Speech recognition software for language learning: Toward an evaluation of validity and student perceptions

Deborah Cordier

University of South Florida
Speech Recognition Software for Language Learning: Toward an Evaluation of Validity and Student Perceptions

by

Deborah Cordier

A dissertation submitted in partial fulfillment of the requirements for the degree of Doctor of Philosophy
Department of World Languages
College of Arts and Sciences
and
Department of Secondary Education
College of Education
University of South Florida

Co-Major Professor: Linda Evans, Ph.D.
Co-Major Professor: James White, Ph.D.
Jacob Caflisch III, Ph.D.
Roger Brindley, Ph.D.

Date of Approval:
July 14, 2009

Keywords: CALL, foreign language software, French pronunciation and speaking, foreign language raters, mixed method

© Copyright 2009, Deborah Cordier
Dedication

To my family, Nicole Cordier, my daughter and inspiration, my parents Mr. Jean J. and Marjorie L. Cordier, my three brothers, David, Derek and Douglas, grandparents Harry and Gladys Cordier (deceased), my uncle Robert Cordier, M.D. (deceased), and adopted grandparents Drs. Anna Zeldin and Petr Stoler, thank you for your years of continued support and encouragement. Without the deep awareness of your love, I would not have been able to continue on this long road to a happy destiny.
Acknowledgements

Thanks to Alain De Connick and Tom Hunsaker, Educational Consultants, Auralog (2004) for accepting my suggestions, and to Tim O'Hagan and Peter Lucas for helping to make the suggestions reality; Dr. Tucker, French grammarian par excellence; my colleague, fellow academic and friend Rob Cooksey, who has continuously inspired me to reach and search, always beyond my limitations; to Rob Summers and the AC staff, Rob Oates, Zac Carter and Krystal Wolfe, who guided my vision of the project and technology skills throughout. To my committee members, Roger Brindley, who early on and at critical periods, inspired me to think creatively, and Dr. Caflisch, who allowed me to follow behind him for several years, listening and absorbing every detail of conversation and exchange (I still have only scratched the surface structure!). Finally, to my co-major professors, Dr. Evans and Dr. White, who believed in my scholarship, research skills and who both helped me beyond measure, to successfully navigate a series of processes through to the end. With sincere gratitude, thanks to all of you.
Table of Contents

List of Tables ...................................................................................................................... iv
List of Figures ...................................................................................................................... v
Abstract .............................................................................................................................. vi

Chapter One:  Introduction ..................................................................................................1
  Background .......................................................................................................................1
  CALL Evaluation and FL Tasks ......................................................................................2
  Statement of the Problem .............................................................................................4
  Research Questions .........................................................................................................6
  Significance of the Study ...............................................................................................7
  Definitions of Terms .......................................................................................................9

Chapter Two:  Literature Review ....................................................................................... 11
  Introduction .....................................................................................................................11
  Foreign Language: Learning, Teaching and Speaking ................................................12
    Automatic Speech Recognition Technology ...............................................................16
    Human Raters for FL Evaluation ................................................................................24
    CALL Research Using Speech Technologies and ASR ............................................32
    CALL Evaluation .........................................................................................................37
    Foreign Language Research and TeLL me More (TMM) ..............................................41

Chapter Three:  Method .....................................................................................................46
  Introduction .....................................................................................................................46
  Research Questions .........................................................................................................47
  Design ...............................................................................................................................47
  Participants .......................................................................................................................48
  Procedure .........................................................................................................................48
    Sentence Pronunciation and ASR Scores .................................................................49
    Digital Data ..................................................................................................................51
    Human Raters ...............................................................................................................53
    Rater Site ......................................................................................................................54
    Rater Training ...............................................................................................................55
    Participant Survey .......................................................................................................57
  Data Collection and Analysis .......................................................................................58
    Statistical Measures - ASR Scores and Rater Data ................................................58
    Inter-Rater Reliability .................................................................................................58
List of Tables

Table 1  Descriptive Statistics for ASR, ra, rb and rc Scores.............................................70
Table 2  Correlation Coefficients: ASR and Raters a, b and c Scores.....................................76
Table 3  Intra-Rater Reliability Correlation Coefficients:
  Time = 1 and Time = 2 for Rater a, b and c ..........................................................81
Table 4  Frequencies for Gender and Duration of French Study ........................................84
Table 5  Frequency Responses for Survey Q 3 and Q 4 .........................................................85
Table 6  Question 7 (a, b, c) Frequency Responses .............................................................88
Table 7  Frequency and Percentages for Survey Question 8(a) and (b) .................................94
Table 8a  Survey Q 9 – ASR Compared to Other Feedback ...............................................99
Table 8b “Perceived Helpfulness” According to Length of French Study .........................101
Table 9  Survey Question 10 - Coded Responses for French Teacher Rating ......................105
Table 10  Rater ASR Scores for Sentences s1-s10...............................................................134
Table 11  Magnitude: ASR and Rater Correlations............................................................143
List of Figures

Figure 1. P25 Sentence 1-10 ASR Scores ........................................................................63
Figure 2. Three Rater Scores for P16 Sentence 1 (s1), s2 and s3 .................................64
Figure 3. Organization of ASR and Rater Scores for Analysis .................................65
Figure 4. Collated Participant Responses for Survey Question Six ......................67
Figure 5. ASR Software “Score” Box ..........................................................................95
Figure 6. Participant Screen View of French Sentence Groupings .......................123
Figure 7. The Rater’s View of the Rating Screen ..................................................131
Figure 8. ASR Software “Score” Box ......................................................................162
ABSTRACT

A renewed focus on foreign language (FL) learning and speech for communication has resulted in computer-assisted language learning (CALL) software developed with Automatic Speech Recognition (ASR). ASR features for FL pronunciation (Lafford, 2004) are functional components of CALL designs used for FL teaching and learning. The ASR features available with the Tell Me More French software provides pronunciation, intonation and speaking practice and feedback. ASR features are examined quantitatively through French student performance of recorded ASR-scored speech and compared with human raters of the same produced speech samples. A comparison of ASR scores to human raters considers the validity of ASR-scored feedback for individualized and FL classroom instruction. Qualitative analyses of student performances and perceptions of ASR are evaluated using an online survey linked to individual pronunciations and performance and examined for positive impact (Chapelle, 2001) and usability.
Chapter One: Introduction

Background

Foreign language (FL) acquisition and teaching have experienced a renewed and vibrant focus, particularly in recent years with explosive growth toward a unified Europe. Globally, there has been attention focused on foreign languages in African nations, Asia and China at both regional and national levels. English acquired as an International Language (Jenkins, 2000) has highlighted the realities of a global economy in which numerous native languages are spoken by people of various nationalities. Jenkins (2000) has researched the pronunciation difficulties experienced by English language learners from diverse backgrounds and found that many were learning English as a third language, in an FL environment. In many countries, people are required to acquire or use English for on-the-job communicative purposes.

Theoretically an academic domain, foreign language learning and teaching has evolved to include the area of Second Language Acquisition (SLA) or the acquisition of a language other than the native language (L1). SLA is a recent evolution from within the disciplines of both theoretical and applied linguistics and has incorporated theoretical foundations from earlier research. For example, the groundbreaking works of the linguist Noam Chomsky (1957) theorized L1 acquisition in children and Roger Brown (1968) studied and documented how young children acquire a first language. Fundamental SLA incorporates both the idea of acquisition of an L1 and, depending on the age of onset,
type and amount of second language (L2), includes both the concepts of acquisition and learning. Indeed, SLA includes important theoretical foundations, methodologies and skills for the acquisition and the teaching of second languages (L2).

The advent of computer technologies and their subsequent applications in various educational disciplines has resulted in the use of computer-mediated communication (CMC) for foreign language acquisition and teaching. Moreover, the technologies used for FL acquisition have paralleled the technological developments within other domains. Foreign language technologies have also been developed, implemented and researched by FL specialists (Plass, 1998). Researchers have been eager to investigate claims that the use of technology can contribute to positive student outcomes.

Foreign language educators, specializing in applied linguistics, SLA and computer-mediated environments, have contributed to a body of work, including theory, processes, products, and research which comprise a sub-area of SLA called computer-assisted language learning (CALL). In the 1990s, the application of foreign language teaching principles informed early CALL which incorporated evolving technologies (Chun, 1994; Warschauer, 1996).

CALL Evaluation and FL Tasks

Today, CALL designs include state-of-the-art software, such as Automatic Speech Recognition (ASR) and internet technologies, and are routinely evaluated by FL professionals for their pedagogical usefulness. CALL evaluation must center on the FL tasks designed for software and computer-mediated environments. In her seminal work, Chapelle (2001) sets forth theoretical foundations and principles to consider for the
design and evaluation of CALL tasks and states that “CALL should be evaluated through two perspectives: judgmental analysis of software and planned tasks, and empirical analysis of learners’ performance” (p. 52). Ideally both types of analyses, judgmental and empirical, need to be conducted.

Chapelle (2001) suggests several evaluation criteria for CALL tasks and an important feature includes task appropriateness. For example, CALL designs that include appropriate FL tasks are evaluated for their language learning potential (focus on the form of the language), ability to engage learners and “positive impact” (positive effects of participation). The task appropriateness criteria can assist developers, researchers and other FL specialists to establish whether the computer-mediated tasks are appropriate for students’ FL learning.

Based on early foreign language teaching methods (Delattre, 1967; Valette, 1965) and technological developments in communications, CALL tasks that have been relatively easy to adapt and develop for the computer-mediated environments were 1) FL pronunciation and 2) speaking skill practice. Interestingly, the science of linguistics was founded on the analysis of phonetic symbols, or the sounds of spoken language and linguists studied the sounds of a language and their arrangement in words to communicate meaning. Phonology remains an important branch of linguistics that is the foundation for understanding speech in any language used for human communication.

Beginning in the 1950s, and used predominately for speech therapy, recording speech and pronunciation had been used for speech analysis and by speech labs to identify, record and correct individual sounds. Today, foreign language computer-assisted pronunciation training (CAPT) is a type of CALL software that includes basic FL
pronunciation training (O’Brien, 2006). Basic CAPT was originally modeled after what students experienced in language labs. In the lab setting, a student could listen to a native speaker model and then try to repeat and record produced speech, or output, based on the voiced features and pronunciation heard in the model.

More recent developments in CALL have included course ware with more sophisticated features. Automatic Speech Recognition (ASR), or the capability of the software to “capture, recognize, and react in some way to human speech is extremely popular” (O’Brien, 2006). Visualization techniques have also been used in course ware in combination with ASR. The most common techniques include the presentation of a pitch curve and a waveform, visual forms that represent voiced speech.

Statement of the Problem

The TeLL me More (TMM) foreign language software products that include automatic speech recognition (ASR), have been in use throughout Europe and globally for several years. For example, Lafford (2004) published a comprehensive review of TeLL me More Spanish and outlined the pedagogical strengths of the ASR software. Later, Burston (2005) recommended the TeLL me More German version and commented on the strengths of the speech and pronunciation features. Several recent conferences have included participant presentations (ACTFL, November, 2005) that reported positive impact (Chapelle, 2001) regarding use of the FL software in education.

The development of CALL software that includes automatic speech recognition used for FL pronunciation and speech practice has contributed to a renewed focus on the development of FL speaking skills. Through the provision of recorded native speaker
models, repetition, voice recording and feedback, ASR has significantly contributed to software designs. FL learners have been able to both learn and practice FL speech and pronunciation in settings where the software has been available for students’ use. Current CALL course ware includes ASR and the entire range of pedagogical skills and activities, several languages and online learning versions.

Foreign language study and interest have expanded within the European Union (EU) community, globally and through technological advances in telecommunications. The ACTFL 2005 Year of the Languages (American Council for Teaching Foreign Language) highlights the importance of concurrent developments in the United States. Important technological developments have significantly added to CALL, FL tasks and the computer-mediated tools made available to foreign language learners.

Following from the recent attention to and visibility of the ASR features of the software, several universities abroad and throughout the US are using TMM versions either as distance learning courses or in blended learning formats (Bonk, 2004) with foreign language classes. University language labs (Rollins College, FL) are providing TeLL me More as stand-alone FL course content or as additions to the in-class curriculum (Bryant University, RI).

Numerous software reviews have documented the use and efficacy of several language versions of the TMM software (Lafford, 2004). Articles submitted to international peer reviewed journals (Calico Journal; Language Learning and Technology) are subject to rigorous CALL evaluation (Hubbard, 1992), including review of both pedagogical and technological features. In several reviews, the speech recognition software used in TeLL me More has been rated a strong and contributory feature to FL
speaking and pronunciation skill practice, by both faculty and students (Barr, D., Leakey, J., & Ranchoux, A., 2005).

Several recent reviews (Lafford, 2004; Burston, 2005) have reported on the positive contributions of ASR to CALL software. Studies have focused on the quality and interpretation of ASR and visualization feedback provided to learners (Chun, 1998; Hincks, 2003). Chun (1998; 2002). Researchers (Pennington, 1999; Hardison, 2004) have suggested that ASR and CALL course ware is ideally suited for research data collection. However, and perhaps due to the relatively recent awareness of ASR among FL educators, there have been no reported research studies on the evaluation of ASR software for improving FL pronunciation and speech with students who have used ASR software.

ASR software is becoming an increasingly important feature of CALL designs and software (O’Brien, 2006; Chun, 2002) and there is a need within the FL community to continue to support and develop technology that contributes to FL teaching and positive student outcomes (Chapelle, 2001). Consequently, there is a definite need for an investigation of FL student learner outcomes and perceptions of ASR for foreign language learning and enhanced pronunciation.

Research Questions

The use of ASR software for FL learning has been reported in numerous studies. The potential efficacy of the ASR as an aid to FL speech and pronunciation has been documented and observed. However, how can the effectiveness of ASR for FL learning be evaluated and can a more conclusive measure or statement of improved FL speech be
determined? What is the contribution of ASR to CALL research and to student learning and performance? Can the results of student performance be compared to qualified, established performance measures? Importantly, how do students view ASR and what is their evaluation of its purpose and effectiveness for their learning? Does the use of ASR software, for pronunciation and speech practice empirically result in improved FL student pronunciation and speech? Does the feedback given actually help students to improve? Are the ASR scores provided to students a valid measure of their performance? How do students interpret the scores and how do they use the feedback for practicing their pronunciation? How does the feedback compare to “traditional” feedback given by raters and how do ASR scores compare to human ratings, of the same speech recordings? The following investigative questions are proposed:

1) When using the same ASR-produced speech for human rating, how do human ratings compare to the ASR scores provided to students and how valid are the ASR scores?
2) How useful did the students find the ASR features were for assisting them with their French pronunciation practice?
3) How effective was the ASR software for helping students improve their mastery of French?

Significance of the Study

The teaching, learning or acquisition of foreign language assumes a full range of activities for improving all aspects of language development, including a focus on the four basic skills of reading, writing, listening and speaking. The skill that is often of
interest to students learning a language is the ability “to speak the language”.

Pronunciation and conversation are important aspects of student language development. Without a comprehensible production (Derwing et al., 1997) of language and speech, learners lack the ability to communicate and to be understood.

The rapid technological developments throughout the world have contributed to the unprecedented expectation of learners for an increased level of sophistication and access to learning. The CALL software currently developed for the FL learning may be, in some cases, incapable of meeting the needs of our current generation of learners who have the desire to collaborate and communicate with other learners and their peers in a virtual world.

Automatic speech recognition (ASR), as a tool for FL speech and pronunciation practice has the potential to meet the needs of FL learners. By providing individualized FL speech practice and instant feedback, ASR can give students access to a form of learning in a simulated, immersive environment.

