprecise or inaccurate use of vocabulary or grammatical structures or be somewhat limited in the range of structures used. Such limitations do not seriously interfere with the communication of the message.	with meaning.
3	The response addresses the task, but development of the topic is limited. It contains intelligible speech, although problems with delivery and/or overall coherence occur; meaning may be obscured in places.	The speech is somewhat hesitant. When he/she searches for patterns and expressions, there are noticeable pauses and unnatural rephrasing that might occur. Speech is generally clear, with some fluidity of expression, though minor difficulties with pronunciation, or pacing are noticeable and may require listener effort at times.	The response is limited in the range and control of vocabulary and grammar demonstrated (some complex structures may be used, but typically contain errors). This results in limited or vague expression of relevant ideas and imprecise or inaccurate connections. Automatically of expression may only be evident at the phrasal level.	Tonal production is not always correct, but the meaning can be understood. Often mispronounces unfamiliar words; may not have mastered all tones.
2	The response is limited in content and/ or coherence is only minimally connected to the task, or speech is largely	Pausing for grammatical and lexical planning and repair is very evident, especially in longer stretches of free production.	Range and control of grammar and vocabulary are severely limited (or prevented) expression of ideas and connections	Frequently mispronounces tones. Inaccurate tonal production that impedes meaning. Difficult to

	unintelligible.	The speech is obscure and causes considerable listener efforts.	among ideas. Some low-level responses may rely on isolated words or short utterances to communicate ideas.	understand even with concentrated listening.
1	The response is very basic with memorized phrases, groups of a few words and formula. The content/coherence is hard to connect to the task, or speech is largely unintelligible.	Can make him/ herself understood in very short utterances. The speech delivery is choppy, fragmented, with long pauses and hesitations. Unclear pronunciation causes great and considerable listener efforts.	Range and control of grammar and vocabulary are extremely limited (or prevented) expression of ideas. Some very low-level responses may rely on only isolated words.	Hasn't mastered accurate tonal productions. Mispronounces most tones in word and sentence levels. The tonal production is not intelligible.

Appendix C

Questionnaire: The Difficulty of Lessons

According to your learning experiences and feelings, please select the degree of difficulty of each lesson/task topic.

Lesson 8: School life **Grammar points** : • After…, then… (他今天早上起床以后，就去朋友家玩儿。) • Doing V1 and V2 at the same time. (我们一边吃饭，一边看电视。) • In addition to……. (我除了学中文以外，还学英文。)	Very easy	Easy	Neutral	Difficult	Extremely difficult
Task Topic: Describe your routine of school life and you are required to use specific sentence patterns.	□	□	□	□	□

Lesson 8: School life **Grammar points** : • After…, then… (他今天早上起床以后，就去朋友家玩儿。) • Doing V1 and V2 at the same time. (我们一边吃饭，一边看电视。) • In addition to……. (我除了学中文以外，还学英文。)	Very easy	Easy	Neutral	Difficult	Extremely difficult
Task Topic: Describe your routine of school life and you are required to use specific sentence patterns.	□	□	□	□	□
Lesson 9: Shopping **Grammar Points**: • Measurement words and colors (一件红衬衫) • Amounts of money (一百五十块三毛四分钱) • Compare (这件衣服跟那件一样大)					
Task Topic: You are required to buy some clothes but you find out you got the wrong item, and you need to return and exchange it.	□	□	□	□	□
Lesson 10: Transportation **Grammar Points**: • First……, then… (先坐红线，再换绿线。)					
Task Topic: You are asked to give directions to the airport.	□	□	□	□	□
Lesson 11: Talking about the Weather **Grammar Points:** • Compare (我比你高多了。) • Again (昨天下雨，今天又下雨了。) (我昨天去打球了，我想明天再去打球。)					

Lesson 8: School life **Grammar points** : - After…, then… (他今天早上起床以后，就去朋友家玩儿。) - Doing V1 and V2 at the same time. (我们一边吃饭，一边看电视。) - In addition to……. (我除了学中文以外，还学英文。)	Very easy	Easy	Neutral	Difficult	Extremely difficult
Task Topic: Describe your routine of school life and you are required to use specific sentence patterns.	□	□	□	□	□
Task Topic: Look at the weather forecast and describe the kind of weather or compare the weather on different days.	□	□	□	□	□

Lesson 12: Dining **Grammar Points:** - 来 (来一盘糖醋鱼。） - Give back wrong change (你找错钱了，多找了我一块。)	Very easy	Easy	Neutral	Difficult	Extremely difficult
Task Topic: You are asked to order dishes with a limited budget but the cashier gives you the wrong change.	□	□	□	□	□
Lesson 13: Asking Directions **Grammar Points:** - Direction and location words (图书馆在 Wells Hall 北边。) (往北走，往右拐。)					
Task Topic: You are asked to read the map and give the directions of how to get to some specific places.	□	□	□	□	□
Lesson 14: Birthday Party					

Grammar points : • To describe the time/place etc. that we know already happened (他**是**去年生**的**。) • Both …and… (他**又**高**又**帅。) • Time duration of an action (我学中文**学了一年**。)					
Task Topic: Describe a person's personal information such as: when was he born and how does he look.	□	□	□	□	□
Lesson 15: Seeing a Doctor **Grammar Points**: • Indicating an extreme degree (我的头疼**死**了。) • Times of an action (昨天我吃了**三次**药。) • 对 (我**对**狗过敏。) • More and more (我的中文**越来越**好。)					
Task Topic: Describe a patient's symptoms and advise him when to take medicine.	□	□	□	□	□
Lesson 16: Dating **Grammar Points**: • Potential complements (法文太难，我**学不会**。) • Directional complements (他走下楼**来**。) • 把+Object+Verb+complement/了 (你**把**这个字写错了。)					
Task Topic: Schedule a time and place to meet with your friend. Have your friend help you move your furniture out of your dorm.	□	□	□	□	□

Appendix D

Questionnaire

Comparison between the old oral interview test and the computerized task-based test

Please answer the following questions according to your testing experiences and feelings.

Open-ended questions

1.What are the advantages and disadvantages of the one-on-one interview speaking test?

2.What are the advantages and disadvantages of the computer-based speaking test?

3.What are the difficulties of taking the one-on-one interview speaking test?

4.What are the difficulties of taking the computer-based speaking test?

5. Which type of tests do you prefer as a final oral test? Why?

Thank you for your help!!

Appendix E

Review sheet

口语考试复习卷

Format of the exam:

In the test, you are required to orally answer the question on each slide. You will hear and see the question first, and you can press HINT to get some information to prepare your answer (no time limit for reading the question). When you are ready to answer it, press START and you have 10 to 60 seconds to answer depending on the question. When time is up, you will see the next question on the next slide.

Questions: (The questions listed here are not exactly the same as those in the oral exam. Don’t memorize your prepared answers but use these questions to understand how to respond to the certain situation.)

Daily life

1. 你平常早上一起床就做什么？几点到学校去上课？

2. 你除了上中文课以外还上什么课？

Go Shopping

3. 要是你在商店买衣服，你要跟服务员怎么说？

4. 要是你买的东西大小或者长短不合适，你要怎么说？

5. 要是你很想买这个东西，可是你带的钱不够，你怎么办？ (use 虽然…可是…)

Get direction

6. 如果你要去机场，你怎么走？

7. 请你说说图书馆在哪儿？

8. 说说从图书馆到学生活动中心怎么走？

Weather

9. 说说今天的天气怎么样？ (use 不但…而且…)

10. 说说下个星期的天气会怎么样？ (use 得多，多了，一点儿)

11. 你最喜欢哪个季节？为什么？

Go to the restaurant

12. 你在中国饭馆吃饭的时候，你点什么菜？点什么汤？

13. 你能吃辣吗？要是你一点辣都不能吃，你怎么告诉服务员？

14. 你平常在学生餐厅吃饭吗？要是你忘了带饭卡，你怎么付钱？

15. 店员把钱找错了，你怎么跟他说？

Birthday Party

16. 你是哪一年生的？属什么的？你今年多大？

17. 你最喜欢哪个电影明星？你觉得他长得怎么样？ (use 又…又…)

18. 你觉得你长得像你的爸爸还是你的妈妈？

19. 要是你开生日舞会，你要开几个小时？

20. 要是你最好的朋友过生日，你要送给他什么生日礼物？

21. 要是你的朋友给你生日礼物，你说什么？

Go to see a doctor

22. 要是你生病了，没去上中文课，你怎么告诉老师？

23. 你对什么过敏？

24. 要是你是医生，你怎么告诉你的病人什么时候得吃药？

25. 要是你的朋友不吃药，你要怎么告诉他得吃药？ (use 要不然, and 越来越)

Dating and Moving out

26. 你和你的朋友要一起去看电影，但是他没有车，你要去接他，你怎么跟他说？

27. 你要去接你的朋友，但是你找不到他家，你要怎么说？

28. 你的朋友要帮你搬家，你怎么告诉她，什么东西得搬出去，什么东西得搬进来？

Appendix F

Chinese 102 SS09 final oral exam questions

Task topic: In the cafeteria

Q1: What do you want to order?