An investigation of ASR features and feedback scores and an examination of FL learner perceptions of ASR as tools for learning could all contribute to the significance of incorporating ASR into CALL designs. Further, the plausibility of using ASR-designed scores for recorded speech and feedback could be documented. A comparison of ASR-scored recordings and human ratings of the same ASR-scored recordings could provide further insights into the applicability ASR for FL speech and pronunciation as well as for programmatic and individualized learning outcomes.
Definitions of Terms

1) ASR: Automatic Speech Recognition
2) CALL: Computer-Assisted Language Learning
3) CAPT: Computer-Assisted Pronunciation Training
4) L1: First language, Mother Tongue (MT) and Native Language (NL)
5) L2: Second language also, Target Language (TL)
6) Native speaker (NS): Speaker of the home country language
7) Intonation: The pitch contour of a phrase or sentence. Pitch movement in spoken utterances that is not related to differences in word meaning.
8) Pitch: In speech, the rate at which the vocal cords vibrate (F-zero) is perceived by the listener as pitch. The auditory property of a sound placed on a scale from low to high.
9) Pronunciation: The utterance or articulation of a word or speech sound, including accurate pronunciation of “neighboring” sounds (allophones).
10) Rhythm: The patterned and recurring alternations of contrasting elements of sound.
11) Speech features: Features are the smallest unit of analysis of phonological structure.
12) Speech synthesis: Uses a computer and graphic representation for the process of generating human-like speech.
13) Stress: Sounds perceived as relatively more prominent as a result of the combined effects of pitch, loudness and the length.
14) Visualization techniques: Graphic representations of a pitch curve and wave-
form, as used in ASR where the computer decodes a representation of the spoken language.
Chapter Two: Literature Review

Introduction

Foreign language learning and teaching has historically been attached to ideas about language and how languages are acquired. Ideas have been put forth by teachers and researchers regarding the L1 or native language and more recently the field of second language acquisition (SLA) has opened doors for understanding L2’s and multilingual environments. Teaching foreign languages has evolved from both the science of linguistics, foreign languages and the development methods designed to facilitate the learning process for students. Methods continue to evolve and improve, often as a result of testing methods and findings about the student learning process.

Technology, in the form of digital and multimedia resources, has been a relatively recent addition to teaching methods and foreign language learning. Computer-assisted language learning (CALL) researchers investigate materials designed to enhance FL learning. Automatic speech recognition (ASR) is one feature that has been added to materials in recent years to assist FL learners with pronunciation and speaking practice. Studies designed to investigate the ASR feature of CALL products, used by students to assist with their FL learning are underrepresented. There is no record, to date, of studies specifically investigating the interpreted meaning of an ASR scoring feature provided by commercial ASR and presented to students using the software.
In the early 19th century, foreign language was an integral part of US immigrant community life as newly arrived families struggled to learn English and to assimilate into a new and vibrant culture. The view of the US as a “melting pot” certainly hinges on the idea of foreign languages and cultures becoming a part of the fabric of a distinctly American language and heritage. Foreign languages of immigrants, US participation in foreign wars and academic interest in understanding and using foreign language texts for literature, industry and commerce were only a few of the factors responsible for the learning of foreign languages within the context of an American English environment.

Teaching of English, and other foreign languages, became an integral part of American education out of a need to understand and communicate with students, their families, teachers and friends. The concept of the “old schoolhouse”, with a few students and a local teacher can still be found in small, remote communities within the rural and expansive western US. The Modern Language Association (MLA) was one of the first of groups of educators involved in developing and disseminating foreign language and translated materials for use within the schools and universities. MLA writing style and guidelines are still used for academic works within a foreign language, social humanities framework.

With the introduction of foreign languages into the communities and schools came a need for trained teachers, teaching methods and materials. Early foreign language teachers, lacking developed methods, were often native speakers of the language or bilinguals capable of managing the FL within the classroom. Still today, many students and foreign language educators feel that the native speaker (NS) model (Long, 1991) is
the model to be aspired to for learning or teaching. William James, an American
psychologist, was educated in France in order to become “fluent” in French (Proudfoot,
2004). He later translated his learning into a universal treatise on the “pragmatics” of
living (Pragmatism). Many of the early learning methods were developed from a need to
understand, communicate, or interpret language and meaning for a distinct purpose.

Linguistics and applied linguistics were the first disciplines, besides the direct
contact with native speakers, to inform FL teaching methods. The science of linguistics
has a long historical tradition with numerous schools of thought. De Saussure, (Course in
General Linguistics, 1959) is credited with writing the first Linguistics text still used and
referenced today. Ironically, the seminal work was compiled from the copious notes of
several of his students, as the maitre was known for “speaking” his lectures and not
writing them down! As was De Saussure, linguists were mainly concerned with the
spoken language and concentrated heavily on phonology and phonetics, or generally the
sounds of language. Applied linguists of the 1960s and early 1970s, predominantly from
the UK, were interested in the applicability of the linguistic frameworks to FL teaching.
One of the earliest methods, grammar-translation (GT) was concerned specifically with a
direct one to one relationship between, for example, a term or word in a first language
(L1) and the translated meaning in the second language (L2).

An interest in first language learning (Roger Brown, 1968) and a subsequent shift
to more psychological and introspective methods for examining the acquisition of
language resulted in a convergence of linguistics, applied linguistics and foreign
language, acquisition, learning and teaching. The sub-field of second language
acquisition (SLA) is a direct outcome of the convergence, whereby the acquisition of a
second language, other than the mother tongue or native language, has come to be viewed as distinctly different from the acquisition of the L1. FL learning and teaching, from both a theoretical and practical framework, were central to the direction, developments and research within SLA.

In the 1980s, the developments within SLA resulted in a renewed focus on foreign language for communication. The communicative language theories and teaching method (CLT) (Canale and Swain, 1980) were influenced by a need to use language for practical purposes and realistic needs. Students wanted to be able to use the language they learned and were often motivated to study and acquire a language for the distinct purpose of using the language with their families or for traveling abroad to visit relatives or friends. FL teachers wanted to be able to motivate, teach and inspire in their students a love for language, a passion that often motivated the teachers themselves.

Dr. Caflisch, Linguistics Professor Emeritus (USF, 2005), has often been praised for his French accent when using his French. He is quick to comment on how he acquired and mastered the accent in seventh grade when his French teacher, a non-native speaker asked for pronunciation assistance from a native French-speaking student in the class. Dr. Caflisch credits the teacher for having the insight to recognize the value of using a native-speaker model for pronunciation practice.

Within the classroom, FL teaching has developed and incorporated or discarded techniques, but the theme of communicative language use has remained a steady framework since the early 1980s. Early FL classroom teaching involved four skills: reading, writing, listening and speaking. Speaking, a main objective for many learners, was difficult for students within a classroom setting, as most students felt anxious
(Horowitz, 1980) and self-conscious about speaking in front of the teacher and fellow students. Teaching speaking, especially with beginners and lower FL level students, was usually in the form of memorized practice, recitation, pronunciation drills and conversation exercises with partners. When speaking was tested, indirect methods were used for discrete pronunciation features or for listening for correct pronunciation (Valette, 1967). Direct methods, whereby the teacher speaks with a student or students, were time-consuming and recording techniques were unsophisticated and costly. As a result of these and other factors, pronunciation and speaking were de-emphasized in the FL classroom.

Speaking as a form of interaction with authentic, real-life materials and native speakers, a main feature of CLT, was re-emphasized and re-vitalized within SLA by the technological developments within society, generally, and specifically through computer-assisted language learning (CALL). Chun (1994) was one of the first SLA researchers to examine interaction with her German students using computers to “speak” by writing to each other and thus communicatively interacting within the FL classroom. Throughout the 1990s, rapid technological advances in hardware, software, audio and video have contributed to CALL tools. The explosion of the internet as a worldwide web of global communication has inspired a desire for social communication and virtual connection. The resurgence of a sense of community has provided new and interesting tools for FL teachers to use. Some teachers have been eager to respond, others have been hesitant and less eager. Regardless, FL students have experienced a bigger world by the time they arrive at the university FL classroom. The FL student of today has the opportunity to explore and learn in unparalleled ways.
One area of computer-assisted language learning that has presented particular difficulties for researchers and teachers and, at the same time, significantly contributed to the learning materials for speech and pronunciation is the use of Automatic Speech Recognition (ASR). The most common and widely recognized example is the TeLL me More (TMM) software by Auralog that has embedded ASR software within a pedagogically-based language learning software. The benefits of the ASR software for students are many and have been documented by numerous researchers and teachers who have used the software (Hincks 2003; Hemard, 2004). FL students can practice pronunciation and speech samples, receive feedback, repeat, listen and hear a native model in an individualized setting.

The limitations of ASR are often misunderstood by the FL teachers who are unfamiliar with the details of how ASR works. There is no need to explain or understand a sophisticated technology, however, there is the need, on the part of the CALL researcher to understand the process in order to re-interpret the limitations for users.

Davies (2006) has provided an excellent reference for language teachers and using a module approach has succeeded in explaining very complex terms in clear and understandable language. According to Davies, the history of how ASR was developed for CALL, where ASR fits into CALL and the state-of-the-art are all traced to the initial efforts in the 1940s-1950s with machine translation (MT). Machine translation as a research area was abandoned, but has contributed to the concepts of artificial intelligence (AI) or the ability of the computer to produce responses equivalent to human, intelligent
response. AI has evolved from the early MT work on the mathematical reconstruction of language, using rules to reconstruct pieces of grammatical structure (Parsing).

Intelligent CALL or parser-based CALL (Davies, 2006), incorporating theoretical linguistic concepts and speech synthesis, has been at the origin of newer domains such as corpus-based linguistics. Speech synthesis, or speech analysis, although similar to ASR, is often confusing for linguists or FL teachers who had previously used speech synthesis for studying sounds of language (Molholt and Pressler, 1987; Laderfoged, 2004). Speech synthesis takes already produced or spoken language speech sounds and analyzes the speech textually, graphically or in other visual formats. A visible form of speech synthesis and transcription is used for producing speech on the TV screen, most commonly for the hearing impaired. Yet, English language beginning learners, who may have only a basic understanding of pronunciation, are assisted in their understanding when they can read the words on the screen. ASR fundamentally attempts to reproduce speech by essentially re-creating the words by “recognizing” or “comparing” sounds for the purpose of providing some spoken information to a user. Speech recognition used for commercial customer service prompts are a common use of ASR.

The incorporation of ASR into CALL, with direct benefits for learners, came with the development of speech recognition engines that could handle continuous speaker-independent speech. The flexibility of this type of ASR allowed for: easy and convenient applications for standard personal computer (PC) hardware, production of complete sentences, no training periods, immediate practice, a focus on fluency, word order and native models, as well as leniency in the recognition of a variety of accents (Davies, 2006). It is important to note that there are also limitations imposed by the system’s
flexibility: “…greater flexibility allows these programs to “recognize” sentences spoken by students with a wide variety of accents which also limits the accuracy of the programs, especially for similar sounding words or phrases. Some errors may be accepted as correct. (p. 13).”

According to Davies (2006), native speakers may be more bothered than students by the trade-off of accuracy for flexibility, although accuracy for most products still ranges between 90-95% (Alwang, 1999 in Davies, 2006). Davies feels that learners are more likely to accept the limitations of ASR than native speakers because of the satisfaction learners experience when the ASR understands, especially beginners', imperfect pronunciation. The different types of ASR and the tools needed for development as well as the cost of speech recognition programs and products makes their use in CALL especially difficult.

An ASR-patented system, by Auralog, Inc., called S.E.T.S (Speech Error Tracking System) is fundamentally a recognition system that incorporates the three main components of a speech recognition system: recognizer, acoustic model and language model. The recognizer is the engine, the computer programmed “driver” of the system. It is designed by computer programmers using programming and mathematical language, using pre-designed algorithms, or specific rule-based programmed commands. With a recognizer, the acoustic and language models can be developed for a specific foreign language, hence the development of several language versions. For example, Auralog provides ASR with nine language versions.

Many commercial software companies purchase the recognizer component and then build the other components. Several academic institutions have been able to design
and develop their own systems. For example Carnegie Mellon University has the computer programming power, financial assets, a large global interest and they have developed their own speech recognizer (Carnegie Mellon, 1999), a system that has been used and tested by CALL researchers (Eskanazi, 1999). Several other university research-based systems have evolved within the past ten years.

In a technical paper obtained from Fifth Generation Computer (FGC) Corporation entitled: *Speaker Independent Connected Speech Recognition* (2008) several speech recognition systems that employ different techniques for the recognition process are described. An explanation which describes the process of ASR could be summarized as follows: a recognizer algorithm takes initially input sequences of phonemes (sound parts) or words and generates acoustic templates. These templates are compared to the “previously acoustically modeled” templates (acoustic model) and a template of each word in the “vocabulary” is stored and input is evaluated for patterned matches. Close or exact matches are then selected. The basic process of ASR, in a simple representation, can be outlined as follows: *Input Speech* ------*Front-end processing*------*Stored Word Patterns*------*Pattern Matching*------*Decision Rule (decides closest match)* ------*Recognized Word* (FGC, p. 12).

In a speaker-independent system: “there is no training of the system to recognize a particular speaker and so the stored word patterns must be representative of a collection of speakers expected to use the system. The word templates are derived by first obtaining a large number of sample patterns from a cross-section of speakers of different sex, age-group and language or dialect, and then clustering these to form a representative pattern for each word. A representative pattern can be created by averaging all the patterns in a
word cluster. Because of the great variability in speech, it is generally impossible to represent each word cluster with a single pattern. So each cluster is sub-divided into sub-clusters and a number of tokens for each word (i.e. up to twelve is stored in the system). All tokens of each word are matched against the input word.” (FGC, p. 12).

Intelligent CALL, or parser-based CALL (Davies, 2006), as a subfield of CALL and of which ASR is a part, is not yet prepared to address the issue of spontaneous, natural foreign language speech. However, FL speech, speaking and pronunciation can be effectively practiced using ASR software. Davies (2006) suggests several purposes for ICALL including: communicative practice, reported increased confidence from students, and research.

Research using CALL software, where the software tracks whole sentences (such as ASR), detects and records errors and includes other types of visual feedback, has been suggested by other researchers (Chun, 1994, 1998, 2002). For example, Chun (1998) suggested using the tracking features of the software for research and for investigating the visual waveforms and pitch curves produced by the software. Chun (1998) proposed training students to interpret the visuals provided with ASR software and then investigating whether the training and assistance contributed to their understanding during practice.

Studies using ASR have found ASR contributes to FL pronunciation practice where the students used the software (Hincks, 2003; Hemard, 2004; Barr et al., 2005). However, there is a lack of studies examining the actual impact of the use of the CALL software generally and ASR software specifically. Part of the reason lies with the difficulty of using the software for data collection purposes. Although student tracking is
provided in the software and in fact research using these features has been suggested (Chun, 1998), the difficulties encountered in extracting the data make studies using this design difficult and time consuming. In many cases the student data can be tracked, but cannot be exported for analysis. For example, TeLL me More (TMM) Automatic Speech Recognition (ASR) provides a visual score to students after each production. The “score”, or green bars, provided indicate a range of pronunciation between 3 and 7 (1 or 2 bars are highlighted as gray, indicating a “below” recording threshold). There is no information on what the score means and it is assumed that “more green bars” is a better pronunciation.

Several reviewers of the TMM - ASR software, of both French (Reeser, 2001) and American English versions (Zahra, R. & Zahra, R., 2007) have commented on the lack of information about the ASR “score”. Students using the software are responsive and excited to see their pronunciation represented by 7 highlighted green bars and will repeat sentences in order to re-confirm, or try again for a “good” pronunciation. What do these scores mean and more importantly what do they represent, in terms of feedback for pronunciation practice or performance, to FL students? How are the “scores” generated, by the ASR software? Is there a special algorithm that produces this product as part of an internally-programmed process of comparison? Fifth Generation (2008) has one short comment that includes the issue of “scores” which suggests this algorithmic comparative process may play a part: “In the one-stage algorithm each word pattern is matched against a first portion of the input pattern using a dynamic time-warping algorithm (DTW). Then some of the best scores together with their corresponding ending position in the input pattern are recorded (p. 13)”. How can this ASR score information be obtained and how can it be interpreted for the FL researchers and the community of users?
Communication and research with Auralog's representatives have been ongoing since 2005. The company has endorsed and been supportive of university-based research. The Auralog website has expanded and includes references and resources, as well as scholarly articles. An article by Wachowicz and Scott (1999) is particularly insightful and informative. Both research-based and commercial CALL speech products are evaluated and reviewed. In some products, pronunciation can be scored (Vocabulary Builder), but only relative to expected or anticipated responses. In reference to pronunciation scoring, Wachowicz et al. (1999) comment: “Based on the performance of ASR in commercial CALL products, it is our opinion that it is too early for recognizers to grade pronunciation as a background process secondary to word-learning games and other response-selection activities in CALL (p. 268)”.