Hint: Please tell the master chef you want to order one meat, one vegetable, and one drink from the menu.

Q2: In addition to what you have ordered, anything else?

Hint: If you don't like MSG, what do you say to the master chef?

Q3: The total amount is $28. Can you give me your meal card?

Hint: Tell the master chef you forgot to bring your meal card. Ask her if you can pay her 30 dollars.

Q4: No problem. One dollar is your change, right?

Hint: The master chef gave your change back but one dollar less. How do you tell her this situation?

Task Topic: Dating and moving out

Q5: It's great to see a movie! But what can I do if I don't have a car?

Hint: Tell your friend you will pick her up. Ask her to wait for you downstairs of her house at 6:30pm

Q6: It's already 7:00pm. How come you haven't arrived?

Hint: you are lost. Tell your friend you are not able to find her house

Q7: Where to put these things?

Hint: After the movie, tell your friend to move the bed upstairs and the computer downstairs.

Task topic: Weather forecast

Q8: How is the weather on Tuesday?

Hint: Use 不但…而且… to describe the weather on Tuesday.

Q9: How is the weather for next week?

Hint: Choose any two days from the above weather forecast and use either 一点儿，得多，or 多了 to compare two days' weather.

Q10: What is your favorite season? Why?

Hint: Describe your favorite season and your reason.

Task topic: Shopping

Q11: Hello! What do you want to buy?

Hint: Tell the salesclerk you want to buy these things.

Remember to say 'quantity' and 'color'.

Q12: Is this pair of shoes ok?

Hint: What do you want the clerk to do? What size do you want?

Q13: This pair of shoes is good. Do you want to buy it?

Hint: Tell the salesclerk this pair is too expensive although it is the right size.

Task topic: Giving directions

Q14: Where is the cafeteria?

Q15: How to get to the computer center from the hospital?

Hint: Please give the direction from the hospital to the computer center.

Task topic: A morning of Little Wang

Q16: What did Little Wang do this morning?

Hint: Describe what Little Wang did this morning.

Q17: What classes did Little Wang have this morning?

Hint: Use 除了….以外，还……

Task topic: Transportation

Q18: How can I get to the airport?

Hint: Describe how to get to the airport by taking bus #3 and subway, and how many stops she has to take.

Task topic: Birthday party

Q19: (1) When was Little Li born? (2) What year does she belong to?

Hint: Use 是…的 for (1)

Q20: (1) How does Little Li look? (2) Who does Little Li look like?

Hint: Use 又…又… for (1)

Q21: How many hours will we hold the dance party?

Hint: Answer with a complete sentence.

Q22: Which one would you like to give to Little Li as a gift?

Hint: Answer with a complete sentence.

Task topic: Advise your patient

Q23: What's wrong with him?

Q24: He feels itchy. Is he allergic to something?

Hint: What is he allergic to? (flower in the picture)

Q25: How many times should he take the medicine? When to take it?

Hint: Tell your patient to take the medicine three times a day after meals.

Q26: What do you say to him if he doesn't want to take the medicine?

Hint: Advise him to take the medicine. Otherwise, he'll get sicker and sicker.

Alameen, G. (2010). The role of video subtitling in listening comprehension. In J. Levis & K. LeVelle (Eds.), *Technology for oral communication* (pp. 28-39), Ames, IA: Iowa State University.

THE ROLE OF VIDEO SUBTITLING IN LISTENING COMPREHENSION

Ghinwa Alameen, Iowa State University

With technological advances, integrating subtitles with videos has shown potential as a useful educational tool, granting students access to the textual representation of the audio message in the video which may enhance language comprehension. Most studies of subtitling have examined the use of subtitles that transcribe the entire text, overlooking the role which modified subtitles may have in mitigating their disadvantages. The present study seeks to investigate the role of three types of subtitles in enhancing listening comprehension for non-native speakers of English. Full-text subtitles will be compared to keyword subtitles. The use of summary subtitles which offer a middle ground between the other two methods, is also examined. They provide more information than keywords without sacrificing linguistic context or overloading learners. Finally, learners' attitudes towards the usefulness of different types of subtitles in enhancing listening comprehension are investigated.

INTRODUCTION

Many researchers have investigated the learning potential of video subtitles, especially their effectiveness in improving listening and speaking abilities and language comprehension in general (Lambert, Boehler & Sidoti, 1981; Graza, 1991; Vanderplank, 1988, 1990; Borrás & Lafayette, 1994; Guillory, 1999; Rozendaal, 2005; Grgurović & Hegelheimer, 2007). These studies, despite their differences, agree on the positive influence of subtitles on learner's comprehension of video. Yet many teachers are still reluctant to use subtitled videos in the classroom because students may become overly dependent on the text and overloaded with information from multiple channels rather than focusing on listening input (Vanderplank, 1988; King, 2002). Most of the studies on subtitling have examined the use of subtitles that transcribe the entire text (hence, *full-text* subtitles) overlooking the role that modified subtitles may have in mitigating disadvantages of full-text subtitles.

Keyword subtitling is one way of modifying subtitles that replaces clauses with one word. Although short and easy to read, they sometimes provide too little information or isolated words that can sound out of context. Another type of modification between full-text subtitles and keyword subtitles is *summary* subtitles. These offer more information than keywords without sacrificing linguistic context or overloading learners with excessive written input. However, no previous studies have examined this type of subtitling and its effectiveness in language classrooms. The present study seeks to investigate the role of those three types of subtitles in enhancing listening comprehension for non-native speakers of English (NNS) and to help find more effective ways to design and use computer assisted language learning (CALL) materials.

Study Rationale

The use of subtitles in this study is founded upon an interactionalist model of second language acquisition (Long, 1996). According to this model, noticing input is an essential condition for acquisition and can be enhanced through modifying input. Chapelle (2003) suggested that multimedia CALL can help in input comprehension by, for example, modifying the input mode from audio to text, which will improve engagement in language tasks. Using subtitles modifies aural and visual input and creates opportunities for noticing. This study contributes to the existing research by examining which type of textual modification of multimedia is more effective for language comprehension.

Subtitles enhance multi-channel processing by allowing students to use "multiple language processing strategies" (Graza, 1991, p. 246). In addition to video images, learners read and connect meanings of the text on the screen to the aural and visual input they are receiving. All these categories of information are "fed simultaneously to an attention moderator in the brain which filters information for the next processing component" (Guillory, 1999, p. 97). According to this model, however, an interruption in the multichannel input can cause input to be otherwise sequential which may result in the loss of information from one or more channel(s). This will ultimately result in a breakdown of comprehension. Such breakdown may happen, for example, when students focus more on reading the textual information offered by full-text subtitles instead of distributing their attention more evenly. Having less text in subtitles will reduce the reading load, consequently enhancing opportunities for comprehension. In this respect, keyword subtitles and summary subtitles may relieve the memory load by offering less text to process.

Literature Review

Proponents of the use of subtitles in second language teaching argue that subtitles can improve language proficiency by helping learners notice language that they might not otherwise comprehend (Price, 1983; Graza, 1991; Vanderplank, 1988, 1990; Guillory, 1999; Pujolà, 2002; Rozendaal, 2005; Grgurović & Hegelheimer, 2007). This is especially true in the case of keyword subtitles, with which words can be made more noticeable, and hence more likely to be unconsciously acquired (Ellis, 1999, p. 48).

In one of the pioneer studies on the use of subtitles in the classroom, Price (1983) found that "viewers, regardless of educational level or language background, benefited significantly from captioning, even with only one viewing" (p. 8). Similarly, Garza (1991) noted a significant positive impact of subtitles on the participants' comprehension of Russian and English videos. In a series of studies, Vanderplank (1988, 1990) investigated the effect of TV subtitled programs on learners' language abilities. Participants found the subtitles helpful to their language development, and were able to develop strategies for using them flexibly and efficiently. More recently, Grgurović and Hegelheimer (2007) noted that students interacted with subtitles more frequently and for a longer time than with transcripts. Since previous research has established the benefit of using subtitles in the classroom, this study will not ask participants to watch videos with no subtitles, but rather focus on the effectiveness of different types of subtitles on listening comprehension.

Assigning students into groups of no subtitles, full-text subtitles, and keyword subtitles, Guillory's (1999) participants in the full-text captions group outperformed those of the two other groups with no significant differences between the full text captions group and the keyword captions group. Similarly, the appearance of individual keywords drew learners' attention to specific content in the video providing more enhanced input in a small-scale study conducted by Rozendaal (2005). Guillory and Rozendaal also investigated students' attitudes toward subtitles. In their studies, students had more positive attitudes toward subtitled videos as opposed to videos with no subtitles and many said that keyword subtitles helped them identify words better than full-text subtitles. This study will survey students' opinions of the three types of subtitling and explore their influence on students' language comprehension.

Although research findings on video subtitles have shown that full-text subtitles may enhance listening comprehension and keyword subtitles may have a similar influence (Guillory, 1999), none has examined the effect of summary subtitles in that respect. This paper compares the three types of subtitling: full text, keyword, and summary; in addition to exploring participants' attitudes to using these types of textual help in a CALL activity. The research questions for this study are:

1. Which type of subtitles has the best positive effect on learners' listening comprehension?
2. What are the participant's attitudes towards the usefulness of the type of subtitles they worked with?

METHODS

Participants

The study used 27 undergraduate ESL students enrolled in a course of listening strategies for non-native speakers of English, at a major research university in the United States. Of the 27 participants, 15 were females and 12 males (14 Chinese, 10 Korean, two Japanese, one Taiwanese, and one Indonesian). Their age range is 18-35 with the majority between 18-21. The participants' English proficiency was determined by considering their TOEFL scores, their current grade in the course and the instructor's evaluation of their listening proficiency. Accordingly, they were divided into 14 lower-intermediate and 13 higher-intermediate students.

Before the onset of the treatment, participants provided information on their experiences with video in ESL classes. Forty eight percent of the students said they used videos for learning English as a second language before, mostly in school. Of all students, 69% thought that using videos in the classroom was helpful, 20% thought it was amusing, while only 11% thought it was boring. It is to be noted, though, that this last group of students had not used videos in the classroom before. Finally, only 46 percent of the participants had some previous familiarity with using subtitles for studying English.

Materials

The main data elicitation tool was a CALL activity designed specifically for this study and based on video recordings of one tertiary-level lecture "Are we alone in the universe? The drake equations" by Steven Kawaler from the Department of Physics and Astronomy at Iowa State University. The lecture was divided into nine 60-90 second segments and each segment was modified in three different ways by adding one type of subtitling: full-text subtitles, keyword subtitles, and summary subtitles. This resulted in 27 video segments, nine for each type of subtitles. The distribution of these types was counter-balanced resulting in three different arrangements so that every student experienced all three types (see Table 1). The difficulty level of the videos was determined to be suitable for the participants by analyzing the videos using VocabProfiler 2.7 (Cobb, 1994; Heatley & Nation, 1994), and consulting the class instructor. The lecture transcript was found to consist mainly of K1 words (83%), 2.6% K2 words, and 5.5% academic words.

Table 1

Participants' Groups According to Arrangement of Subtitle Types

Group 1	Full-text subtitles	Summary subtitles	Keyword subtitles
Group 2	Keyword subtitles	Full-text subtitles	Summary subtitles
Group 3	Summary subtitles	Keyword subtitles	Full-text subtitles

All subtitles appeared at the bottom center of each video segment in white san serif font on a black background, and were synchronized with the audio. Full-text subtitles were divided into phrases of 10-12 words and remained on the screen for the whole duration of the spoken utterance (Schelinger, 1968). Keyword subtitles and summary subtitles are modifications to full-text subtitles that are intended to provide the students with less text to read, yet more time to process the aural message. Keywords were selected by a preliminary study adopted from Guillory (1999). Two native speakers of English worked on determining keywords of the video segments according to their significance to the understanding of the utterance. All words agreed upon by at the two raters were kept as keywords resulting in 11% of the original transcripts. Every keyword was kept on screen for one second (Marleau, 1981) giving students enough time to read, and keeping the screen clean for the majority of the video play time.

Summary subtitles were decided upon by the same panel of raters. They summarized the transcript of the video using complete sentences and preserving the original wording of the text whenever possible. Hence, students would not be distracted by the appearance of new vocabulary not included in the video. Furthermore, having grammatical simple sentences placed words in a familiar syntactic context that students were probably more comfortable reading than longer complicated sentences or isolated words. Raters eliminated unnecessary information that could potentially be distracting, such as paraphrases of ideas. The resulting text was approximately 50% of the original. The summary subtitles appeared at or a little ahead of the onset of the corresponding original statement and stayed there for 50% of the time of this statement. As a result, the video segments had balanced periods of text and no text on the screen. As in full-text subtitles, they were divided into phrases of 6-9 words to allow for comfortable reading by the students.

Every video segment was followed by two multiple-choice questions that required students to use information from the video and to draw conclusions based on their understanding of the video. Students' answers were used to determine their comprehension level. Eighty percent of the questions were adapted from Grgurović (2005). The resulting 18 questions were judged to be proper for the students' proficiency level by their instructor; in addition, most of them were piloted on two students of a comparable listening proficiency (Grgurović, 2005).

Procedures

The study was integrated into the syllabus of the listening class and conducted as a part of a regular class period. Students were given the option of participating in the study and their data were only used upon having their informed consent. The study took place over one class period of 50 minutes. Participants completed the pre-listening questionnaire and went through the videos and the accompanying questions followed by the post-listening questionnaire. They had minor control over the activity and could only progress in a linear fashion. Once the play button was clicked, the video started playing followed by the comprehension questions. They received verbal instructions from the researcher on how to navigate the activity which were accompanied by illustrations on the lab projector. When students had answered all questions, a report of their score was displayed and saved on the computer. Student data was saved on Moodle, a course management system.

RESULTS AND DISCUSSION

The first research question in the study investigated which type of subtitles has the best effect on learners' listening comprehension. To determine this, descriptive statistics were used. Table 2 outlines the means and standard deviations of the participants' scores in the three types of subtitles. The statistical results detect a difference in means between full-text subtitles and keyword subtitles on one hand, and summary subtitles on the other. This difference has a medium effect size on a Cohen-d measure (Cohen, 1988). Correct answers to questions corresponding to summary subtitles exceeded those corresponding to the other two types.

Table 2

Descriptive Statistics of Participants' Scores

	N	Mean	Standard Deviation
Full-text	27	2.37	1.275
Keyword	27	2.26	1.023
Summary	27	2.85	1.292

A one-way ANOVA was applied to the participants' comprehension scores ($\alpha = 0.05$). The results of that analysis (Table 3) showed no statistically significant difference among the three categories. The results are in line with Guillory's (1999) finding in terms of similar efficacy of full-text and keyword subtitles. A post hoc analysis (HSD) to determine whether participants' language proficiency level influenced the outcome was unnecessary since there was no significant difference among the different types of subtitles.

Table 3

ANOVA Results

	Sum of Squares	df	Mean Square	F	Sig.
Between Groups	5.358	2	2.679	1.851	0.164
Within Groups	112.889	78	1.447		
Total	118.247	80			

Overall results indicate that all types of subtitles have comparable effects on learner's listening comprehension. Drawing on previous research finding (Guillory, 1999; Rozendaal, 2005) that keyword subtitles help learners identify words better than full-text subtitles, summary subtitles can be seen as more effective when considering their conciseness. Learners were able to attain better scores with less reading required than full-text subtitles. Hence, summary subtitles appear to be helping learners equally well while reducing mental load.

The second research question in the study examined student attitudes toward the role that different types of subtitles play in scaffolding listening comprehension. When asked about the amount of video they understood, 89% of the students said they understood the linguistic input a little, while 4% understood spoken language well, and only 7% did not understand the language at all. This indicates that the majority of students comprehended a moderate amount of video language. Only 8% of the participants preferred to watch the videos with no subtitles, 38% favored videos accompanied by full-text subtitles, 31% favored summary subtitles, and 23% had keyword subtitles as their preferred type. In terms of the type of subtitling that would allow for maximum focus on listening as opposed to reading subtitles, 64% of the participants chose keyword subtitles, while 24% and 12% felt more comfortable with summary subtitles and full-text subtitles, respectively.

By looking at previous research and participants' comments, it seems that different learning styles, speed and amount of reading and listening input and enhanced noticibility of linguistic input are possible reasons for the discrepancy of student scores in the different subtitle categories. What some students see as a help tool to understand language, others might conceive of as a hindrance. Full-text subtitles were seen by the majority of participants as too long to be read thoroughly in the period of time they remained on the screen. This made it difficult for students, if not impossible, to read them all. Not being able to capture the whole written message distracted their attention. A number of participants, however, expressed preference for full-text subtitles which, despite their length, provided an opportunity for scanning for the details. Not all participants appeared to be equally comfortable with scanning, this may have added to the difficulty of full-text subtitles. Summary subtitles, on the other hand, provide students with shorter text that can be read in a shorter amount of time giving them more time to process the aural message. However, not having parallel and simultaneous audio and textual messages was confusing for a couple of students who expressed their inability to "think and listen at the same time because subtitle is not following what lecture say", as one of them expressed. This reflects differences in learning styles as well as proficiency level since many higher-intermediate students favored summary subtitles because of the shorter processing time they offered.

Longer subtitles, therefore, pose higher chances for dependence on one channel (written) at the expense of the aural channel (Vanderplank, 1988). The use of full-text subtitles may "sacrifice listening strategy training such as guessing and inferring meanings from visual clues" (King, 2002, p. 517). This may explain why 64% of the participants chose keyword subtitles as the type that allowed for maximum focus on listening. One student expressed difficulties at multi-channeling since, in addition to the textual, aural, and linguistic messages, she experienced an extra processing level of L1 translation "full subtitles are likely to make me not listening and make me try to translate in my native language."