Wachowicz and Scott (1999) also reviewed an earlier product by Auralog (Auralang 1995) and their evaluative findings are important for this study. “Auralang offers activities focusing on word boundaries and pauses in a sentence (Auralog, 1995)”. The authors continue: “Speech interactive courseware allows the learner to focus precisely on the desired portion of the sentence The system also allows the learner to “see” his or her voice in the form of a sound wave and compare it with the waveform of a model native speaker. We found this kind of activity interesting and potentially instructive. It is not subject to the errors made by ASR when attempting to categorize utterances as this or that word or phrase”. The authors caution readers, “However, we did not test the Auralang activity with students, nor did we find teachers familiar with the software who could give an impressionistic assessment (p. 268)”.

22
Wachowicz et al. (1999) contend, as does Chun (1998) that the waveform is a valuable tool and although teachers may be hesitant, “there is a new audience of technologically sophisticated learners (p. 268)”. Although somewhat dated (ten years have passed since the article was published) the realities of conducting foreign language research with commercial speech-interactive CALL, despite, or because of technological advances, are daunting for CALL researchers and consequently for teachers whose methods they seek to inform.

Auralog’s educational representatives have been helpful and provided information about the accuracy rate of the speech recognition engine used for their ASR products:

The error rates that are usually used in speech recognition engines are related to native speakers speaking in their own language. The speech recognition engine that Auralog is using has an accuracy rate of 99% (error rate of 1%). However this does not reflect the reality of the language learning domain where the speaker is not a native. If a sentence that is extremely poorly pronounced is misrecognized, is it an error from the speech recognition engine (that should account in the error rate) or the consequence of the bad pronunciation? We do not have a simple error rate that would show our performance in our specific domain. We use multiple different indicators that measure the various possible problems: an average pronunciation that gets a good score, a good pronunciation
that gets a poor score, etc. We try to reduce them all knowing that they are contradictory and that the final tuning is based on a good balance between all of them. (Personal communication from Tim O’Hagan, Auralog, May 2009).

The issues relating to automatic speech recognition (ASR), the components designed into the system and the CALL researchers attempting to investigate how the system can be used for instructional FL speech-interactive purposes are faced with considerable challenges. The challenges continue to emerge from several domains such as computer speech recognition designs and advances (Rodman, 1999), the renewed interest in natural language processing (NLP) and computational linguistics, and a focus on foreign languages and consequent sociolinguistic developments worldwide. CALL researchers are challenged as never before to make sense of the rapid and sometimes daunting changes occurring within the CALL domain (Chapelle, 2007).

**Human Raters for FL Evaluation**

Early foreign language classroom teachers could be considered the first human raters as they were charged with evaluating the speech samples produced by their students. Valette (1965) in a seminal work on modern language testing, written for FL classroom teachers, describes the design and evaluation of speaking tests. Referencing Lado (1964), considered the “father” of language testing, Valette describes how teachers can conduct speaking tests and how students’ pronunciation, intonation and other speech features can be evaluated. In each instance, the teacher is considered the judge or evaluator and is cautioned to understand the specific problems faced by students. For
example, English-speaking students often have problems with multiple-syllable word stress and the English stress system allows word stress on the first syllable, second syllable as well as weak stresses, depending on the word. Students, in their early FL learning, apply English word stress patterns to FL words. Misplaced stress can result in a word that is incomprehensible to or misunderstood by a native speaker (Valette, p. 93).

Valette (1965) addressed the issues of teacher-rater reliability with regards to speaking tests and provided a clear discussion of ways for increasing score reliability and consistency by limiting the rating to a specific feature or aspect of an utterance. Suggestions are made on how to design scoring systems or evaluation categories and forms for rating. The issues of rater reliability that were pertinent to FL teachers in early FL classrooms remain important, if not paramount, to teachers and human raters of FL speech today.

Valette’s (1965) discussion highlights the problems that are inherent in human rating, problems which have remained basically the same over the years, despite the development of CALL materials and technological advances. Standard protocols for insuring rater reliability and consistency in testing situations, that is where human raters are used to evaluate FL speech production and outcomes, have been developed. Protocols for human rating are routinely used in any case where human error has the potential to contribute to standard error of measurement (SEM) related to the test items or test format, in any testing situation (Weir, 2005).

Rater training is considered necessary to insure that raters are instructed in the features of the rating scale and practice using the scale to rate actual speech samples. Rater training is one way believed to increase the consistency of the ratings raters assign
to the speaking samples (McNamara, 1996; Weir, 2005). Practice rating and training are suggested at all levels of FL speaking rating and may be especially important for beginner levels where speech may be undeveloped. Targeting specific speech or pronunciation features, in short sentences designed to elicit those features, as suggested by Valette (1965), can assist in training raters, and in their actual rating of FL speech, in both direct and indirect methods.

Although Valette (1965) suggested limiting the features of an utterance for teacher evaluation, many language test developers proceeded to use interviews, as a direct method of rating FL speech (Underhill, 1987). Any direct method of rating speech involves having the rater, or raters, present with the examinee and rating is done within the context of the speaking test situation. Indirect speaking test methods involve audio or video recordings of FL speech or pronunciation that are evaluated outside of the test environment. Indirect methods were available in the 1950-1960s. However, the lack of mechanical and technological sophistication made for a time consuming process for recording and playback. This in turn made listening and re-listening difficult and frustrating for raters.

The American Council of Teachers of Foreign Languages (ACTFL) (1985) was an early proponent of the direct method and the development of the ACTFL-Oral Proficiency Interview (OPI) was an attempt, on a large scale, to standardize a form of direct speaking test. The ACTFL-OPI met with considerable controversy (Lantolf et al., 1987) and the methods and evaluation criteria were severely criticized by foreign language researchers. Despite the early criticism, the ACTFL-OPI remains a regularly used instrument for evaluation. As well, ACTFL has established itself as the preeminent
FL organization for foreign language teachers and researchers. ACTFL has developed a sophisticated, time and cost intensive process for training human raters, with the purpose of standardizing the rater training and scoring processes.

ACTFL has also been important in the development of scoring rubrics and descriptors used for evaluating FL tasks. Initially, ACTFL (1982) borrowed and adapted the numerical scoring system developed by the Interagency Language Roundtable (ILR), a system originally developed for the Defense Language Institute to teach foreign languages to military personnel in an immersive and accelerated program. ACTFL, however, addressed different needs—those of professors teaching FL at the university level. University students were learning under very different conditions from military personnel. ACTFL dropped the numerical scale (1-7) and used the ILR foundational framework to develop descriptors marking attainment levels. Eventually, ACTFL collapsed levels and assigned low, mid and high to both the Novice and Intermediate Levels to accommodate learning levels observed in university FL courses. In 1999, in response to a defined interest and more diversified need for evaluating speaking, ACTFL developed the ACTFL Guidelines-Speaking 1999. A rubric and descriptors were developed to address the need for a direct method of evaluating FL speech and speaking performance that was not specific to an interview setting.

In the 1980s, with the work of Canale and Swain (1985) and their proposition of a communicative approach to language teaching, the shift toward language for communication resulted in a focus on authentically-based texts and resources. Throughout the 1990s, technological advances in multimedia, including audio and video, significantly contributed to the development of computer-assisted language learning.
CALL developments also contributed to the expansion of materials that included considerations for ways to include not just recording and playback of speaking skills, but also spontaneous speech, intonation and related features such as pitch and stress, pronunciation, conversation and discourse. Chun (1994), in her early study, concluded that the words students wrote using the computer were a form of spoken communication or dialogue. Subsequently Chun proposed a more inclusive view of FL speech that included intonation and discourse features. Chun (1998) focused on CALL and the possibilities for FL student learning. Chun’s vision and suggestions were significant to the direction of the current study. Her critical review of the nature and functions of discourse and intonation (Chun, 2002) are significant contributions for the direction of CALL, FL speaking skills, discourse and testing.

Chun (2002) critically reviewed the ACTFL guidelines and evaluated the current descriptors. Chun found a distinct lack of focus on evaluating and rating speech, intonation, pronunciation and related features. The descriptors did not address features critical to speaking except, marginally, at the higher attainment levels. The novice and intermediate levels, levels most often found in students studying FL at the university level, were disturbingly incomplete and provided few or no descriptors and guidelines for FL educators to use to evaluate students at these more beginner levels.
Chun (1998; 2002; 2007) suggested a need for more research using CALL materials and suggests the need for more complete scales and descriptors developed for evaluating speech by raters, as well as for teachers. If the state-of-the-art for human raters has remained essentially the same for the past ten years, given the established and prevalent elements of human error in rating, research needs to be directed toward addressing the lack of methods and paucity of qualified rubrics for scoring and evaluating speech data. Technological advances have provided ways for CALL and multimedia materials to collect speech data, however the components necessary for evaluating speech are more complicated and require understandings and expertise in areas such as natural language processing (NLP) and automated scoring. Previously, these areas have been unexplored by CALL developers and FL researchers (Chapelle et al. 2008).

Advances in multimedia and the advent of the internet in the 1990s, have contributed significantly to foreign language speech and speaking both learning and teaching. Automated and instantaneous feedback, speech recognition, voice-over internet protocols (VOIP) such as SKYPE and online communities such as Second Life, have all impacted advances in learning and have contributed to a more participatory and inclusive virtual world. Global educational and environmental developments in Asia, Africa and the emergence of a unified European community have all contributed to the multicultural and multilingual online world of learners. The thrust toward a more social and cultural world fabric have resulted in a renewed focus on understanding cultural similarities and differences, regional languages and dialects, as well as national and regional boundaries. Online communication and transaction are commonplace exchanges at all levels of society and among all ages.
Internet-based and online FL tests that can record, provide feedback and evaluate speaking skills are used routinely for placement purposes, in courses, for certifications, work-related training and admissions. In high stakes testing, such as the Educational Testing Service (ETS), Test of English as a Foreign Language (TOEFL-ETS) internet-based test, a speaking section was added in 2005. During the test, foreign language examinees produce task-related speech samples that are recorded and then sent to ETS raters for rating. The raters are ETS certified and trained and examinees are notified of their speaking score about 7-10 days after the actual test. Speaking scores are then reported as part of the total test score.

Within the high-stakes testing arena, recent work by Cambridge ESOL researchers into the IELTS (International English Language Test) speaking test resulted in the first-time addition of pronunciation criterion and a pronunciation scale (Develle, 2008). Early studies using raters revealed problems with raters interpreting the pronunciation scale. The findings during a two-phase mixed method study resulted in revisions that included a larger point range as well as half point values being added. The revised pronunciation scale became fully operational in August 2008 and monitoring of Speaking test performance data, including functioning of the pronunciation scale, is part of an ongoing research agenda. Using both correlation analysis and feedback questionnaires, ESOL researchers were able to refine a pronunciation scale used at the international level. Cambridge ESOL has been at the cutting edge of speaking and pronunciation testing.

Interestingly, even at a high stakes level of testing such as the iBT-TOEFL, raters are used for rating non-native speaking tasks. The speech is produced by examinees, the subsequent sound files, or recorded speech samples are sent via the internet to the raters
for rating. Unfortunately, the quality and sophistication of the hardware used to collect
the test audio samples are to some extent dependent on the headset/microphone used for
recording and listening and on the internal computer sound card. The six TOEFL
speaking tasks, the audio samples, rated by ETS raters are between 45 seconds to 2 1/2
minutes long. Likewise, the raters are limited by hardware specifications. Efforts are
made to minimize or eliminate any reduction in sound quality.

Researchers at ETS have been exploring the possibility of using automated speech
scoring. In a recent document, Xi (2008) reported on the development of Speech Rater v
1.0, a speech recognizer (a main component of an automatic speech recognition system)
used for automated rating of speech produced by students using a TOEFL speaking
practice software. To date, this is the first automated speech scoring instrument that has
been developed or used by ETS. The Xi (2008) report was directed toward the need to
establish scoring validity criteria before deploying the instrument in an actual test
situation.

An established international language testing body, such as ETS, and the eminent
researchers, as well as many foreign language researchers and language test specialists,
are directing attention and resources to the development of measuring and evaluating FL
speaking and speech. Using raters to evaluate speech remains an important feature of the
developments. However, automated scoring and the use of speech recognition to evaluate
speech is, as reported by numerous researchers, relatively new and unexplored terrain.

Given the potential benefits to FL students and teachers in an informal
instructional or testing context such as the classroom or language lab, how does FL
speech rated by trained raters compare to an automatic speech recognition (ASR) score?
Is the score produced a face validity score for evaluation purposes? What does the ASR score indicate to students and how do students interpret the score? What form of feedback does the score represent for students? The current study proposes to investigate the scores produced by automatic speech recognition software by providing FL raters with the same recorded speech samples participants produced with the ASR software.

**CALL Research Using Speech Technologies and ASR**

Second language acquisition (SLA) and applied linguistics research have expanded greatly on foreign language teaching and learning theories and methods. Communicative competence in foreign language (FL) learning (Savignon, 1997) that is authentic language use for communication, and the more recent use of computer-mediated communication (CMC) in foreign language teaching (Warschauer, 1996), has resulted in a more inclusive view of FL learning. Warschauer interpreted this view as one that included “empowerment” with a distinct focus on learner control and language production. Computer-assisted language learning (CALL) frameworks (Warschauer, 1997), multimedia software applications (Lafford, 2004), and developments in speech recognition technology (Eskenazi, 1999), have all contributed to enhanced FL communication, listening and speaking skills and student motivation (Hémard, 2004).

The concurrent and residual developments within SLA are those experienced in many other domains where a lack of research or an inability of researchers to proceed as quickly as the technological developments, has hindered our ability to conduct research that can inform methods and materials. Many areas of SLA research are of interest. However, as CALL is the domain of SLA researchers specifically interested in
technology and in using technology for teaching, it appears that the rapid technological developments most concretely affect CALL research. Several researchers, over the past ten years, have suggested using specific CALL products and materials for research (Chun, 1998; Levis, 2007; Xi, 2008), to inform teaching, to contribute to student learning and to design better CALL materials.

Materials that have contributed to CALL development are multimedia for teaching and practicing pronunciation (Clear Speech, 1995, 2003; Tell me More, 1995, 2003) and improving speaking skills (Streaming Speech, 2005). Other FL multimedia has incorporated video, audio and numerous learning activities incorporating all four skills. Many of the technological features included in CALL materials are the domain of computer scientists and instructional technologists, as well as programmers and information technologists and trainers. It is this overlap of domains within CALL that is most problematic as regards rapid technological developments. CALL researchers need the technological expertise to interpret the developments, both academically and for the foreign language community of learners and teachers. As well, CALL researchers are called upon to create a bridge of understanding and communication with the other domain specialists. Commercial FL software developers need information regarding actual users of their software (Nielsen, 2006) to inform future iterations. However research-based investigations are procedurally-defined, process-oriented and time consuming. The area entitled human-computer interaction (HCI) has grown from the need to understand how users interact with the computer, specific programs or software and to observe and research the interactions. As well, commercial products are confined by patents and other legally exclusive contracts that can make it difficult for researchers
to conduct investigations. Likewise, the internet and a vast array of diverse resources, has produced another challenge for CALL researchers willing to provide teachers and students with quality and timely learning materials.