The rich content of multimedia material may also have a role in the amount of comprehensible input students get from the video lecture. The subtitles themselves may overload participants' capacity to comprehend language (Graza, 1991) by providing "too many words" as one participant commented. Manipulating subtitles so that they only contain ideas essential to the understanding of message and eliminate unnecessary information is a way to mitigate the heavy load students get with full-text subtitled videos. Summary subtitles, in particular, help students focus on the main points "which the speaker wants to tell, and understand the lecture even though I do not hear some parts" as a participant commented. Modifying input in such a manner may create more opportunities for noticing which, in turn, improves conditions for comprehension and acquisition (Chapelle, 2003).

Unlike keyword subtitles, summary subtitles provide a syntactic context familiar to learners and a scaffold to help interpret new vocabulary. Keyword subtitles, although more salient and less distracting for many students, may introduce isolated new vocabulary, which means that the student needs to deduce the meaning from the aural message. This process can be more complex for lower-proficiency students or simply those who are not trained to do so.

CONCLUSION

Although participants seemed to favor the use of full-text and summary subtitled video to assist them in comprehending the audio message, their performance using all three types of subtitles was similar. The fact that no one type of subtitles significantly outperformed the other should be looked at after taking into consideration the differences in learning styles and background students come from. In conclusion, to decide which type of subtitling to use with videos, there is no one-size-fits-all solution. Students can be made aware of differences among subtitle types and then given the option of choosing the one that fits their proficiency level, learning style, and last but not least, the objectives behind the activity at hand.

The possibility to modify subtitles allows the teacher to provide scaffolding appropriate to the level of learners. This will help the teacher use authentic video with learners from beginning levels as long as the linguistic message can be modified to suit their level. Different types of subtitling can be used to enhance noticing of certain elements in the linguistic input. They can be used to draw viewers attention to individual words (such as in keyword subtitles), or main ideas (such as in summary subtitles) and can enhance CALL material design and its use by learners and teachers.

If this study were to be replicated, several limitations should be addressed. First, the small sample size of participants did not allow for generalization of the results. Additionally, video materials, although thought to be suitable for students' level, appeared to be difficult for the majority of participants. Lastly, the study design and analysis did not account for the supplemental material that appeared on parts of the video. These contained equations and brief main ideas that the lecturer pointed at during the lecture. Such materials could have affected participants' performance on the study questions.

Work on subtitle modification can contribute significantly to CALL material design and evaluation. Future research can explore the use of other types of modifications such as summary subtitles that do not contain the exact words of the original script. Researchers and designers can also look at possibilities of enhancing noticibility of subtitles by using colors or font variation. The timing of subtitle appearance on the screen is another factor that may affect learner's listening comprehension and can be looked at. Finally, more research is needed on training learners on how to choose the type of subtitling that fits their level and needs and encourages an effective use of learning strategies.

Appendix

Pre-Listening Questionnaire

I. Background Data

Name ………………………………………….

Home country ………………………….. Native Language ………………………………

Gender Male Female

Age …………………….

Degrees/Major …………………….

Period of time living in USA ………….

Number of years of English training in your home country ………..

English classes taken at ISU ……………………….

……………………….

TOEFL score (if applicable) ………...............

II. Use of Video in learning a language

- Have you used video in learning a second language?
- If yes, where did you use it (school, self-study, etc)?
- What type of language activities did you use videos for? Listening, reading, speaking, writing, other?

Please circle all that apply to you:

- Using video in the classroom is: amusing waste of time boring helpful
- When video is used in language activities, you:
  - are distracted from understanding language,
  - understand language better,
  - do not watch the video at all.
- Do you like reading subtitles in videos in your native language?
- Do you like reading English subtitles in videos?

Post-Listening Questionnaire

Have you used videos with subtitles for studying English before?

If yes, where? ……….

For which class? ………..

What activities have you used them for?

Please check the answer that applies to you best and write any comments you have about it.

1) After watching the video

- ○ I did not understand the spoken language at all.
- ○ I understood the spoken language a little.
- ○ I understood the spoken language well.

Comments

2) You prefer to watch this video

- ○ Without any subtitles
- ○ With full subtitles
- ○ With keyword subtitles
- ○ With Summary subtitles

Comments

3) You focus on listening more than reading subtitles when the video has

- ○ Full subtitles
- ○ Keyword subtitles
- ○ Summary subtitles

Comments

References

Borrás, I. & Lafayette, R. (1994). Effects of multimedia courseware subtitling on the speaking performance of college students of French. *The Modern Language Journal*, 78(1), 61-5.

Chapelle, C. (2003). *English language learning and technology: Lectures on applied linguistics in the age of information and communication technology*. Amsterdam: John Benjamins.

Cobb,T. *Web Vocabprofile*. (1994). [accessed 29 October 2007 from http://www.lextutor.ca/vp/].

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Hillsdale, NJ: Erlbaum

Ellis, R. (1999). Factors in the incidental acquisition of second language vocabulary from oral input. In R. Ellis (Ed.), *Learning Second Language through Interaction* (pp. 35-61). Philadelphia: John Benjamin.

Garza, T. (1991). Evaluating the use of captioned video materials in advanced foreign language learning. *Foreign Language Annals,* 24 (3), 239-258.

Grgurović, M. (2005). *Research into the use of help options in a multimedia listening unit.* Unpublished MA thesis, Department of English, Iowa State University, Ames, IA.

Grgurović, M. & Hegelheimer, V. (2007). Help options and multimedia listening: students' use of subtitles and the transcript. *Language Learning and Technology*, 11(1), 45-66.

Guillory, H. (1999). The effect of keyword captions to authentic French video on learner comprehension. *CALICO Journal*, 15(1-3), 89-108.

Heatley, A. & Nation, P. (1994). *Range*. Victoria University of Wellington, NZ. [Computer program, available at http://www.vuw.ac.nz/lals/.]

King, J. (2002). Using DVD feature films in the classroom. *Computer Assisted Language Learning*, 15(5), 509-523.

Lambert, W., Boehler, I, & Sidoti, N. (1981). Choosing the languages of subtitles and spoken dialogues for media presentations: Implications for second language education. *Applied Psycholinguistics*, 2, 133-144.

Long, M. (1996). The role of linguistic environment in second language acquisition. In W. C. Ritchie, & T. K. Bhatia, (Eds.), *Handbook of second language acquisition* (pp. 413-468). San Diego: Academic Press.

Marleau, L. (1981). Les sous-titres … un mal nécessaire. *Meta: Journal des Traducteurs, 27* (3), 271-285.

Pujolà, J.-T. (2002). CALLing for help: Researching language learning strategies using help facilities in a web-based multimedia program. *ReCALL*, 14 (2), 235-262.

Price, K. (1983). Closed-captioned TV: An untapped resource. *MATSOL Newsletter*, 12, 1-8.

Rozendaal, C. (2005). *Fluency through Friends: Authentic video, subtitle modification, and oral fluency*. Unpublished MA thesis, Department of English, Iowa State University, Ames, IA.

Schelinger, I. (1968). *The syntactic constituent as a unit of decoding: Sentence structure and the reading process*. The Hague: Mouton.

Vanderplank, R. (1988). The value of teletext sub-titles in language learning. *ELT Journal*, 42(4), 272-281.

Vanderplank, R. (1990). Paying attention to the words: Practical and theoretical problems in watching television programs with uni-lingual (CEEFAX) sub-titles. *System*, 18(2), 221-234.

Rodrigues, P. (2010). Determining L1 & L1 degree of accent from phonetic transcription. In J. M. Levis & K. R. LeVelle (Eds.). *Technology for oral communication* (pp. 40-55). Ames, IA: Iowa State University.

DETERMINING L1 & L1 DEGREE OF ACCENT FROM PHONETIC TRANSCRIPTION

Paul Rodrigues
Indiana University, Linguistics Dept.
University of Maryland Center For Advanced Study of Language

This paper introduces a new computational method for the automatic grading and classification of L1 accent from IPA transcription of L2 speech in Computer Aided Language Learning (CALL) systems using statistical classification methods called Naïve Bayes Classification and Bayesian Belief Networks and speaker measurement using multidimensional scaling**.** The IPA transcriptions available in the Speech Accent Archive (Weinberger, 2005) were used to construct a corpus designed for this study. Forty-five speakers from five languages (Spanish, Portuguese, Arabic, Russian, and English) were selected for the analysis. Comparison of thirteen variations of the current corpus suggests that by filtering out vowels and by leaving only consonant phonotactics, a reliable native language classification system could be developed that performs with an F-Score of .86 and classifies L1 accent categories with 84.44% accuracy.

This research serves as an alternative to typical Digital Signal Processing techniques, in which the acoustic data is analyzed directly for accent signals. In this system, language transcription could be an output of a speech recognizer and accent identification is performed on this output. Having linguistic tokens act as targets for speech recognition allow for solving various problems of grading accent.