Speech technology, and specifically automatic speech recognition (ASR), was designed to recognize produced foreign language speech and is used in CALL for teaching foreign language intonation and pronunciation to second language learners. Pronunciation and intonation are features of FL speech and discourse that have gained renewed emphasis both in FL teaching and research. Chun (1998) outlined several reasons for the recent developments in speech recognition software for teaching pronunciation. For example, Chun (1998) cites the sophistication of the visual and digital audio technology for the production of speech sounds, more sophisticated models for intonation and speech processing and contextualization of speech and learner performance feedback.

Chun’s (2002) recent work with discourse and CALL has centered on the shifts in communicative language technology and research that have resulted in a renewed focus on speech intonation as a fundamental part of discourse fluency and FL proficiency. Chun (1998) noted the development of contextual, discourse-based FL multimedia software and suggested a need for research on the effectiveness of computer training materials (p. 79). In her work, Chun theorized that FL communicative methods and computer-mediated communication (CMC) can contribute to more discourse-based teaching of pronunciation and intonation.

Increasingly, the view of intonation as an integral component of language communication (Chun, 1998) has contributed to CALL software designs developed to...
include the practice of speech discourse beyond the sentence. Constructed oral dialogues used in CALL designs, as models for conversation and “authentic speech”, have enhanced the productive possibilities available to FL students. The development and addition of ASR to CALL designs has further enhanced the content and quality of performance feedback. Chun (1998) suggests more inclusive CALL research to assess the types and effective combinations of relationships between oral feedback and production. The essential questions in any CALL evaluation (Chapelle, 2001), including speaking skills and speech production, centers on student usefulness and effectiveness for production of FL communicative discourse.

Levis and Pickering (2004) demonstrated how intonation functions in discourse using speech technology and discussed the implications for teaching elements of intonation. Brazil’s (1992) model was used to categorize degree of engagement (1=low, 5= high) where each level was characterized by different tonal qualities. Levis et al. found that text-based readings were associated with higher levels of engagement than sentence-based readings and demonstrated how the context of discourse affects intonation.

Research on speech recognition software used to evaluate foreign language pronunciation and intonation demonstrates that student’s production of communicative discourse can benefit from skills practice. McIntosh et al. (2003) used an online web-based product (WIMBA) to record FL student speech samples and oral activities. Students were evaluated by the teacher on discrete pronunciation features and spontaneous speech. Responses to the use of the tool were positive and students reported feeling more confident with FL language use.
Automatic speech recognition (ASR) includes production, practice and evaluation of FL speech. Student perceptions of speech production and ASR have contributed to evaluations of the effectiveness of speech recognition software for the improvement of FL intonation and pronunciation. Hincks (2003) evaluated several speech recognition software programs for student perceptions on feedback and pronunciation features. As ASR is designed, using underlying algorithms, to compare “recognized speech” to an underlying model, it is only an effective evaluation tool when produced speech is compared to the underlying model. ASR algorithms are effective for determining how much speech samples differ but cannot determine the ways speech samples differ from the models. Feedback evaluation methods using ASR are still in development phases.

However, Eskenazi’s (1999) FLUENCY project at Carnegie Mellon University (CMU) and her pilot experiments have demonstrated the use of speech technology for effectively recording prosody, or pronunciation features of speech responsible for intonation and rhythm. Contextualized ASR for meaningful FL communication is one of the many practical recommendations, as the development of speech prosody is considered an important feature of communicative discourse.

Eskanazi’s (1999) work highlights the importance of research on ASR and the practical applications for FL pronunciation, for CALL designs and for FL teaching and student performance. The development of ASR can significantly contribute to the FL curriculum and as an adjunct to the teacher’s role of FL speaker model for discourse and conversation (Eskanazi, 1999).
CALL Evaluation

In an effort to determine the effectiveness of foreign language CALL materials, Chapelle (2001) developed appropriateness criteria considered fundamental for CALL design. Jamieson, Chapelle and Priess (2004) demonstrated the use of the criteria in an evaluation of the design of Longman English Online. Judgmental evaluation was used to examine several instructional variables and results indicated that in an evaluation of interactive features, interaction “between people and between people and the computer” (p.412) were good. However, the intrapersonal interaction which includes learner initiated choices was weak. Can CALL that includes ASR for FL oral skill practice contribute to the intrapersonal interaction Jamieson et al. (2004) found weak in the English program? Warschauer (1996) concluded that student choice and empowerment were the definitive features required for effective foreign language learning in a computer-mediated communicative (CMC) environment.

A feature of CALL evaluation (Chapelle, 2001) includes, among other features, positive impact or questions related to the tasks designed for CALL and student responses. Chapelle (2001) suggests that CALL must support evidence of positive impact through an examination of questions related to student learning and experiences. Research designed to qualitatively investigate FL learner observations and experiences, within the context of task use, is essential to an evaluation of the positive impact of CALL.

Hémard (2004) designed a CALL program for use with French language learners and evaluated the effectiveness of the audio and interactive features from the perspective of the user interface design and student experience. Using user-centered evaluation
methods such as questionnaires, user walk-throughs and focus groups, Hémard demonstrated student participation in the interactive learning experience. Hémard concluded that evaluating usability in the context of use is critical for generating new and effective materials for inclusion in CALL.

Plass (1998) created criteria for a domain-specific (FL reading), user interface evaluation and used a cognitive approach to design. Using CALL software, Plass applied his criteria to an evaluation of Cyberbuch/Ciberteca (Plass, 1998) and demonstrated how the model of cognitive design included a FL domain-specific, user-centered approach. Through an identification of the mental processes involved in SLA-related reading activities, Plass demonstrated how a cognitive design contributed to the effectiveness of multimedia tools “that function as aids to cognitive processes involved in the development of particular linguistic and pragmatic competency” (p. 45).

Handley and Hamel (2005) support evaluation of speech synthesis systems and believe research on FL student user satisfaction can contribute to the inclusion and use of speech synthesis in CALL. Building on the evaluation criteria developed by Chapelle (2001) Handley et al. (2005) evaluated several products in an effort to establish benchmarks, or essential requirements, for the use of speech synthesis as an aid to acquisition of spoken language.

Using the criteria of comprehensibility, naturalness and accuracy (Handley et al.), as indications of whether speech recognition software effectively manipulated speech output, Handley concluded that different language functions require different elements of speech recognition. Consequently, some elements could be more useful for pronunciation development than for conversation. Handley et al. suggest that understanding various
speech features and language skill modes through user-centered evaluative criteria can help to determine the acceptability of computer training materials for CALL inclusion.

Smith and Gorsuch (2004) used a university usability lab (UL) and related technologies to examine synchronous CMC (SCMC) within English second language pairs. The capabilities for data collection provided by the UL allowed a comprehensive examination of audio, video, and computer interaction combined with student interaction and response. The interactive, communicative nature of speech was documented and analyzed. In audio recordings of screen captures Smith et al. (2004) found evidence for verbalizations or private, “barely audible self-speech that occurred among a few participants” (p.564). Smith et al. (2004) defined the findings as “verbalized private speech, cross-modality uptake, or practice” (p.564). Smith suggests that future research on ‘verbalizations’, private speech samples, “may provide insights into meaning-based tasks…and into what participants have learned or are learning through an interaction” (p.572). Could ASR used in CALL software help researchers to evaluate the productive or oral nature of private speech? Did Smith et al. (2004) find examples in their study for the phenomenon Jamieson et al. (2004) described as “intrapersonal or within person” interaction?

In a discussion of the tutor/tool model of CALL software, Hubbard and Siskin (2004) provide evidence for a return to tutorial CALL as an effective tool for learner-directed language learning. Central to the model is the issue of whether FL software is used as a teacher replacement or a tool for learning. Hubbard et al. (2004) convincingly document a case for convergence of the dichotomous model and no longer see the categories as mutually exclusive. A more inclusive view of tutorial CALL, and one that
includes learner-centered designs, includes “one of the most current and active areas in tutorial software development, pronunciation” (p. 459). Further, Hubbard describes the features of the multimedia models from the user-learner perspective and states these programs “allow the user to record and compare, and can provide various forms of sound visualization including wave forms and pitch contours” (p.459). Hubbard believes that feedback from users and the results of informed research will be combined to enhance future CALL applications.

FL student perceptions of learning and performance are essential to an effective design and evaluation of software developed for learning. Barr et al. (2005) cite research from Burnage (2001) that supports students’ recognition and appreciation of CALL software designed with FL pedagogical value. Barr (2005) found that students understood the limited nature of ASR for practice of conversational speech. The findings resulted in the conclusion that “technology is perhaps best kept out of free conversation and integrated more into pronunciation drilling and the development of associated skills” (p.72).

Research studies have used CALL methods (Jamieson, Chapelle & Preiss (2004) to design and investigate multimedia product effectiveness and learning. Technological developments and refined tools, such as speech recognition software, have contributed to more sophisticated methods for collecting data on FL speech production. Research using speech recognition software to provide native speaker examples, record and score student speech, provide feedback and collect FL produced speech could contribute to enhanced or improved student learning and performance.
Auralog, a French software company, has been a forerunner in the development of speech recognition software. Several researchers (Handley, 2005; Hincks, 2003) have used TMM versions for various tests or evaluations. Several studies examined and evaluated CALL from a FL user perspective.

Several researchers have used TMM foreign language software in studies either as content, activities or resources. For example, Barr et al. (2005) conducted a pilot project at the University of Ulster (Ireland) using TeLL me More (version 5, Auralog). The oral activities were conducted and data collected using the speech recognition features available in the software. The researchers suggested that the design of the CALL software (TMM French) seemed to reflect a progression of practice stages beginning with pronunciation drilling and progressing to simulated interactive role-plays (p. 60).

A pilot project conducted by Barr et al. (2005) is a significant contribution to CALL and the fields of second language acquisition (SLA) and applied linguistics research. First, the pilot study was the first part of a longitudinal research project designed to examine the use of multimedia software for oral language learning and conversation over the course of several years. Initial conclusions reported that the multimedia lab was an effective venue for tutorial, practice and assessment phases of oral work. Barr et al. made a case for the use of technology for certain phases of oral skills teaching and learning (p. 61).

Secondly, in Barr et al.’s (2005) initial CALL experiments both quantitative and qualitative measures were used. The quantitative research focused specifically on an evaluation of oral language learning (French), with task and skill comparisons.
Qualitative analysis documented student support for learning using technology. Students made little reference to problems with learning or with technology use. The findings are inconclusive, but students reported positive impact (Chapelle, 2001) with the use of computer technology to enhance their oral skills (p. 72).

Finally, Barr et al. (2005) expect to expand, further refine and develop the study following a thorough consideration of the initial findings. The TMM software is adaptable and flexible enough to be used for specifically collecting FL speech data. In the future design, Barr et al. plan to include individualized learning paths as diagnostic tools and to allow for student differences in foreign language proficiency.

Hincks (2003) used *Talk to Me: The Conversation Method* (an Auralog ASR program, without the full course features of TMM) as an additional pronunciation resource for adult learners of English in a second language environment in Sweden. The program was found to be beneficial for students with strongly accented (intrusive) pronunciation as measured on a standardized oral assessment, *PhonePass*. Ordinate designed the assessment to be administered by telephone. The student responses were recorded and rated. Hincks used Chapelle’s (2001) CALL criteria as a framework to evaluate the efficacy of the CALL tasks designed into the *Talk To Me* software. Further, Hincks (2003) evaluated the ASR and described several problems that are a result of the inherent limitations of ASR, one of which includes the ability to recognize and reproduce spontaneous speech. However, the results indicated that ASR-based training was useful to beginning learners (p. 17).

Handley and Hamel (2005) designed a study to examine the possibility of developing a method for adequacy tests (benchmarks) for speech synthesis software used
in CALL. For this case study, a conversation corpus was designed and a simulated ASR-based dialogue from *Talk to Me French* was used for the “20 consecutive turns corpus” (p. 109). Although the study did not specifically investigate features of the ASR, participant’s ratings of accuracy and appropriateness of the corpus and simulations demonstrated that contextualization of ASR for oral activities could assist with an understanding of which roles and utterances are acceptable for inclusion in CALL applications (Handley et al.). The recent studies using automatic speech recognition, and specifically the ASR software used in TeLL me More, point to the sophisticated nature of the software and its perceived potential for improving FL learner oral performance.

The research literature points to a distinct interest in the development of computer-mediated tools and activities designed to assist FL learners at all levels of skill learning. Students have through time consistently expressed an interest in learning to speak the language and to use their language for useful purposes. Teachers have a responsibility to assist them by providing resources and materials that are directed toward the development of these skills.

Speech and pronunciation practice for communicative and authentic interaction is an important goal in a progression toward effective language use. Documented research has demonstrated that FL students are receptive to and empowered by the development of their ability to use the language, especially in the early stages of learning. Automatic speech recognition (ASR) can be used as a powerful tool to assist students with speech production and use, both within academic contexts and in the real and virtual world.

Through an examination of scores produced in a computer-mediated environment using ASR software, scores that are then given to students as feedback on their produced
speech, teachers can further meet the needs of their students. Teachers have the responsibility to provide information on the meaning of the scores, whether in a testing situation, online or in the classroom. Information on the meaning of the scores can assist teachers and researchers with the interpretation of the feedback and scores for the FL community as well as for materials developers.

Human raters have been used for many aspects of research and especially in rating FL speech where their language skills and expertise have been and continue to be needed. By having human raters rate FL speech, samples of audio feedback produced by ASR software, rater insights can provide an unprecedented dimension to a previously unexplored aspect of their work.

By combining human rater perspectives, ASR scores and audio produced by FL students, as well as considering student perspectives, a more complete examination of the value and usefulness of the ASR software as a contributory factor in FL pronunciation and speech learning can be made. As CALL continues to incorporate and use features that are technologically inventive and responsive to the needs of users, ASR will continue to add to the resources and materials available to all participants in a multilingual and global community of learners.

In response to the need for an investigation into the value and interpretation of the ASR score feature and by using expert human raters, this study was designed to challenge the limits of CALL research within the constraints of current technological practices. The difficulties of design and administration using ASR software to investigate French learner input and outcomes are detailed in order to assign value to a numerical representation and for future replication studies. Participant-learner feedback is presented as evidence for
consideration in future CALL evaluation and designs. Ideally, the study will serve as a model for CALL researchers and further our understanding of the usefulness of ASR for future designs and to assist with an understanding of ASR within the FL domain.

Importantly, it is hoped that the study will be interpreted as a call for attention from FL student-users, to FL teachers and researchers, to address their documented interest in learning to pronounce and 'speak the language'.
Chapter Three: Method

Introduction

Several pilot studies have examined FL student use of the Automatic Speech Recognition (ASR) software. Each study provided insights, observations and direction for the current method design. The pilot studies have significantly contributed to the final design of the study and actual testing of the instrument, as well as preliminary data collection and analysis. The pilot studies, the design features, considerations and conclusions that have directly influenced the methods for the current study are described, where necessary and appropriate, in the following sections. An account of the pilot studies can be found in Appendix A.

The use of Automatic Speech Recognition (ASR) software for FL learning has been reported in numerous studies and the efficacy of the ASR as an aid to FL speech and pronunciation has been reported and observed (Hincks, 2003; Levis & Pickering, 2004; Levis, 2007; O’Brien, 2006). However, many questions have remained unanswered and empirically unexplored. For example, what would the results of student performance using ASR reveal when compared to qualified FL raters using established performance measures? Importantly, how do students view ASR and what is their evaluation of its purpose and effectiveness for their foreign language practice and learning? Finally, what is the contribution of ASR to CALL research and to student learning and performance?
Research Questions

The following investigative questions were proposed and have been used as the research questions to conduct an analysis of the actual and potential pedagogical value of ASR for student FL practice, performance and learning:

1) When using the same ASR-produced sentences for human expert rating, how do human ratings compare to the ASR scores provided to students and how valid are the scores?

2) How useful did the students find the ASR features were for assisting them with their French pronunciation practice?