INTRODUCTION

Computational approaches to accent identification are often found in the digital signal processing and speech science literature, but approaches to accent identification and measurement based upon segmental transcription have been understudied. In this paper we adopt classification architectures common in speech recognition research and compare performance across phonetically transcribed corpora in various segmental phonotactic situations. We then take these transcription corpora, and perform multidimensional scaling to produce a graph. Measurement on this graph between an individual speaker and a prototypical cluster of speakers shows a speaker's degree of accent from that cluster.

Many systems have been designed to classify the language of orthographic text, but these language identification systems typically classify into a discrete language category broad enough to represent the entire language. For example, a segment of Japanese text would be classified as Japanese, not English, and assigned to the Japanese textual bin (Figure 1).

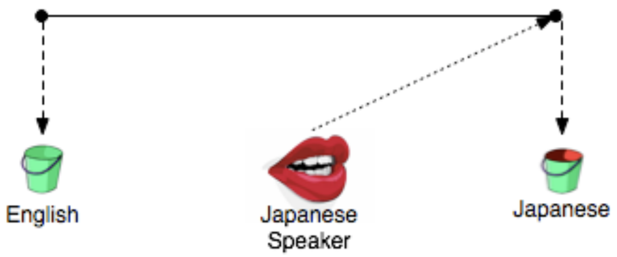


Figure 1. Two discrete language buckets with clear categorization for a native speaker.

Previous language identification studies compared languages that look and sound very different (Table 1).

Table 1

Example of difference between two languages.

	Orthography	Transcription
English	Please call Stella	pʰl.i:zkʰɑ:lstɛlʌ
Japanese	ステラに電話してくださ。	Sʊtɛɾʌnɪdɛɴwʌɕɪtɛkʊdʌsʌɪ[1]

Figure 2 displays an example of language identification for Japanese accented English. What would one of these language identification systems do if it were evaluating the speech of this variety of English? Would the system place the speaker in the English bucket or the Japanese bucket?

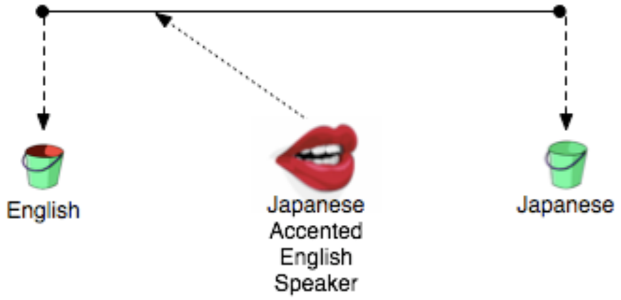


Figure 2. Japanese accented English with two discrete buckets. Not clear which category.

A further problem with language identification is the issue of linguistic variation. Language identification papers often do not mention whether accent or dialect is controlled for. It is

[1] Kansai dialect

usually not clear whether all speakers are of a uniform variety of a language, or are from different dialectal regions. In a transcription-based language identification task, would the phonology of the varieties of English spoken in Hindi-predominant regions and the English spoken in North America cause a splintering of the English category, or correct classification as in Figure 3. Similarly, in a language identification system based on speech signal (such as spectral coefficients), would the spectral coefficients of these varieties of English be correctly grouped together, or might one be divergent enough to be classified as another language entirely? Compared to the latter two pictures, Figure 1 is a simple problem to solve.

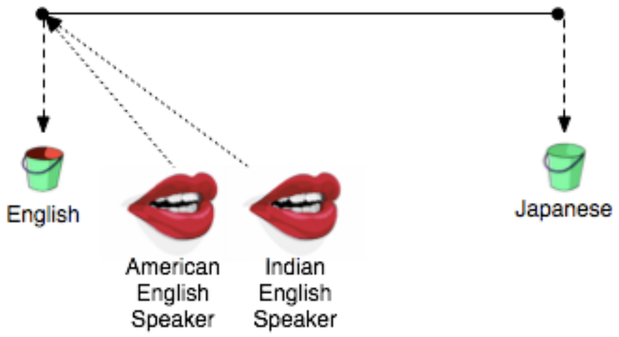


Figure 3. Two English native dialects that belong to the same language bin, but sound different to an English speaker.

The speaker variation shown in the second two situations is where our interest lies. While language identification is comparing two very different languages (as either spectral coefficients of speech or orthographic representations), accent and dialect identification are comparing more subtle differences.

Language, accent, and dialect identification

Various measures have been used to identify the language of a text or audio sample in the computational linguistics and speech recognition communities. Some of these algorithms may be useful for accent identification.

A popular method to measure the distance between words is to use Levenshtein distance. Heeringa et al. (2006) measures dialect by using Levenshtein distance calculations off of phones, with the goal of estimating geographic distance between spoken dialects. They reported reasonable results with geographically close dialects.

Kondrak (2005) introduces a string similarity algorithm based upon the calculation of the longest common span of characters between strings. Though no experiments were performed on dialects or accent, experiments in two cognate recognition tasks, and one pharmaceutical name confusion measurement task, showed that the more complex approach used in Kondrak (2005) could perform better to measure word similary and distance than Levenshtein.

Nagy, et al. (2005) used the Complete Link Algorithm to cluster phonetic feature vectors, resulting in a hierarchical view of dialect relationships. The study relies on an English dialect database of 179 phonetic features that show whether or not each feature could appear in a particular English dialect. These features include vowel features, vowel distribution features, consonant features, and prosodic features. They use ten dialects of English (e.g. Philadelphia, Cajun English, Tok Pisin, Australian), and limit themselves to a handful of words (e.g. kit, dress, goat, fleece, etc.). They show that vowel features alone clustered dialects better than consonant features alone, or to vowels and consonants combined.

Classification

Two classification approaches will be compared in this paper: a Naïve Bayes Classifier and a Bayes Belief Network Classifier. A Naïve Bayes Classifier is a simple fixed-structure network of nodes, made up of two levels, a parent level with a node for each class, and the child level, which contains the features. (Zhang, 2004; Manning, Raghavan and Schuetze, 2008). Each feature is treated independently, and the ordering of the features does not matter. The presence of one consonant substring does not require the presence of another consonant substring, but the presence of each feature adds to the likelihood that the feature belongs to some class. (Larose, 2006). Hence, the presence of both ŋks and tɹ add to the likelihood that the L1 language is Spanish while strings such as ʧ and tr add to the likelihood that the L1 language is Arabic.

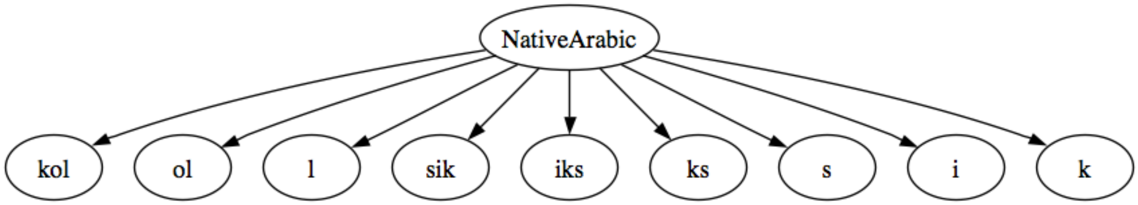


Figure 4. A graphical sample of a Naïve Bayes network.

This structure is called “naïve” as there is no graph structure to learn, there is only the setting of the weights. A graphical representation of what a Naïve Bayes network might look like on a set of strings can be found in Figure 4. The classification formula for Naïve Bayes follows in Formula 1.

$$c_{map} = \arg\max_{c \in C} \left(logP\prime(c) \sum_{1 \leq k \leq n_d} logP\prime(t_k|c) \right)$$

Formula 1. Naïve Bayes Classifier (Manning, Raghavan and Schuetze, 2008)

A Bayesian Belief Network classifier (e.g. Bouckaert, 2008) uses a set of features to create a probabilistic directed acyclic graph. As it is a directed graph, its features are dependent. We use the K2 algorithm with Bayesian metric scoring in order to learn the structure of the graph. K2 is very sensitive to the ordering of features (Cooper and Herskovits, 1992). The internal representation that a Bayesian Belief Network might generate could be demonstrated with a graph such as in Figure 5.