3) How effective did French students perceive the ASR software to be for improving their mastery of French?

Design

A hybrid method, incorporating both quantitative and qualitative measures, was used (Tashakkori & Teddlie, 2003). A quantitative analysis (Hatch and Lazaraton, 1991; Mackey and Gass, 2005) using descriptive statistics and correlation were used to analyze the variability and relationship of the ASR scores and human ratings of recorded French sentences. Qualitative measures (Brown, 2001; Dornyei, 2003; Patton, 2002; Strauss and Corbin, 2008) were used to analyze the qualitative data obtained from an online survey. The survey questions were designed to elicit responses from participants regarding their experiences using the ASR software to record and listen to French sentences.
Participants

Participation, using IRB-approved informed consent protocols, was sought from 30 undergraduate French language students. Students in the beginner levels (I and II), enrolled in or having completed one semester of French were included in the sample. Participants attended an orientation to the ASR software at the foreign language computer lab prior to the data collection session. There are 15 computer stations with the software available at the FL lab. However, in an attempt to insure an environment as free as possible of extraneous noise, approximately ten students participated per data collection session. During the data collection all participants were randomly assigned a number and a password to access the pre-designed learning path of ten French sentences. The same participant number was used for access to the qualitative online survey administered at the end of the session. In this way, the individual participant sentence recordings and scores were cross-checked with online survey responses. Additionally, participant composite data, including scores and perceptions are available and provide the basis for post-hoc individual case analysis.

Procedure

Using the ASR features of the multimedia foreign language software TeLL me More, the participants used an assigned password to enter the pre-designed learning path leading to ten French sentences for oral production and repetition. Logitech headsets with attached microphones were used by each participant for listening and recording. Following the oral production and recording, participants were given web access to enter a secured, online survey.
The survey was created online using the survey tool provided through the university academic computing department. Upon completing the survey, participants submitted the survey online for data collection. A sample of the survey is included in Appendix D.

**Sentence Pronunciation and ASR Scores**

Using a pre-designed learning path consisting of 10 French sentences (Appendix B) taken from the 540 sentence pronunciation corpus provided by the software, each participant orally produced and recorded each sentence at least one time. Each sentence repetition is ASR scored however the software was preset to save only one recorded repetition of each sentence as a separate audio file. The data collection session included ASR scores, one score for each of ten sentences for a total of ten ASR scores and ten accompanying audio recordings of the same ASR scored sentences, per participant. The 30 participants data sets resulted in ten ASR scores and ten audio recordings per participant, for a total of 300 ASR scores and 300 accompanying audio files.

The ten sentences selected from the 540 beginner level corpus were chosen for pronunciation features identified for beginning learners. Several sentences include the same pronunciation features, for example the “r” found in the word français. Several textbooks and workbooks were consulted for the purpose of aligning the sentences found in the software with comparable sentences for beginner level French learners. The ten French sentences were selectively chosen from several of the beginner units to reflect a generalized “French language” learning theme.
As the sentences were designed into the beginner level software corpus, it was necessary to design the study using sentences from this group. However, the following detailed selection process was designed to identify the ten sentences, from the 540 sentence group, and subsequently used for the study.

1) Review all 540 sentences within the beginner level.

2) Identify sentences from specific units that located the French “r” is several positions within sentences and words.

3) Choose 25-30 sentences that included both the “r” in varying positions and identify a generalized overarching theme.

4) Consult with both a French language and a linguistics expert for input on the first draft of 25 sentences.

5) Diminish the list to 10-15 sentences based on sentence examination, suggestions review and finalize a general theme. Review the final ten sentence choices and translations with experts.

6) Using the TeLL me More software “Tutor Tools”, set-up and trouble-shoot the pre-defined learning path of the ten chosen sentences for the study.

Hardison (2004) also used French sentences for her pronunciation study. However, the sentences used in the study had been previously recorded in a presentation by learners and were later extracted from the recorded student productions. Unlike the French sentences used in Hardison’s (2004) seminal study, the ASR software contains corpora from which the sentences were chosen.
Digital Data

During the data collection session the sound files and ASR scores were saved in the software, on the main computer at the lab. The main lab computer served as a server for the 15 computer stations that contain the software. Following the data collection sessions, all participant audio files were identified, copied, extracted and saved to CD-ROM. The participant numbers identified the sound files associated with the participant recordings. As the audio files are saved in an open source format (.ogg files) a conversion program was used to “batch” convert the .ogg files to a more accessible MP3 format. Once converted the files were coded according to participant and sentence recording (e.g. p1s1 = participant 1, sentence 1).

During an earlier pilot study (April 2008), audio files collected were used to test the identification, copying, extraction and secure saving of audio files. During this pilot testing the coding system (p1s1, p1s2 etc.) was developed and the audio files were coded using this system. As well, the open source files were converted to MP3 format using an accessible conversion program. The conversion process digitally compresses the sound/audio to the MP3 format so that the audio files can be played on an embedded Flash Player. The Flash Player was embedded into the design of the Rater Site so that the raters were able to quickly and conveniently listen to and rate the participant sentences. After the sound files were extracted from the lab server computer, converted to MP3 and coded the files were then saved to a web space, then “connected to” the Rater Site and made accessible to the human raters for rating.

Technically, the process involves a small amount of HTML code written and input to the survey to direct each sentence to be rated to the corresponding audio file. The title
Rater Site is used as the name given to the online model the raters used for listening to and rating the ASR sentence recordings. Raters input their ratings of the individual audio recordings to the Rater Site using a mouse. An adapted numerical descriptor scale (3 novice – 7 advanced) was used for rating and a paper copy of the scale was available during the rating session. Raters used headsets to listen to the sentences and a mouse to input the ratings to the Rater Site.

The ASR scores are assigned one of seven numbers, low to high, 1 through 7. However, in the pilot studies the ASR scores assigned to the sentence samples (120 audio samples) there was only one ASR score = 1 and only three ASR scores = 2. It was observed that in the case of the low numbers it appeared the audio files were poor recordings and not that the student pronunciation was poor or incomplete. For this reason the #1 and #2 ASR score were collapsed into #3 for raters. Several researchers (Luoma, 2004; McNamara, 1996; Weir, 2005) have suggested that raters are only able to reliably handle rating scales that contain four to six levels. Additionally, the adapted rating scale of 3 to 7 was found to be adequate for the purposes of this study.

Data collected included: participant identifier number, sentence identifier number, ASR scores per sentence and human rater scores for each sentence. All data collected was transferred to an Excel data file for analysis. In order to obtain ASR scores, the three expert French professor raters, two native, and one near-native, speakers produced the same ten sentences as the participants. The raters ASR scores verified their expert status as raters with ASR score = 7, for 27 of the total = 30 ASR scores.
Human Raters

Three French professor raters, including two native and one near-native speaker adults, listened to and rated the recorded sentence audio files for each participant for a total = 300 ratings per rater. Two of the raters had rated during the pilot study (April 2008) and had been previously introduced to the rating procedures. A third rater, also a native speaker, offered to rate for the study.

Raters used a scale (Appendix C), adapted from both the earlier ILR (Interlanguage Roundtable) scale (Appendix C) and the ACTFL Guidelines-Speaking 1999 (Appendix C). Currently, the ACTFL Guidelines are a widely used and accessible foreign language descriptive scale for rating FL speaking. Seven levels of the Speaking 1999 guidelines have been adapted and were used for rating according to the following descriptive levels: Novice: no response or low = (1-2) - 3, mid, high = 4; Intermediate: Low = 5, Mid, High = 6; and Advanced = 7. A copy of the adapted rating scale descriptors was available to the raters during the rating session. The 7-level descriptive rating scale was designed into the Rater Site, using “buttons”, so that the raters needed simply to “select and click on” their numerical choice for each participant audio sample.

The raters were trained, accessed the “sites”, and rated the audio sentences using a laptop (Dell Windows XP, 2006) and headphones, in a quiet, comfortable space in the university library. The researcher was present for the training and during rating to answer questions and to observe. The same process was used during an earlier pilot study and proved efficient and comfortable for the raters. In order to determine an intra-rater consistency check for each rater, raters were asked to re-rate ten random participant
sentences following the rating session. A random sample of sentences was pre-selected and put into a separate Rater Site entitled: Re-Rate Sentences. The following random sequence of ten participant sentences \((N = 10)\) was used to check intra-rater consistency where \(P = \text{participant and } s\# = \text{sentence number:}\) P16s10, P18s8, P23s6, P26s4, P29s2, P34s1, P37s3, P40s5, P43s7, P45s9.

**Rater Site**

The Rater Site was a model created for the purpose of ordering and labeling the sound files recorded by the participants. Further, the site design provided a simple, direct and consistent way for raters to listen to the sentence pronunciations and then input the ratings for each participant. The survey was designed with the university-based survey tool, was accessed online, and was designed for rater ease of use, consistency and efficient organization. In this way, the design, layout and organization of the Rater Site, as a model for rating FL speaking assisted the raters with the concentration and rating tasks required of them for listening and rating numerous pronunciations.

In a previous pilot study (April 2008) the Rater Site prototype was developed using actual sentence sound files obtained from beginner French students. The pilot study was designed for the sole purpose of field testing the sentence audio file tracking, copying, extraction and saving. Although the pilot sentences and audio files were not the focus of the current study, the sound files collected proved indispensable for developing and testing the Rater Site prototype.
Rater Training

The rater training was conducted for approximately fifteen minutes, prior to the rating session. During the training raters were oriented to the Rater Site, the location of the sentence sound files, how to input ratings and use of the headset for listening. The content and descriptors of the rating scale were discussed and raters were presented a paper copy to read and review. Using a separately prepared rater site entitled: Rater Training, the raters practiced using the scale to rate sentence pronunciation samples.

The Rater Training consisted of an initial fourteen participants where each of the fourteen produced all, or several, sentences and audio files and completed the online participant survey. In some cases, due to an initial problem with the software options settings for saving the audio files, only one sentence, sentence 10 was recorded. Thus, the Rater Training consisted of actual audio files for rating practice. To complete the Rater Training, raters were given a printed page with the participant/sentence codes and the ASR scores and were instructed to review, compare and write in their ratings. Ratings were subsequently reviewed with the researcher. Any questions regarding the ASR scores, sentence audio samples or the individual rater’s concerns regarding their ratings were answered and discussed with the raters prior to the rating session. On average, the rater training lasted approximately 15 to 20 minutes.

Following the training the raters were instructed to enter the Rater Site and to rate the audio files, to ask questions if necessary regarding any aspect of the Rater Site, audio files or rating scale. Raters had access to the rating scale for consultation during the rating period. After completing the rating, raters were asked to enter the Re-Rate Sentences site and re-rate ten participant samples. At the end of the session, raters were asked to answer
five questions relating to their rating experience. The entire time period for training, rating, re-rating and answering five questions was approximately an hour and a half.

The quality of the sound files produced for the pilot studies was not considered important. Nonetheless, it was determined that the quality and clarity of the audio recording was significantly reduced when students used a headset with an external microphone. The audio recorded with a headset (earphones with an attached microphone) is far superior to the external microphone set-up. External microphones available at the lab do not filter extraneous noise and can interfere with the ASR recordings, an observation that highlights the critical reason for using headsets with attached microphones for the study. Because the microphone is inserted or connected to the sound card located on the computer, both the quality of the microphone and sound card can affect the quality of the audio recording.

The lab computers where the software was installed were Pentium IIIIs (Windows 2000) with limited speed and storage and consequently it was expected that the quality of the audio would be compromised. However, the use of headsets by the participants helped minimize these detrimental effects.

The procedures described above and the instruments tested in the pilot study (April 2008) were repeated and used for the current study. Specific modifications included: 1) fourteen separate participants for the Rater Training, including sentence sound files for practice rating; 2) thirty French student participants producing ten sentences, selected from the software corpus, for a total of 300 audio files for rating; 3) expanding and further developing the Rater Site to accommodate the larger number of participants and audio files; 4) collecting ASR scores for ten sentences from the three
raters; and 5) a pre-chosen random sample of ten participant sentences and creation of a
Re-Rate Sentences site used for collecting intra-rater data for reliability calculations.

Participant Survey

Qualitative measures (Patton, 2002) included a descriptive online participant survey designed to elicit responses relative to: the ASR-recording experience; repetition and pronunciation of the sentences; ASR-scores; listening to native models; use of the waveform and pitch curve visuals and listening to self-produced speech. The university-based survey tool was used to design the survey. Participants accessed the survey using a link provided at the end of the sentence recording session. The participant’s randomly assigned number identified and linked the qualitative responses to the ASR scores and ratings for the same participant.

Several online surveys have been conducted (2005, 2006, 2007) to discover different facets of French student software users. The online survey used for this study (Appendix D) is an adaptation and modification of earlier piloted surveys. The qualitative survey questions were further developed for a distinct focus on the participant experience using the ASR software for recording French sentences. As well, the survey questions were designed to elicit the participant’s perceptions of using ASR software for practicing their French pronunciation. The descriptive survey components were designed to: 1) collect information regarding student perceptions, and experience, of the ASR software features; 2) to provide deeper insight into the participant’s experience; and 3) to investigate the ways in which the participant descriptions triangulate with the ASR scores they obtained for the sentence pronunciations.
Data Collection and Analysis

Statistical Measures - ASR Scores and Rater Data

The ASR scores were analyzed using descriptive statistics to summarize the basic characteristics of the scores based on the participant group. Variability in ASR scores was captured through calculating the standard deviation and the typical ASR score was assessed by calculating a group average ASR score. Participant-level mean scores were calculated as well. The same statistical procedures were conducted for human rater scores. Next correlation coefficients were used to examine the consistency or relationship between ASR scores and scores recorded by human raters. All descriptive statistics were calculated using SAS v9.1.3. to further examine the ASR scores and to summarize the basic characteristics of the scores based on the participant group. Standard deviation was calculated, as well as, a calculation of mean scores. Correlation calculations were used to calculate the consistency of the rater’s rating as compared to the ASR scores.

Inter-Rater Reliability

Inter-rater reliability was calculated using SAS and was used to examine the consistency and reliability of the human ratings across raters for the recorded sentence samples. Correlation calculations were used to calculate the inter-rater consistency.

Intra-Rater Reliability

The ratings from the Rater Site (first scoring) and the Re-Rate Sentence ratings (second scoring) were checked for intra- , or within-rater reliability. Due to the small
sample size, only ten scores were collected and compared, correlation coefficients were calculated, but were not expected to be significant. The rater training, questions addressed during the training review, researcher observations and the rater responses to the five questions provide insights into the rating experience. These insights are discussed in the analysis, as well as under directions for future research, and may help to explain the first and second score differences.

**ASR Score Validity**

Depending on the outcomes of the rater scores and following a determination of inter-rater and intra-rater reliability, it may be possible to consider the ASR scores as valid, where it can be demonstrated that the ASR scores are consistent with the human raters (personal communication, White, 2008). A case for ASR score validity as a measure of pronunciation performance and practice will be discussed.

**Qualitative Data - Participant Survey**

The online survey completed and collected from the participants at the end of the data collection session was saved in an electronic format for data organization. The online survey provides for data export features and the ability to systematically organize survey data. Based on the topics and questions included in the survey (Appendix D) content coding of questionnaire responses (Brown, 2001; Dornyei, 2003; Strauss and
Corbin, 2008) and further data analysis (Miles and Huberman, 1994) proceeded as outlined below.

Survey Data – Coding and Analysis

Specific attention was given to coding responses, as related to content features investigated in the research questions. The survey questions addressed participant use and perceptions of the features related to the ASR software. Response formats provided for both closed and open-ended responses (Dornyei, 2003). The closed questions included participant data and demographics and naturally were more limiting and in many cases did not allow for participant comments.