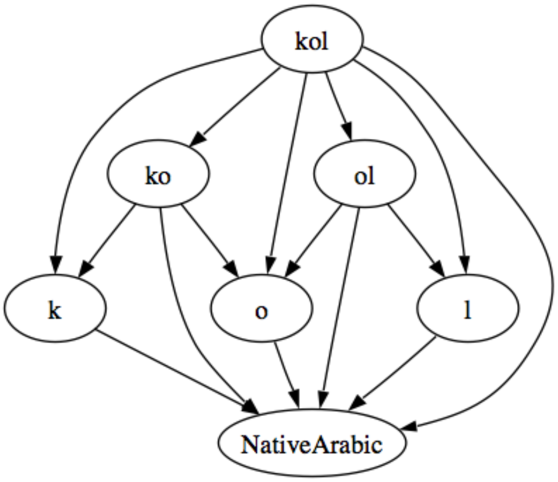


Figure 5. Graphical Representation of Bayesian Belief Network for one trigram in Trigram With Vowels

The classification formula for a Bayesian Belief Network Classifier follows in Formula 2.

$$c_{map} = \arg\max_{c \in C} \left(\prod_{c \in C} p(c|parents(c)) \right)$$

Formula 2. Bayesian Belief Network Classifier (Bouckaert, 2008)

Because features are dependent in the Bayesian Belief Network, the presence of one consonant substring cooccurring with the presence of another consonant substring increases the likelihood that those two together could help identify the class. (Larose, 2006) Thus, the presence of both *ŋks* and *tɹ* occurring by the same speaker add to the likelihood that the L1 language is Spanish, while *ʧ* and *tr* occurring by the same speaker add to the likelihood that the L1 language is Arabic.

METHOD

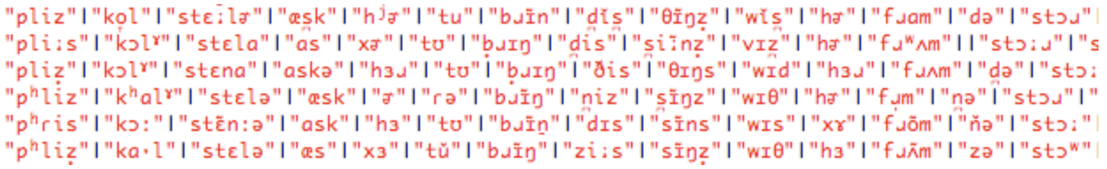
A custom corpus was created using the Speech Accent Archive (Weinberger, 2005) as a foundation. The Speech Accent Archive (SAA) is a database of 1065 speakers from around 200 L1 languages speaking the same paragraph of English.

> "Please call Stella. Ask her to bring these things with her from the store: Six spoons of fresh snow peas, five thick slabs of blue cheese, and maybe a snack for her brother Bob. We also need a small plastic snake and a big toy frog for the kids. She can scoop these things into three red bags, and we will go meet her Wednesday at the train station."

The SAA provides the original speech audio and the corresponding phonetic transcription of each speaker (using the International Phonetic Alphabet).

The first nine native speakers listed in the SAA from English, Spanish, Arabic, Portuguese, and Russian, were selected for the current study, resulting in a total of 45 speakers.

In order to compare phonetic transcriptions across speakers, the words of each speaker were aligned with every other speaker. In the SAA, sometimes a reader skipped a word, or read a word twice. Sometimes a transcriber did not segment the transcription properly. In order to align each word across the speakers, words boundaries were adjusted. If a reader skipped a word, a blank word position was added for that speaker. If the SAA did not separate a word from the next word, or if the speaker missed a word, a separation was added to the transcription (see Figure 6).



```
"pliz"|"ko̞l"|"stɛːlɚ"|"æs̪k"|"hʲɚ"|"tu"|"bɹĩn"|"d̪ĭs̪"|"θĩŋz̥"|"wĭs̪"|"hɚ"|"fɹɑm"|"də"|"stɔɹ"|
"pliːs"|"k̆ɔlˠ"|"stɛlɑ"|"ɑ̆s"|"xɚ"|"tʊ"|"b̥ɹɪŋ"|"d̪ĭs"|"s̪iĭnz̥"|"vɪz̪"|"hɚ"|"fɹʷʌm"||"stɔːɹ"|"s
"pliz̥"|"kɔlˠ"|"stɛnɑ"|"ɑskə"|"hɜɹ"|"tʊ"|"b̥ɹɪŋ"|"ð̆is"|"θɪŋ̊s"|"wɪ̆d"|"hɜɹ"|"fɹʌm"|"d̪ə"|"stɔː
"pʰlĭz"|"kʰɑlˠ"|"stɛlə"|"æsk"|"ɚ"|"ɾə"|"bɹ̆ĩŋ"|"n̪iz"|"s̪ĩŋz"|"wɪθ"|"hɚ"|"fɹ̥m"|"n̪ə"|"stɔɹ"|"
"pʰris"|"kɔː"|"stɛ̃nːə"|"ɑsk"|"hɜ"|"tʊ"|"bɹĩn̠"|"dɪs"|"sĩns"|"wɪs"|"xɤ"|"fɹõm"|"ň̆ə"|"stɔː"|
"pʰliz̥"|"kɑ·l"|"stɛlə"|"æs"|"xɜ"|"tŭ"|"bɹĩŋ"|"ziːs"|"sĩŋz̥"|"wɪθ"|"hɜ"|"fɹʌ̃m"|"zə"|"stɔʷ"|
```

Figure 6. Example Corpus Subset after alignment

Both speaker errors and transcription errors were corrected by the author. Speaker errors are reading errors such as skipping words, repeating words, and stumbling over words. Transcription errors include such things as improperly segmenting speech and overt transcription errors. Certain languages have more alignment/transcription errors than others. The same set of transcribers may have contributed to these errors, but information associating the transcribers to the transcriptions is not publically available.

Thirteen variations of the corpus were created from the foundation corpus controlling for the number of transcription symbols (or n-gram; where n can be 5-, 4-, 3-, 2-, or 1-symbols long), presence of vowels, and position of vowels. All conditions consider the word segmentation character as a component of the n-gram, and several include the diacritics, which are non-phonemic, secondary phonetic characters[2]. Table 2 gives examples of the thirteen corpus variations. Vowels may be in their original position in the string (unmodified), converted into a generic vowel placeholder ("substituted with a hyphen"), or be a chopping point where the string is broken up and only the consonantal clusters remain (split). When a vowel is left unmodified, or substituted with a hyphen, an important side effect is that the position of the vowel in the string can be used to help classify the string (respected). When the string is split into smaller pieces, the position of the vowel is lost (lost).

Trigrams were used because three phonemes is a minimum representation of a syllable onset, nucleus, and coda when diacritics are not present. 5-grams were selected because it is the average length of an English word. The 5-gram substring could also capture a full syllable with two vowel diacritics or it could capture a word-final coda followed by word segmentation character and a word-initial onset. One example of where this could be useful is with vowel epenthesis (anaptyxis) in syllable onset position of words in Spanish L1 speakers (e.g. initial vowel in /estela/, /ɛspunθ/, /əsneɪk/).

Trigram With Vowels and **5-gram With Vowels** use unmodified substrings of the word segmented SAA corpus. Both were controlled for substring distance only. **Trigrams Vowel**

[2] For example, in the transcription kʰɑlˠ, two diacritics appear: ʰ, and ˠ

Placeholder and **5-gram Vowel Placeholder** substituted every vowel with a place-holding character (a hyphen).

The vowel placeholder variations were created on the assumption that consonantal phonotactics alone could determine native language accent. We know empirically that vowels are more difficult to transcribe than consonants. Pollock and Berni (2001) explain the problem:

> "The transcription of vowels is more difficult than the transcription of consonants for several reasons. For example, vowels are both less discrete than consonants and more variable across dialects. In addition, there is not widespread agreement on the best categorization and transcription system for vowels...."

Shriberg et al. (1997) verified this by comparing the transcriptions of 33 child talkers created by two experts and one amateur. It showed that the agreement for broad transcription, or the phonemic transcription of the speech, was 92.7% for consonants and 86.5% for vowels. For narrow transcription, the transcription of speech with both phonemic and non-phonemic details, agreement was 80.6% for consonants and 77.8% for vowels. When processing transcribed speech, reducing vowels in the corpus reduces the error and ambiguity.

Trigram Maximized Consonantal and **5-gram Maximized Consonantal** took substrings of consonants up to 3 or 5 phones where the n-gram was not interrupted by a vowel. For the Maximized Trigram Consonantal, 1-, 2-, and 3-grams were used. For the Maximized 5-gram Consonantal 1-, 2-, 3-, 4-, and 5-grams were used. Conditions **2-gram Maximized Vowel**, **3-gram Maximized Vowel**, and **5-gram Maximized Vowel**, were added to compare against the consonantal results.

Four conditions for 1-grams were also used to compare the prediction quality of consonants to vowels: **Single Vowel No Diacritic**, **Single Vowel or Diacritic**, **Single Consonant or Diacritic**, **Single Consonant No Diacritic**.

Tables 2 and 3 explain what the conditions mean in light of these characteristics.

Table 2.

Examples from one speaker saying "Please Call Stella" as #pʰliz#kʰalˠ#stɛlə#

Corpus variations	Example
Trigram With Vowels w/Diacritic	#pʰl, pʰli, ʰli, iz#, z#k, #kʰ, kʰɑ, ʰɑl, alˠ, lˠ#, ˠ#s, #st, stɛ, tɛl, ɛlə, lə#
Trigram Vowel Placeholder w/Diacritic	pʰl-, ʰl-, -z#, z#k, #kʰ, kʰ-, ʰ-l, -lˠ, lˠ#, ˠ#s, #st, st-, t-l, -l-, l-#
Trigram Maximized Consonantal w/Diacritic	p, pʰ, pʰl, z#, z, z#k, #k, k, zkʰ, l, lˠ lˠs, ˠs, st, l
5-gram With Vowels w/Diacritic	#pʰli, pʰliz, ʰliz#, liz#k, iz#kʰ, z#kʰɑ, #kʰɑl, kʰɑlˠ, ʰɑlˠ#, ɑlˠ#s, lˠ#st, ˠ#stɛ, #stɛl, stɛlə, tɛlə#
5-gram Vowel Placeholder w/Diacritic	#pʰl-, pʰl-z, ʰl-z#, l-z#k, -z#kʰ, z#kʰ-, #kʰ-l, kʰ-lˠ, ʰ-lˠ#, -lˠ#s, lˠ#st, ˠ#st-, #st-l, st-l-, t-l-#
5-gram Maximized Consonantal w/Diacritic	p, pʰ, pʰl, z#, z, z#k, #k, k, zkʰ, l, lˠ lˠs, ˠs, st, l
5-gram Maximized Vowel w/Diacritic	ʰ, i, ɑ, ˠ, ɛ, ə
3-gram Maximized Vowel w/Diacritic	ʰ, i, ɑ, ˠ, ɛ, ə
2-gram Maximized Vowel w/Diacritic	ʰ, i, ɑ, ˠ, ɛ, ə
Single Vowel or Diacritic	ʰ, i, ɑ, ˠ, ɛ, ə
Single Vowel No Diacritic	i, ɑ, ɛ, ə
Single Consonant or Diacritic	p, ʰ, l, ˠ, z, k, s, t
Single Consonant No Diacritic	p, l, z, k, s, t

Table 3. Thirteen variations of corpora used in the current study.

Corpus variations	Length of strings	Vowels	Position of vowels
Trigram With Vowels w/Diacritic	3	unmodified	respected
Trigram Vowel Placeholder w/Diacritic	3	substituted with a hyphen	respected
Trigram Maximized Consonantal w/Diacritic	3, 2, and 1	split	lost
5-gram With Vowels w/Diacritic	5	unmodified	respected
5-gram Vowel Placeholder w/Diacritic	5	substituted with a hyphen	respected
5-gram Maximized Consonantal w/Diacritic	5, 4, 3, 2, and 1	split	lost
5-gram Maximized Vowel w/Diacritic	5, 4, 3, 2, and 1	unmodified	respected
3-gram Maximized Vowel w/Diacritic	3, 2, and 1	unmodified	respected
2-gram Maximized Vowel w/Diacritic	2 and 1	unmodified	respected
Single Vowel or Diacritic	1	unmodified	respected
Single Vowel No Diacritic	1	unmodified	respected
Single Consonant or Diacritic	1	split	lost
Single Consonant No Diacritic	1	split	lost

These corpora were then converted into a feature matrix for each speaker by assigning a Boolean value to the existence of each substring within the 45 speaker's readings of the stimulus paragraph. Each of these thirteen variations was trained on their own Bayesian Belief Network and Naïve Bayes classifiers with 10-fold cross validation because cross-validation has been shown to be a preferable method of testing small data sets (Goutte, 1997).

RESULTS

Table 4 summarizes the results of the thirteen classifiers. "Percent correct" is defined as the number of correctly classified speakers divided by the number of speakers (45 in this study). When broken down into precision and recall, precision is the percentage of results in that category that were correctly classified as that category, and recall is the percentage of results that were classified into that category, out of all the results that should have been in that category. F-Score is defined by Formula 3.

$$\frac{2(precision * recall)}{precision + recall}$$

Formula 3. F-Score

Out of the thirteen variations, the 3- and 5-gram "Maximized Consonantal" versions outperformed the others by a wide margin.

Table 4

Outcomes from Thirteen Corpus Variations.

Experiment	**BBN % Correct**	**BBN F-Score**	**NB % Correct**	**NB F-Score**
Trigram Vowel Placeholder w/Diacritic	57.78	.54	57.78	.55
Trigram With Vowels w/Diacritic	66.67	.66	66.67	.66
Trigram Maximized Consonantal w/Diacritic	77.78	.77	77.78	.77
5-gram Vowel Placeholder w/Diacritic	51.11	.51	44.44	.45
5-gram With Vowels w/Diacritic	55.56	.54	35.56	.35
5-gram Maximized Consonantal w/Diacritic	82.22	.81	80.00	.78
5-gram Maximized Vowel w/Diacritic	57.78	.56	55.56	.54
Trigram Maximized Vowel w/Diacritic	62.22	.59	53.33	.52
2-gram Maximized Vowel w/Diacritic	64.44	.64	60.00	.58
Single Vowel or Diacritic	53.33	.53	53.33	.53
Single Vowel No Diacritic	20.00	.18	20.00	.18
Single Consonant or Diacritic	55.56	.55	57.78	.58
Single Consonant No Diacritic	44.44	.42	44.44	.41

BBN=Bayes Belief Network, NB=Naïve Bayes

The best performing corpus variation was the 5-gram Maximized Consonantal using Bayes Belief Network, classifying 82.22% of the speakers into their correct L1 with an F-Score of .81. Table 5 shows the confusion matrix for the BBN 5-gram Maximized Consonantal classification and Table 6 shows the performance for each language.

Table 5

Bayes Belief Network Confusion Matrix

	Classified As				
Native Language	English	Spanish	Arabic	Russian	Portuguese
English	9	0	0	0	0
Spanish	0	8	1	0	0
Arabic	0	0	9	0	0
Russian	0	0	2	7	0
Portuguese	1	2	1	1	4

Table 6

Performance on Bayes Network, Maximized Consonantal 5-gram.

Native Language	**Precision**	**Recall**	**F-Measure**
English	0.90	1.0	0.95
Spanish	0.80	0.89	0.84
Arabic	0.69	1.0	0.82
Russian	0.88	0.78	0.82
Portuguese	1	0.44	0.44

The Naïve Bayes network reached 80% correctly classified with an F-Score of .78 on the 5-gram Maximized Consonantal condition. Table 7 shows the confusion matrix for this condition and Table 8 shows the performance for each language.

Table 7

Naïve Bayes Confusion Matrix

	Classified As				
Native Language	English	Spanish	Arabic	Russian	Portuguese
English	9	0	0	0	0
Spanish	0	8	1	0	0
Arabic	0	0	9	0	0
Russian	0	0	2	7	0
Portuguese	1	3	1	1	3

Table 8

Performance on Naïve Bayes, Maximized Consonantal 5 gram.

Native Language	Precision	Recall	F-Measure
English	0.90	1.0	0.95
Spanish	0.73	0.89	0.80
Arabic	0.69	1.0	0.82
Russian	0.88	0.78	0.82
Portuguese	1.0	0.33	0.50

Portuguese was the most improperly categorized L1 accent, with 6 out of 9 speakers misclassified across all the L1 accents on both the Naïve Bayes and the Bayes Belief Network classifiers. We see on both that while Portuguese had perfect precision, it had disappointing recall. On the other hand, Arabic had perfect recall, but had poor precision. While the 5-gram Maximized Consonantal condition performed very similar across the two algorithms, some conditions differ greatly. It was found that the NB performed better for the Single Consonant or Diacritic condition, but the BBN classifier met or surpassed the BN classifier for all of the other conditions.

Corpus noise

The system described in this paper utilized 45 speakers that are speaking 45 different variations of the same target language. These data have been transcribed by multiple transcribers with their own individual L1 biases, as well as academic linguistic backgrounds and theories. Ikeno (2007) shows L1 biases to be statistically significant when perceiving accents. The Speech Accent Archive speakers read with varied speeds and are not controlled for prosody. Prosodic information can affect a listener's bias to accentedness just as much as segmental information (Munro,1995; Munro and Derwing, 1998). With natural speech, human biases, the difficulty of phonetic transcription, and the difficulty of collecting data on such a large scale with open submission, we have some very noisy data.

Multidimensional Scaling Analysis

In order to visualize the variation between speakers, we condense over 750 features of the Maximized Consonantal 5-gram down to two dimensions using multidimensional scaling. Multidimensional scaling uses distance to respect the similarity and dissimilarity between two vectors (Borg and Groenen, 2005). Figure 3 displays the results of multidimensional scaling analysis of the 5-gram Maximized Consonantal condition.