Initially, all descriptive survey responses were grouped and transferred verbatim to an electronic document (Brown, 2001). Responses were openly coded (Strauss and Corbin, 2008) by question, for emergent patterns and categories. Content-themes, relative to individual student responses, were categorized as they relate to the ASR features. Codes were assigned to categories. Specific attention was given to axial coding (Strauss et al.) of responses as related to subsequent emergent themes investigated in the research questions.

Miles and Huberman (1994) outline three iterative and interactive processes for qualitative analysis: data reduction, display, and triangulation for verification and drawing conclusions or identifying patterns. For data reduction, coding was also used for confirmation of observed themes within and across survey responses. A data display matrix, in the form of an adapted Checklist Matrix was used to systematically summarize
the “clustered” coded responses (Patton, 2002). The matrix included coded content, informative emergent patterns and response themes used for analysis (see Appendix E).

A readily available word count software tool and the word count feature in the qualitative software Atlas/ti were used, in some cases, to cross-check word codes, categories and themes. Brown (2001) used a similar approach in a study where a survey was developed and used to investigate language researchers. Both high and low word counts were used successfully for cross-checking and analysis. The actual word count tool Brown used in his study is no longer available.

Miles and Huberman (1994) have suggested both within-case and cross-case display and these displays were applicable for a description and explanation of data for both individual participant survey responses as compared with the individual ASR-scores. Additionally, where applicable, the human rater’s ratings were included for comparison and explanation of individual survey responses.

Triangulation, or the use of multiple data collection methods was used to cross-check and examine multiple data sources including participant ASR scores, perceptions, rater scores given for the same participant audio sentences, and observations. Discussions in the form of “researcher bias checks”, with other research colleagues, proved helpful and provided further insights for the analysis.
Chapter Four: Results

Introduction

Several types of data were collected for the study in order to answer the research questions. As the study applied a mixed method, two distinct data analysis processes were applied.

A quantitative analysis was conducted to answer Research question 1: When using the same automatic speech recognition (ASR) produced sentences for human expert rating, how do human ratings compare to the ASR scores provided to students and how valid are the scores? The audio sentence data was collected and both the ASR scores and the subsequent rater scores were analyzed using correlation. As well, the inter- and intra-rater reliability indices were calculated. The score results are discussed and address questions of validity, given that no validity evidence exists relative to the ASR-generated scores.

A qualitative analysis was conducted and a ten question survey questionnaire was used to collect data regarding student perceptions for Research question 2: How useful did the students find the automatic speech recognition (ASR) features were for assisting them with their French pronunciation practice? Several survey questions addressed the utility of the software for assisting with pronunciation practice. Other survey questionnaire items were designed to elicit descriptive participant data that was collected for Research question 3: How effective did French students perceive the automatic speech
recognition (ASR) software to be for improving their mastery of French? Responses were evaluated to determine whether participants perceived the software could be effective for improving their mastery of French.

In order to conduct the quantitative analysis of ASR and rater scores, the score collection and analysis organization process was conducted in three stages. First, the ASR scores for each participant (P) were extracted from the software and inserted into an Excel file. The Figure 1 screen shot is an example of the software interface view of the ten French sentence audio recordings and ASR scores. In this view, at the top of the screen, beside “Student's audio recordings” the P25 indicates that this is participant 25. Under “Recording” and “Sentence Pronunciation” a speaker icon indicates that these are sentence recordings. Beside each icon we see a French sentence, a date and time (called a date-time stamp) and under “Result” an ASR number score out of 7 total.

**Figure 1.** P25 Sentence 1-10 ASR Scores
Stage 2 involved extracting the rater scores for each sentence from the rater website where they were recorded and inserting the rater scores into the same Excel file. The Figure 2 screen shot illustrates the rater website and screen view of the rater scores. In this screen, beside numbers 1, 2, 3 (delete) with date and time, identifies the rater (to the researcher only). The “speaker icon” with the right pointed arrow indicates that the audio could be played in this view (by the researcher). The code “p16s1” indicates that these are rater scores for participant 16 sentence 1 (s1), s2 and s3. For example, “rater 1” assigned p16s1 = 5 (= score), p16s2 = 5, and p16 s3 = 6.

![Figure 2. Three Rater Scores for P16 Sentence 1 (s1), s2 and s3](image)

In stage 3 the ASR scores and rater scores for each participant were organized and input to an Excel spreadsheet for quantitative analysis. Figure 3 illustrates the Excel spreadsheet organization where “id” indicates that score data for participant 16 through 25 is displayed with scores reported for sentences 1 through 3, asr, raters a, b and c. For example, P16 s1asr = 7, s1ra = 6, s1rb = 5 and s1rc = 5.
In each of the three stages used for extracting, organizing and then analyzing the quantitative score data, cross-checking of scores was done at least three times by the researcher, with assistance from two colleagues.

For the qualitative analysis, the survey questionnaire was organized by participant survey to begin initial coding. Each participant survey was saved and printed. The survey tool used to create the questionnaire allowed for a print-out of responses, according to survey question. The training participant surveys (P1-P14) were reviewed and initial codes were assigned. All surveys, that is the 14 surveys collected from the training participant's and the thirty study participant surveys (= 44 surveys) were collected, reviewed and coded.

Subsequently, a separate electronic Excel file was created for each survey question where only the thirty study participant responses were organized and collated. Codes and categories were assigned and final analysis was conducted based on a question by question organization of survey response data. Figure 4 is an illustration of the electronic file created by collating the thirty participant responses for survey question 6.
As illustrated in Figure 4, the initial survey questions that included biographical data such as participant number, gender, French level, years of study were summarized under each question in order to identify each set of descriptive data by participant.

The bio data collected in the first two survey questions defines the participant sample and indicated (summarized in Table 5) that the thirty participants were predominantly female (male = 8), beginning French students who had been studying for a year or less at the university level. Survey questions 3 and 4 requested experiential information related to using computers and new software, specifically speech recognition software (summarized in Table 6). The reported experiences revealed participants were comfortable using computers, however they had not previously used automatic speech recognition software (ASR).
Figure 4. Collated Participant Responses for Survey Question Six

16m I 1-2 s I liked the software very much. I hope to use the software to help me enunciate the French words properly. The time frame for each sentence seemed so short for me, but that helped me understand the proper flow of each sentence.

17f I > 1 s I enjoyed it because it allowed me to listen to a native speaker and right after try to pronounce the same words on my own. I believe if I worked on it longer, my pronunciation would subtly improve.

18m I 1-2 yr I though my pronunciation was awful.

19f II 1-2 s Frustrated!!! but I like that you can go back and repeat it until you get it right. I felt that the sentences were a little hard for me. They spoke so fast and I felt like my tongue couldn't keep up with theirs.

20f II 1-2 yr It was interesting, I had a hard time following what the people where saying when they read aloud the sentences. I found it really fast.

21m I 1-2 s Software was very good and actually being able to see the sound waves helped tremendously. I could hear the "ups and downs" of the language, what had more emphasis, and what was not really pronounced. Great help!

22f I > 1 s I like the software. It really helped having someone say the sentence first and then repeating it. I also liked how it instantly showed your progress to the side. It helped me to compare how my tries were going.

23m I 1-2 yr I thought my pronunciation was awful. I really liked this software. It was easy to use.

24f other >1s I was kind of nervous because I have not had french for many years and I know that my pronunciation is not the best. It was weird to see that my words were so off from that of a native speaker.

25f other >1 s I wish I could do this more often, it helps to learn the correct pronunciations, the way the pitch of our voice is recorded and the correct way is recorded above it is very helpful and a really cool idea!!

26m I 1-2 yr It was very helpful and useful.

27m other >1 s Not very comfortable sometimes

28f other 1-2 yr I was nervous since other people were in the room and could hear me, but I thought it was fun and helped me alot.

29f I 1-2 s I felt like it was very neat and a good learning tool. A few of the sentences moved a little fast though so I felt like I needed to speak just as fast which made my words slurr together some.

30f I 1-2 s I enjoyed using the software but I felt that the time to get accustomed to the sentence was very short.

31f I 1-2 yr I enjoyed. I was a little intimidated at first but I eventually got past that once everyone else started talking.

32m I 1-2 yr nervious at first, then got used to it.

33m I 1-2 s Oddly, I did feel slightly uncomfortable using this because I realized just how bad my pronunciation is. I could see this being beneficial and enabling me to be more comfortable speaking with native french speakers.

At some points, when the words were unfamiliar to me, I felt the recordings didn't allow enough time for me to get the full recording in. My mouth just didn't want to move that fast.

34f other <2yr I didn't know how loud I needed to speak in order for my voice to be understood, so I made the assumption it would only need to be slightly over a whisper so that I would not be heard by the few other individuals in the room.

Once I realized it was not working very well, and I started to get more comfortable, I was engaged and excited to speak as loud as needed without considering others around me.

This session was very useful because I realized a lot of what is holding me back is my intimidation, but that it is possible for me to speak. I need to realize I read much more than I speak, thus my speaking is naturally going to suffer.

There were a few words within the sentences that I knew the meaning to but had never read aloud. When I read it in my head I thought it would have been pronounced in a certain way, but was proven wrong by the native speaker.

35f other 1-2 yr It was a good experience. I was impressed with the software.

36f I 1-2 s I thought it was fun and it really helped my with pronunciation.

37f I <2 yr It was interesting, I wish I had more time to practice with it more.

38f other >1 s It was easy and simple enough. Since we were using headphones, I forgot that there were others who might hear me, so I wasn't too anxious.

39f I 1-2 s I felt that the software has done its purpose but it was a bit too fast for me.

40f II 1-2 s I was a little frustrated with how poorly I was doing but eager to experiment with ways the software could potentially help my French pronunciation.

41f other >1 s It was alright, it will be better to get feedback and learn what I need to work on in my pronunciation. The red highlighting helped as I was going through because I knew what I needed to focus on.

42f I 1-2 s It was a new experience that I have not done before. I enjoyed being able to record my voice and understand how different I sound compared to a real french speaker.

I am excited to further my education in the subject and work on my sentence pronunciation.

43f I 1-2 yr easy to understand

44f I 1-2 s easy to use and I would like to use it again.
The organization of survey questionnaire data was critical for coding of categories and themes by hand and iteratively. It is important to note that the data coding, categorizing, journaling, researcher reflections, frequency and word counts were done by hand. Both a freely available word count tool (Bonkenc) and ATLASi were used experimentally with a few survey questions to cross check word counts with assigned codes and categories. However using the word count tools proved redundant and time-consuming, particularly once the majority of the coding and analysis had been done by hand.

Organizing the data, both ASR scores and qualitative survey data, was a complex process because the different types of data needed to be carefully extracted from various electronic sources. Using Excel files to organize both types of data, that is score data and survey data, provided a consistent and complete score and survey data set for each participant. Setting up the data and creating a database in this way provided for individual case observation and analysis, future combined case analysis and a myriad of other post hoc studies.

Analysis

Research Question 1: When using the same ASR-produced sentences for human expert rating, how do human ratings compare to the ASR scores provided to students and how valid are the scores?

To answer the first part of the question, that is how do human ratings compare to the ASR scores provided to students, the ASR scores produced by the ASR software for the participant’s sentence recordings and the rater assigned scores for the same recordings
were subjected to a number of statistical tests. Descriptive statistics were used to examine the ASR score and rater a (ra), rater b (rb) and rater c (rc) score distributions. Correlation analysis was used to determine the relationship between ASR scores and rater scores. Inter- and intra-rater reliability was also calculated using correlation.

In order to examine the second part of the research question, *how valid are the scores* and following from the statistical analysis of the ASR scores and rater scores, the validity of the ASR scores are discussed. As there has been no known analysis of the ASR scores produced by the software, prior to this study, any validity statements are made solely on the basis of the findings presented.

**Quantitative Analysis: ASR Scores and Rater Scores**

*Descriptive Statistics*

For the purposes of descriptive data analysis, each set of scores - ASR scores for 30 participants = 300 ASR scores and three sets of rater scores (ra = 300, rb = 300 and rc = 300) are considered separate variables, resulting in four variables: ASR, ra, rb, rc. SAS was used to examine the distributions of each variable, individually, prior to examining the relationship between the variables. The descriptive statistics for the four variables can be found in Table 1.
<table>
<thead>
<tr>
<th>Statistics</th>
<th>ASR scores</th>
<th>ra scores</th>
<th>rb scores</th>
<th>rc scores</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>N</strong></td>
<td>300</td>
<td>296</td>
<td>299</td>
<td>298</td>
</tr>
<tr>
<td>Mean</td>
<td>4.45</td>
<td>4.46</td>
<td>4.47</td>
<td>5.74</td>
</tr>
<tr>
<td><strong>SD</strong></td>
<td>1.44</td>
<td>1.14</td>
<td>1.09</td>
<td>1.95</td>
</tr>
<tr>
<td>Skewness</td>
<td>-0.13</td>
<td>-0.003</td>
<td>-0.14</td>
<td>-1.46</td>
</tr>
<tr>
<td>Kurtosis</td>
<td>-0.38</td>
<td>-0.25</td>
<td>0.43</td>
<td>3.52</td>
</tr>
<tr>
<td>Range</td>
<td>7.0</td>
<td>6.0</td>
<td>6.0</td>
<td>6.0</td>
</tr>
</tbody>
</table>

*Table 1 Descriptive Statistics for ASR, ra, rb and rc Scores*

*N explained.* The slight reduction in the *N* (*N* = 300 total scores) for *ra* = 300-296 = 4, *rb* = 300 - 299 = 1 and *rc* = 300 – 298 = 2, is the result of several audio files not opening during the rater’s scoring period. Although all files were checked, double-checked, and were working audio files, it appeared that the files were not able to open for technical reasons associated with the wireless network system. Missing data is recorded as follows: P19s4ra, s4rb, s4rc; P25s9ra; P37s3rc; P44s8ra, s9ra (*N* = 7 missing scores).
Mean (M) and extreme scores. An average score for each set of scores is commonly reported in Second Language (SL) studies (Mackey and Gass, 2005), and describes the typical score of the set of scores. The M, as an average, can be misinterpreted if there are extreme scores reported. There were a small number of low/extreme scores, considered low where ASR score = 0 or 1, reported for a few specific audio files that had been inadequately recorded due to poor quality of the specific computer sound card. The extreme reported ASR scores are listed for participant and sentence as follows: P19s4asr = 0; P20s9asr = 1; P27s2asr = 1, s4asr = 1, s8asr = 1; P31s4asr = 1. In one case, the rater scores were also low as a result: P27s2asr = 1, s2rc = 1; s4asr = 1, s4rb = 1 and rc = 1. In a few cases, although the ASR score < 3, rater scores were reported as extreme. The participant, sentence and ASR score, as well as the extreme rater scores are as follows: P25s6rc = 1; P34s2asr = 5, ra, rb and rc = 1; s6asr = 6, ra and rc = 1; P35s6asr = 5, rc = 1. Although statistically these scores are considered extreme, they were not removed from the analysis. A total of 15 extreme scores are reported.

It is important to note however that the extreme scores in these cases were the result of several factors. In some cases the ASR score was a direct result of the poor recording and functioning of the individual computer station. Sentence recordings were on average less than 3 seconds in duration. Any delay, hesitation or interruption, whether created by the computer, or microphone, or the participant interacting with the features of the ASR software, were difficult to identify. The audio sentence recordings were contributed as part of the participant data set of ten sentences, where each sentence

71
resulted in an ASR score and separate audio file. In order to maintain the integrity of the participant data set, for rating by the raters, no extreme scores were removed.

On several occasions, during the rater scoring sessions, an audio file would not play. The only explanation was that there was a temporary interruption in the wireless network. All files were checked, re-checked and accessible. Also, the temporarily unavailable audio file was different for each rater, except for P19s4 where audio data was missing entirely (P19 may have not recorded or may have skipped or missed sentence 4). Again, given the unexplained variability relating to the low/extreme scores it was decided to include all scores as reported and collected.

*ASR and rater score distributions.* The mean (M) for the ASR scores, ra and rb scores are close. However, M = 5.74 for rc is significantly higher. Likewise for rater c, S = 1.95, and skewness, sk = -1.46. Clearly, the score distribution for rater c is non-normal and indicates that rc, on average, rated the participant sentences higher than ra or rb. On average, rc rated participant sentences 1.29 points higher than the ASR and 1.28 and 1.27 points higher on the rating scale (1 = low, 7 = high score) than ra and rb respectively.

The kurtosis value for rater c, k = 3.52 indicates a peaked, very leptokurtic, non-normal distribution. The highly peaked distribution indicates rater c scored within a very narrow point range and significantly violates the underlying normality assumption. It is questionable whether rater c scores are reliable measures.
Important considerations: rater C. Rater c was the third rater, a native speaker (NS) who had not participated in the pilot study whereas both rater a and rater b had rated previously (April 2008). Rater c received the same rater training as the other raters, however it is uncertain whether having rated previously gave rater a and rater b an advantage. It is possible that ra and rb may have been more experienced raters, in the ASR context. Clearly, ra and rb had the experience of “practice rating” (although they did not rate the same participant audio) several months earlier.

Variability in the scores is most reliably demonstrated using the Standard Deviation (S) and is a reliable index of variability. However, it is important to examine the S and the means in relation to each other in order to determine how the scores are spread. A small S, in relation to the means, demonstrates, generally, that the mean has captured the scoring behavior of the ASR and raters. For example, rb M = 4.47, where S = 1.09, demonstrates a proportionally small spread. Rater c distribution reported a larger spread, S = 1.95

The shape of the distribution of each set of scores, as indexed by the skewness and kurtosis for each variable, are an important contribution toward an understanding of the degree of normality displayed by the distribution. The skewness values indicate that the ASR, ra and rb distributions are slightly negatively skewed, where the scores are concentrated at the higher end of the rating scale. Rater c (sk = -1.469) indicates a high negative skewness value.
The range of scores, based on the ASR scores assigned and the rubric developed and used by the raters to evaluate the sentences was designed on a $1 = \text{low}$ though $7 = \text{high}$ rating scale. The ASR reported a range $\text{ASR} = 7$. The range for the rater $a$, $b$ and $c = 6$.

In a discussion of statistical analyses and language tests Brown (2001) suggests that violations of the normality assumption is not as problematic for criterion-referenced measures. Rater $c$'s skewed distribution might indicate, for example, that participants pronounced the sentences successfully. Also, Brown (2001) suggests reporting the mean and median for skewed distributions because a few skewed scores (representing a few $P$ sentences) can significantly affect the mean, but has less effect on the median (middle score $= 50\%$ above, $50\%$ below or midway between two middle scores). Violations to the assumption of normality that underlie correlation are obviously problematic for rater $c$, in particular. A comparison of the variable distributions, that is ASR scores and $ra$, $rb$ and $rc$ scores, was used as the basis for the correlation and other measures.

Bachman (2005) suggests that a common misunderstanding is to interpret the means of score groups as suggesting that there is a relationship between the scores and he suggests that researchers often incorrectly interpret findings. The description of the distributions is necessary for a correlation analysis and statements about the strength and degree of relationships can only be made subsequent to conducting a correlation.

Correlation Analysis

In correlation research attempts are made to determine the relationship both between and among variables. Usually, it is practical to use scatter plots in order to
graphically observe the degree of the relationship between the variables. Scatterplots of the ASR scores with each rater's scores were not visually helpful, as no relationships were established. However, correlations were calculated for the four variables - ASR, ra, rb, re scores - in order to examine the relationship between ASR and rater scores. The correlation coefficients for the ASR scores and rater a, b and c scores are presented in Table 2.

**Relationship between ASR scores and rater scores.** Results report positive, statistically significant, yet low moderate relationships between variables. Rater a (ra) is a non-native, female French literature professor responsible for French teaching assistants (TAs) and graduate students. Rater a had been oriented to rating ASR-produced audio during the pilot and although not entirely comfortable with technology, felt comfortable and unhindered by the rating required for the study. As a non-native speaker rater a has a command of French pronunciation and considers pronunciation enhancement and practice a necessary element of teaching and learning for non-native TAs and her students.

The relationship between ASR scores and rater b = rb, indicate the strongest relationship, where \( r = 0.452 \). Rater b had scored ASR-produced audio during the pilot and had been previously exposed to the scoring procedures and rating scales which lends some support for the stronger relationship observed with the ASR scores.
Table 2 Correlation Coefficients: ASR and Raters a, b and c Scores

<table>
<thead>
<tr>
<th>Correlations</th>
<th>ASR</th>
<th>Rater a</th>
<th>Rater b</th>
<th>Rater c</th>
</tr>
</thead>
<tbody>
<tr>
<td>ASR</td>
<td>( r = .99 )</td>
<td>( r = .36 )</td>
<td>( r = .45 )</td>
<td>( r = .30 )</td>
</tr>
<tr>
<td></td>
<td>( p &lt; .0001 )</td>
<td>( p &lt; .0001 )</td>
<td>( p &lt; .2027 )</td>
<td>( p &lt; .0001 )</td>
</tr>
<tr>
<td>Rater a</td>
<td>( r = .36 )</td>
<td>( r = .42 )</td>
<td>( r = .57 )</td>
<td>( r = .40 )</td>
</tr>
<tr>
<td></td>
<td>( p &lt; .0001 )</td>
<td>( p &lt; .2260 )</td>
<td>( p &lt; .0001 )</td>
<td>( p &lt; .0001 )</td>
</tr>
<tr>
<td>Rater b</td>
<td>( r = .45 )</td>
<td>( r = .57 )</td>
<td>( r = .71 )</td>
<td>( r = .49 )</td>
</tr>
<tr>
<td></td>
<td>( p &lt; .2027 )</td>
<td>( p &lt; .0001 )</td>
<td>( p &lt; .0193 )</td>
<td>( p &lt; .0001 )</td>
</tr>
<tr>
<td>Rater c</td>
<td>( r = .30 )</td>
<td>( r = .40 )</td>
<td>( r = .49 )</td>
<td>( r = .46 )</td>
</tr>
<tr>
<td></td>
<td>( p &lt; .0001 )</td>
<td>( p &lt; .0001 )</td>
<td>( p &lt; .0001 )</td>
<td>( p &lt; .0001 )</td>
</tr>
</tbody>
</table>

Rater b (rb) is a native French speaker with ten years teaching experience in the U.S. and has worked with French second language learners and homogeneous groups of
American English university classes. As well, rb has worked with University-trained French student-teachers in local, American high schools. Another factor that could have contributed to the relationship found with ASR and rb scores is the fact that of the three raters, rb was the most highly trained technology user. Rater b was also the most comfortable with technology and had even recently taught a graduate level FL technology course.

Rater c (rc) was also a native speaker and a younger female graduate student who had limited experience teaching in the L2 environment. Rater c had also not rated ASR during the pilot and questioned the rating scale prior to rating for the study. All three raters were aware that they were using a scale to rate French sentences, student produced audio that had been previously scored by the ASR. Raters did not have access to the ASR-produced scores, except during the training using separate participants (P1-P14).

**Rater reliability.** Critical to the analysis of ASR scores, is the fact that, to date, a relationship between ASR scores and any other type of FL scoring has not been established. In this investigation ASR scores were examined for French sentence pronunciations. Raters a, b, and c scored exactly the same participants and their French sentence audio pronunciation files. In order to establish rater reliability, an examination of consistency between (inter-rater), and within individual (intra-rater), raters a, b, and c was necessary.

**Inter-rater reliability.** Using correlations as indicators of the degree of consistency between the three raters: rater a = ra, rater b = rb and rater c =rc, Table 2
indicates the correlation coefficients calculated. The total $N = 300$ French sentence pronunciations, that is 30 participants x 10 sentences = 300 audio files for rating, by each rater. The correlation between raters a, b and c are listed in Table 2.

From the correlations between raters, all of the correlations were positive and the relationship between rater a and rater b was the strongest, where $r = .57$. Both ra and rb had rated French audio files in the pilot study. It is reasonable to assume that their previous experience rating may explain the observation that their scoring patterns, although only moderately related, were the strongest of the three.

Previous experience rating similar ASR audio (ra and rb) was more highly correlated with the raters than was being a native French speaker (rb and rc). Based on the correlation coefficient between asr scores and rb scores, $r = .45$, rb correlated more highly with ASR than the two raters, ra and rc, correlated with each other.

Rater c and rater a had the least strong relationship, where $r = .4$. It is important to note that these two raters were the most dissimilar. Rater c had not rated before and was a native French speaker, while rater a was a non-native speaker and academic, who had previously rated.

**Intra-rater reliability.** In order to check the consistency of each individual rater, that is the consistency with which a rater scored the same sentence pronunciations from the initial rating period (time = 1) to a later time period (time = 2), ten random participant sentences were selected for re-rating.
The $N=10$ is an insufficient number for reliable statistical coefficient measures. However, it is clear from rater b results for time 1 and time 2, $r = .718$ that rater b rated the most reliably when compared to ra and rc. The intra-rater reliability for rb, even given a small $N$, lends support for rater b's training, experience and expertise when rating French audio sentences in the context delineated in the study.

The ten random sentences were accessible through the *Re-Rate Site* and were re-scored following the initial rating session. Note from Table 3 rater c initially had problems opening one audio file for Time = 1 and thus $N = 9$. However, during the re-score when the rater entered the *Re-Rate Site* the file was able to be opened and re-scored. Again, the explanation for this discrepancy is that there was interference from the wireless network in the library, during the scoring time period. Interestingly, this was not the case for rater a and b who both scored several days prior to rater c when all of the files were able to be accessed accordingly. Thus for ra and rb, $N = 10$ consistently.

The small $N = 10$ for the intra-rater reliability correlations is a drawback to actually finding consistency within individual raters. There were several reasons such a small number of audio files were chosen and input to the *Re-Rate Site*. First, raters were asked to attend a pre-scoring training session. During the training session, raters scored approximately 14 “training” participants. Several training participants included all ten sentences, and some contained only one sentence for a total of 80 “training” audio files. The rater training participant set ($P = P1$ through $P14$) were not part of the data set of 30 participants ($P = 30$, consecutively $P16$ through $P45$).

Second, the rater training included a review of the scores the individual rater assigned to the “training” participants and a comparison to the ASR scores. Questions the
rater's had regarding individual “training” participants, sentences, audio files or scoring were discussed during the training. The training process lasted approximately a half hour.

Once the “training” scores had been reviewed with the rater, by the researcher, the raters entered the *Rater Site* and began rating the 30 participants (300 audio sentences). Each individual rating period lasted approximately 1-1½ hours.

Third, once the rating period ended, the raters were asked to enter the *Re-Rate Site* and to re-rate ten sentences. By this time the raters had been rating for a considerable time. As well, the ten sentences were repetitive and the raters had begun to experience an overload. Cognitive overload is common in foreign language speaking rating and results from several factors both individual and situational: rater concentration and fatigue, physical discomfort, and repetitive factors. Clearly, based on researcher observation, the raters had become fatigued at this point.

The ten participant sentences were re-rated by rater a, b and c and correlation coefficients were calculated. The results for time 1 and time 2 for each ra, rb and rc are presented in Table 3.
Table 3  Intra-Rater Reliability Correlation Coefficients: Time = 1 and Time = 2 for Rater a, b, and c

<table>
<thead>
<tr>
<th>Time 1 and Time 2</th>
<th>N</th>
<th>r</th>
<th>p</th>
</tr>
</thead>
<tbody>
<tr>
<td>ra1 and ra2</td>
<td>10</td>
<td>r = .42069</td>
<td>p = 0.2260</td>
</tr>
<tr>
<td>rb1 and rb2</td>
<td>10</td>
<td>r = .71808</td>
<td>p = 0.0193</td>
</tr>
<tr>
<td>rc1 and rc2</td>
<td>9</td>
<td>r = .46916</td>
<td>p = 0.2027</td>
</tr>
</tbody>
</table>

All of the correlations conducted for the quantitative analysis are summarized in Table 11 (p. 112). The individual tables 1-3 report the correlations of each variable grouping (ASR scores – rater scores, inter- and intra-rater reliability). The summary table was created in order to further the discussion relative to: the validity of the ASR scores as generated by the software; consideration of the ASR scores and rater assigned scores as valid and as practical indicators of pronunciation ability; and inter- and intra-rater considerations.

Survey Questionnaire

In order to obtain information for answering research questions 2 and 3, data and descriptive responses were elicited through the use of a survey questionnaire. The survey
was linked to the participant sentence recordings (the same participant number was used) and was administered online, immediately following the sentence recording period. While the complete questionnaire results were analyzed, and provided valuable data, not all of the survey questions were directly applicable to research questions 2 and 3.

**Research question 2:** How useful did the students find the ASR features were for assisting them with their French pronunciation practice? Survey responses were examined and analyzed as they related to student perceptions of the utility of the software for assisting with pronunciation practice.

**Research question 3:** How effective did French students perceive the ASR software to be for improving their mastery of French? Descriptive participant data was collected to evaluate whether participants perceived the software could be effective for improving student mastery of French.

All of the survey question results are presented and reported. However where the survey item results relate directly to the research questions, the survey results are organized under and support the respective research question. The complete survey questionnaire is attached in Appendix D and is available for reference.

**Results - Survey Questionnaire**

The survey questionnaire consisted of ten questions and the organization of the survey questions was important for eliciting specific experiential qualities and for triangulation to other responses. A number of questions consisted of several parts. Survey questions 1 and 2 were used to collect bio data. They were closed questions and inquired
about gender and length of French study. The frequencies for gender and duration of French study are outlined in Table 4.

Survey question 2 also gave participants the opportunity to elaborate or “explain further” about their French study experiences. Twenty-one of the thirty participants felt it important enough to explain further about their experiences. The responses were related to high school French study where nine participants related experiences of studying French in high school and explained gaps in their study and learning periods. For example, the explanation provided by one participant was a common remark and includes: “This is actually my first semester taking a french class since 10th grade in High School” (P23). Three participants commented that they had spent time in or had traveled to France. One participant commented: “i grew up in canada where I was taught french from 4th grade on” (P20). Finally, participants added comments specifically about their past and future French learning and the following comments reflect the responses for participant 36 and 38: “I studied french in high school but do not remember a thing”(P36); “... rarely got to practice the speaking aspect during high school and haven't taken a French course for two years on top of that” (P38).

Survey questions 3 and 4 inquired about comfort using technology and experience using language learning or speech recognition software, respectively. The responses were coded for yes = comfortable, no = uncomfortable and other = sometimes and comments on comfort level other than yes or no. For question 4 answers were coded for yes = have used software before, no = have not used before and Conditions = comments beyond yes or no regarding conditions for learning. Frequencies and conditions for Q3 and Q4 are listed in Table 5.
Table 4  Frequencies for Gender and Duration of French Study

<table>
<thead>
<tr>
<th>Survey Responses</th>
<th>Frequency</th>
</tr>
</thead>
<tbody>
<tr>
<td>Male</td>
<td>8</td>
</tr>
<tr>
<td>Female</td>
<td>22</td>
</tr>
<tr>
<td>French Study</td>
<td></td>
</tr>
<tr>
<td>French I</td>
<td>15</td>
</tr>
<tr>
<td>French II</td>
<td>8</td>
</tr>
<tr>
<td>Other/Unspecified</td>
<td>7</td>
</tr>
<tr>
<td>Semester/Year</td>
<td></td>
</tr>
<tr>
<td>&gt;1 semester</td>
<td>6</td>
</tr>
<tr>
<td>1-2 semesters</td>
<td>12</td>
</tr>
<tr>
<td>1-2 years</td>
<td>8</td>
</tr>
<tr>
<td>&lt; 2 years</td>
<td>4</td>
</tr>
</tbody>
</table>

Survey questions 5 and 6 were designed as a “warm-up” to encourage participants to reflect on past French learning experiences. The intended purpose of survey Q5 was to have participants describe how they perceive their affective responses to learning situations and to relate their interactions. This question was designed to allow participants
to express both positive and negative learning experiences, in the hope that they would more freely relate to survey Q6 which followed.