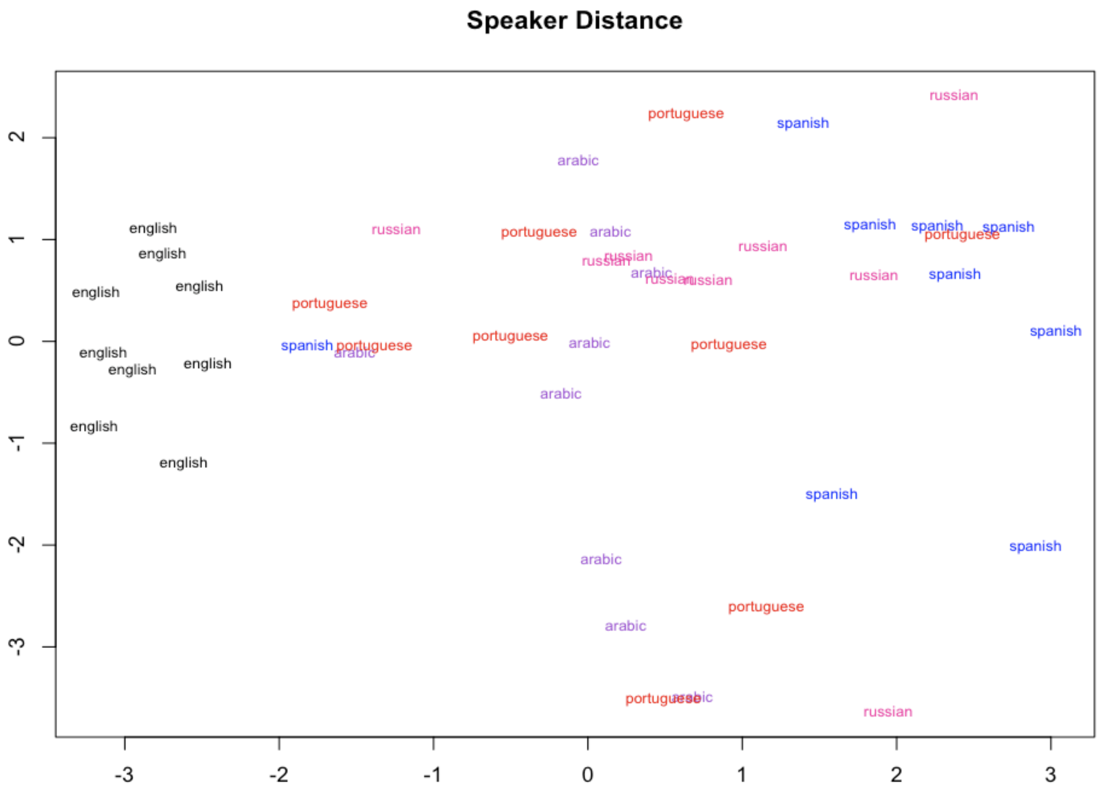


Figure 7. Multidimensional Scaling Analysis of 5-gram Maximized Consonantal

By flattening out our feature vectors, we find there are overlapping, but noticeable clusters of L1 speakers (Figure 7). English native speakers form a dense cluster with no overlap along the left, Spanish is diffuse along the top right except for one speaker, Arabic forms a vertical cluster going straight down the center, and Russian overlaps Spanish and Arabic in the central top right quadrant, except one speaker appears in the bottom right. With the exception of two outlying speakers, Portuguese is diffuse across the top half of the figure.

If we consider the English cluster of speakers to be "prototypical English," the distance between the English cluster's central point and an individual speaker can be used as the basis for a measurement of that speaker's phonological accent on English speech. The speakers of each native language appear to cluster together. This may be due to the biases of native-English speaking transcribers, or it may be a legitimate phonological distance. If this distance appears in a speech recognition system for use with accent scoring, it would have to be normalized in some way to account for the fact that some L1s have more phonological divergence than others.

DISCUSSION

Even with all that variation of 45 speakers speaking English, the most robust language category was native speakers of English.

We found consonant clusters a better indication of accent. On the other hand, Nagy et al. (2005) found that vowel features, rather than consonants, were a better indication of dialect in their word list. Though this is not a direct comparison, the difference requires some attention because we believe accent analysis and dialect analysis to be similar problems. Nagy's data on vowel features appears to be far more discrete than IPA transcription. Each of the features is a direct yes or no choice about whether a particular sound can appear in a particular dialect with a prototypical example for comparison, and a discrete classification. Each choice is made deliberately by analogy, and may require more attention to encode. For example "possible variants of the vowel of the word /KIT/ are identified as (1) "canonical" high front [I]; (2) raised and fronted variant phonetically identified by the symbol [i], (3) centralized [ə], and (4) with an offglide, e.g. [Iə/iə]." In the corpus used for the current study, transcribers type a symbol that represents what they think they hear, in an open-ended classification using a continuous 3-dimensional vowel space. Research (e.g. Shriberg et al. 1997 ; Li et al. 2000) shows that this kind of identification is difficult and that even trained transcribers sometimes show inconsistency.

CONCLUSION

This paper introduced a new computational method to identify L1 and degree of accentedness of speakers from phonetically transcribed corpora. The results from 45 speakers showed that the system can correctly identify native English speakers from a set of all English speakers with 90% precision (Table 8), and could classify speakers into native speaker group with 82.2% accuracy by using 5-gram Maximized Consonantal with Diacritic condition and classifying by using Bayesian belief networks (Table 4). We show that if we do not wish to assume dependence between strings, we can adopt a naïve Bayes classifier that works nearly as well. The Naïve Bayes and Bayesian Belief Networks perform similarly across the 13 variations the corpus, showing that the results are fairly stable.

These results suggest that narrow transcription has enough information to correctly classify accent. It also suggests that much of the accent information could be found in the consonantal phonotactics, and that transcribed vowels add little value to classification. This last point lies in contrast to the results Nagy et al. (2005) has shown, and is likely because vowel transcription is much more difficult than vowel analogy.

Furthermore, the vowel classification results here indicate that transcribed vowels do not aid in classification of speakers to their linguistic affiliation, yet transcribed consonants can. This could be tracked to transcriber biases, or that consonantal differences are more stable and less variational. Regardless of the answers to those two empirical questions, the underlying question these results bring up is if the 3-dimensional vowel system is really capturing and communicating the linguistic differences that linguists rely on it to do.

REFERENCES

Borg, I. and Groenen, P.J.F. (2005). *Modern multidimensional scaling*. 2nd edition. New York: Springer.

Bouckaert, R. (2008). Bayesian Network Classifiers in Weka for Version 3-5-7. Retrieved from http://www.cs.waikato.ac.nz/~remco/weka.bn.pdf on May 15, 2009.

Cooper, G. R., & Herskovits, E. (1992). A Bayesian method for the induction of probabilistic networks from data. *Machine Learning, 9*, 309-347.

Goutte, C. (1997). Note on free lunches and cross-validation. *Neural Computation, 9*, 1211-1215.

Heeringa, W., Kleiweg, P., Gooskens, C., & Nerbonne, J. (2006). Evaluation of string distance algorithms for dialectology. In Proceeding in the Linguistic Distances Workshop at the joint conference of International Committee on Computational Linguistics and the Association for Computational Linguistics (pp. 51-62).

Ikeno, A., & Hansen, J. H. L. (2007). The effect of listener accent background on accent perception and comprehension. *EURASIP Journal on Audio, Speech, and Music Processing, 2007,* Article ID 76030.

Kondrak, G. (2005). N-Gram similarity and distance. Proceedings of the Twelfth International Conference on String Processing and Information Retrieval. 115-126.

Larose, D. T. (2006). *Data mining methods and models.* Hoboken, NJ: John Wiley & Sons.

Li, A., Zheng F., Byrne, W. J. , Fung, P., Kamm, T., Liu, Y., Song, Z., Ruhi, U., Venkataramani, V., & Chen X. X. (2000). CASS: A phonetically transcribed corpus of Mandarin spontaneous speech. *ICSLP-2000*, *1*, 485-488.

Manning, C. D., Raghavan, P., & Schuetze, H. (2008). *Introduction to information retrieval.* New York, NY: Cambridge University Press.

Munro, M. J. (1995). Nonsegmental factors in foreign accent. *Studies in Second Language Acquisition*, *17*, 17-34.

Munro, M. J., & Derwing, T. M. (1998). The effects of speaking rate on listener evaluations of native and foreign-accented speech. *Language Learning, 48*, 159-182.

Nagy, N., Zhang, X., Nagy, G., & Schneider, E. (2005). A quantitative categorization of phonemic dialect features in context. In A. Dey, et al. (eds.). *CONTEXT 2005 Lecture Notes in Artificial Intelligence 3554*. New York: Springer-Verlag. pp. 326-338.

Pollock, K. E., & Berni, M. C. (2001). Transcription of vowels. *Topics in Language Disorders, 2,* 22-40.

Shriberg, L. D., Austin, D., Lewis, B. A., McSweeny, J. L., & Wilson, D. L. (1997). The Percentage of Consonants Correct (PCC) Metric: Extensions and reliability data. *Journal of Speech, Language, and Hearing Research, 40*, 708-722.

Weinberger, S. H. (2005). *The Speech Accent Archive.* George Mason University, Fairfax, Virginia. Available from http://accent.gmu.edu/

Zhang, H. (2004). The optimality of naïve Bayes. In V. Barr, & Markov, Z. (Eds.) *Proceedings of the 17th International FLAIRS Conference*. AAAI Press.