*Table 5  Frequency Responses for Survey Q 3 and Q 4*

<table>
<thead>
<tr>
<th>Frequency Responses</th>
<th>Yes</th>
<th>No</th>
<th>Conditions</th>
</tr>
</thead>
<tbody>
<tr>
<td>Survey Q 3 and Q 4</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Q 3 - Comfortable using</td>
<td>23</td>
<td>1</td>
<td>Yes, if or as long as = 4</td>
</tr>
<tr>
<td>computers/new software?</td>
<td></td>
<td></td>
<td>Sometimes = 2</td>
</tr>
<tr>
<td>Q 4 - Used language learning</td>
<td>7</td>
<td>23</td>
<td>Online Spanish = 1</td>
</tr>
<tr>
<td>speech recognition software</td>
<td></td>
<td></td>
<td>Other software = 3</td>
</tr>
<tr>
<td>before?</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Survey question 6 specifically asked participants to describe, affectively, the short experience (in many cases 15 minutes or less) of recording their pronunciation of French sentences using the automatic speech recognition (ASR) software. This question was designed to capture the overall quality of the participant experience and to provide a means for triangulating to the individual experience with the specific features of the software and the pronunciation recording period.

The results for survey Q6 revealed positive user experiences. Initial coding resulted in two basic categories: 1) comments relating to the software and 2) how the
features (for example the native speaker) or a feature of the software related to the individual. For example, the following quote by one participant demonstrates the 'positive experience' theme: “I enjoyed it because it allowed me to listen to a native speaker and right after try to pronounce the same words on my own.” (P17). Another participant commented: “I enjoyed being able to record my own voice and understand how different I sound compared to a real French speaker.” (P43). The same theme was expressed about the graphic display by another participant: “Software was very good and actually being able to see the sound waves helped tremendously. I could hear the ‘ups and downs’ of the language, what had more emphasis, and what was not really pronounced.” (P21).

On a continuum of general 'quality of experience' for Q6, from exciting to frustrating, twenty-seven participants related that they found the software: fun, exciting, interesting, helpful, enjoyable. Of the twenty-seven participants, seven were coded more toward the middle of the continuum where participants reported feeling initially uncertain or nervous, yet were positive overall about the experience. Participant 28’s response was characteristic: “I was nervous since other people were in the room and could hear me, but I thought it was fun and helped me a lot.” (P28). Three participants were coded at the frustrating end of the continuum and indicated they were not comfortable, for example participant 31 reflected simply “Not very comfortable sometimes.” (P31).

The continuum of 'quality of experience' theme related directly to the participants' impressions of the software. So, where the experience was positive, the software was perceived as “easy to understand, or easy to use”, and the features such as, listening to the native speaker, being able to repeat and to say each word separately were viewed as helpful. Where participants had expressed frustration with the experience, the overriding
comment was that the speed of the speech and pronunciation of the native speaker was very fast and as one participant described: “They spoke so fast and I felt like my tongue couldn't keep up with theirs.” (P19).

Survey Q6 was an open-ended question, the first question that requested a description by the participant relating to the ASR experience. Only following the analysis of later questions (especially Q7 and Q8) did it become clear that participants provided their initial descriptions and impressions of the software in Q6. After coding open question Q6 (and Q7 (d)), it became clear participants provided complete descriptive responses about their impressions and experience in Q6. This is understandable as the focus and wording of each question was similar, yet Q6 could be considered a “first request” for an open, individualized evaluation of the experience. Regardless, participant responses, comments and descriptions were consistent with the individual's described experience.

**Automatic Speech Recognition (ASR) Features**

*Research question 2: How useful did the students find the ASR features were for assisting them with their French pronunciation practice?* Survey questions 7 and 8 were designed to address participant perceptions regarding the specific features of the ASR software. Responses were examined and analyzed as they related to perceived usefulness of the software for assisting with pronunciation practice.

Survey question 7 consisted of four parts directly relating to the features of the ASR software. Data responses for three parts were quantitatively-oriented and the final part was an open-ended question. Survey Q7 was organized as follows: Based on the
French pronunciations that you recorded today did you find it helpful: a) To listen to your own pronunciations? b) To listen to the native speaker model? c) To look at the visual graphs? d) Please tell us more about your impressions and your experience with these software features.

Table 6  Question 7 (a, b, c) Frequency Responses

<table>
<thead>
<tr>
<th>Helpful</th>
<th>Yes</th>
<th>%</th>
<th>No</th>
<th>%</th>
</tr>
</thead>
<tbody>
<tr>
<td>a) to listen to your own pronunciations?</td>
<td>27</td>
<td>90%</td>
<td>3</td>
<td>10%</td>
</tr>
<tr>
<td>b) to listen to the native speaker model?</td>
<td>30</td>
<td>100%</td>
<td>0</td>
<td>0%</td>
</tr>
<tr>
<td>c) to look at the visual graphs?</td>
<td>23</td>
<td>76%</td>
<td>7</td>
<td>24%</td>
</tr>
</tbody>
</table>

88
For survey Q7 (a, b, c) frequency responses were coded yes = helpful, and no = not helpful for 'perceptions of usefulness'. The pronunciation features of the ASR software examined were: 1) listening to self, 2) listening to the NS model and 3) looking at the visual graphs. The frequency responses for each feature are listed in Table 6.

Native speaker model. Participants responded positively to the native speaker model presented for each of the ten sentences (Helpful = 30 participants) where all participants reported it was helpful to have the NS model. The native speaker (NS) model is based on a standard French speaker from the northern, Parisian region of France and referred to as Parisian French (Valdman, 1976). The native models used in the ASR software are representative of the French spoken in that region of France and include males and females, young and older adult speakers. Although the participants could read each French sentence presented to them, on the computer screen, they felt that being able to listen to the sentence pronounced by the NS speaker, any number of times, either before or after recording, was very helpful. Although there may have been participants hesitant about pronouncing the sentences, seeing the words on screen and listening to the pronunciation by the native speaker contributed to their confidence and ability to orally repeat the sentence pronunciation.

Listening to self. Participants reported that listening to their own pronunciations was helpful. However, it is unclear whether they listened to a pronunciation and then reproduced the sentence based on the NS and their own pronunciation or based on the "score" they received. In other words, when participants listened to their own recordings
were they attempting to improve the sentence pronunciation or their ASR score? Or did they see both their sentence pronunciation and the ASR score as the same? For example, one participant commented: “I wish that it would save your recordings so you can shut down and come back in later to listen to yourself.” (P37).

**Visual graphs.** The visual graphs were also reported as being helpful however there were participants who clearly felt they were not as helpful and were perhaps even confusing. The visual graphs are presented as a waveform and a pitch curve in the ASR software and a screen shot was put into the survey for clarity. The visual graphs are sophisticated representations of sound, and researchers (Chun, 1998; 2002) have commented that where they are used in software for language learning, students need to be trained to understand and interpret their meaning. Otherwise, it has been suggested they not be used.

Interestingly, one of Chun's (1998) early suggestions for research was just this topic - train some students using visual graphs for speech synthesis and then leave others to interpret on their own, an excellent suggestion for a research study. However, the TeLL me More (TMM) software options do not allow “turning off” (or making unavailable) the visual graphs for some and not others. It is considered a global option, meaning it is an option that occurs throughout the program. Nonetheless, Chun's work (1998; 2002) is an important contribution to FL speech synthesis and CALL.

Participants were impacted by the visual graphs and in the following case examples did understand and interpret the graphs for themselves: “I really liked the ability to go back and listen to what I can improve on, the fact that the proper way of
speaking the sentence can be accessed, and the graphical display to show the stressed part of the word(s)” (P16); and “The graphs were extremely helpful, it would tell you where or when you need to fix your pronunciation of a word or phrase, loved it!!” (P25); or “The visual graphs were an important part in pronouncing the words right” (P17).

Of the participants who responded “no” the visual graphs did not help (P = 7), only two commented specifically about the visual graphs: “The program would give me a great score, but the visual graph looked totally different than the native speaker” (P32); “The visual graphs were not helpful...I felt like they were tracking voice inflection, not syllable accuracy” (P40).

More impressions about the ASR experience. The open-ended feature of survey question 7 (d) asked participants to “tell us more about your impressions and your experience with these software features”. Participants responded by describing more about their interactive experiences with the features of the ASR software. Comments from participants 28, 39 and 44 related specifically to the NS model and how participants felt it helped them, for example, “I liked the native speakers were the voices heard on the software. It really helps hearing the correct accent.” (P39); “I really enjoyed seeing how I sounded compared to the model.” (P44); and “It is very fun and challenging and helped me pronounce words I didn't know how to pronounce before” (P28). Participants were able to see and hear the French sounds and make phonetic connections for their own French pronunciation.

Although all of the participants reported that they found it helpful to listen to the native speaker model, comments in Q7 part (d) related to the NS model. Participants
reported they felt the NS spoke too fast and suggested they would have liked the speaker to slow down, or to speak slower. Participants 20 and 27 gave definitive responses: “great and they speak too fast” (P27) or “It is interesting, but they need to speak a little slower” (P20). The NS model was perceived to be helpful, maybe even necessary, and participants suggested how they perceived it could be more helpful for their French pronunciation learning.

Pronunciation practice - initial impressions. Given the short time span for recording, participants responded positively to the features of the software and their interaction with the ASR. Although not all commented specifically on the features, comments were directed to how they felt the software could help them practice or improve. “It’s good to help us practice and listen to how we speak” (P26) and “It was amazing! I didn't realize how off I was when I spoke” (P23). Participants 33 and 37 respectively, commented on ways to improve or change the software to be more helpful to them, for example, “I wish that it would save your recordings so you can shut down and come back in later to listen to yourself.” (P37) and “I would have liked the speaker to slow down, or perhaps have an option where you can click on individual words for pronunciation” (P33). Interestingly, the software does provide this feature option. However, it was not an included option for the study sentences.
ASR Scores - Perceptions of Usefulness

Survey question 8 consisted of several parts addressing the ASR “score” box with green bars. Survey Q8 was organized as follows: Tell me More ASR provides a score (represented by 'green bars') for each of your sentence pronunciations: a) Were the ASR scores helpful for your pronunciation? b) Do you think the ASR scores reflect your French pronunciation ability? c) We are most interested in the ASR score as it is presented to you. Could you please comment on exactly what this score meant to you? The first two parts of Q8 were closed (yes/no) questions. The frequency and percentages for both parts (8a and b) are presented in Table 7.

Emergent Themes

Survey question 8(c) was open-ended and asked participants to comment on exactly what the score meant. A screen shot of the box marked 'Score' was also presented as a visual aid for reference, next to the question. Three significant themes emerged from the analysis of 8(c) and included three separate meanings participants assigned to, or associated with the score feature: 1) visual, 2) pronunciation and 3) score meanings.
Table 7  Frequency and Percentages for Survey Question 8(a) and (b)

<table>
<thead>
<tr>
<th>Frequencies for Survey Question 8 a and b</th>
<th>Yes</th>
<th>%</th>
<th>No</th>
<th>%</th>
</tr>
</thead>
<tbody>
<tr>
<td>a) Were the ASR scores helpful?</td>
<td>28</td>
<td>93</td>
<td>2</td>
<td>7</td>
</tr>
<tr>
<td>b) Do you think the ASR scores reflect your ability?</td>
<td>25</td>
<td>83</td>
<td>5</td>
<td>17</td>
</tr>
</tbody>
</table>

**Visual Meaning**

The visual aid presented with the question was assigned a meaning and represented a common reference. A visual marked with the name score was assigned an evaluative value and was considered some form of evaluation. Participants used terms such as: bars, score chart, scale and guideline. The colors gray and green were also mentioned, where green was equated with good. In the ASR software screen the score is visually represented by small, rectangular and vertically positioned bars. Figure 5 is a screenshot of the “Score” box, with green bars, in the ASR software and as seen on the Speech Recognition activity screen.
The positioning of the score box and the green bars that appear within the box (following a recording) are important from a visual graphics standpoint. The score box uses color, rectangular “bars” and vertical movement on the screen to create the appearance of addition, more or better. Comments included: “It felt good when I looked and I had five bars, since the last time it was in the gray which was not very motivating” (42) and “I never wanted to get a gray color bar, green meant good to me” (P44).

Figure 5. ASR Software “Score” Box

Pronunciation Meaning

In terms of pronunciation, participants referenced features of pronunciation and associated meaning to specific qualities of pronunciation. Two distinct and important features of pronunciation are speed and accuracy and related features such as timing and clarity. One participant commented: “This score showed me that I need to work on my speed and accuracy” (P39); another commented “The scores made me realize that some sentences were easier to pronounce that others” (P17). One participant (French II)
commented on the vocabulary used for the sentence pronunciations: “I was unfamiliar with the vocabulary and would get tongue-tied by the words I had never seen before” (P40). These participant comments indicated a sophisticated understanding of the qualities necessary for interpreting pronunciation meaning.

It is appropriate that speed of pronunciation was identified and triangulates to previous comments referenced earlier (Q7) regarding the “fast” NS speaker model. Also, in survey Q8, the NS was a reference point for comprehensibility (Munro & Derwing, 1997) or understanding of French pronunciation. Participant 32 interpreted the score to be a comparison to the NS. For example, “A score above the bar was acceptable. Meaning that a native French speaker would understand what I said. Maybe not perfectly, but understandable” (P32). The perception expressed (P32) indicated that the participant had linked their pronunciation performance to being “comprehensible” to a NS.

**Score Meaning**

The word 'Score' on the top of the “score box” was the most concrete example of associations with grades and participants quantified and qualified the score with meanings. The familiar dichotomies associated with evaluation were apparent and several categories emerged such as: good/bad, pass/fail, high/low, easier/harder. As well, participants interpreted a higher score (or more green bars) as better, or more acceptable.

More important, perhaps, was the fact that participants interpreted the score to indicate progress or ‘need improvement’, and an indication to work or to try harder. This self-evaluative view of the score meaning took the form of statements about how they viewed their speaking skill, or where they saw themselves, “It meant I need a lot of help...
speaking” (P28), or their goals, for example, “It gave me a better idea of how well I spoke certain sentences. I enjoyed knowing where I was at and where I should be” (P31); and “It showed me that I need to work on certain areas of pronunciation and some areas I do well on” (P43).

It is apparent from the analysis of survey Q8 that the feature of the ASR software entitled “Score” impacts the students' feelings about their pronunciation. It appears from the responses that the green bars appearing and indicating a positive result, or pronunciation, can spontaneously result in a good feeling and provided motivation to continue recording the same sentence or to move on to the following sentence. Participants want to know how well they were producing, pronouncing and speaking the language and how they could work to improve. Even in a short period of interaction, participant responses described their perceptions indicating that the ASR software was a tool that did and would help them practice their French pronunciation. Participant 23’s comment was particularly instructive and informative: “It actually made me want to work to get the bars higher. It was amazing how much it affected me to want to get a little green bar to rise. I think that this tool will make student work harder because it actually shows you where you are and where you could be” (P23).

Participant responses and the subsequent analysis demonstrate conclusively that their perceptions of the usefulness of the software for assisting with pronunciation practice emerged. Not only did participants perceive the software to be useful for pronunciation practice, they also reported a perception that the software engaged their motivation and stimulated an interest, particularly for their current FL learning and for future practice. Given the elements of positive feedback, motivation and interest
expressed by participants in response to their brief use of the software, it appears imperative to consider how and in what ways the software can be made available to students for their use with their FL pronunciation and speaking skills.

Perceptions of Effectiveness of the ASR Software

Research question 3: How effective did French students perceive the ASR software to be for improving their mastery of French? Descriptive participant data was collected to evaluate whether participants perceived the software could be effective for improving student mastery of French. Participant responses to survey questions 9 and 10 were analyzed and responses were examined for participant perceptions of the effectiveness of the ASR software for improving French speaking skill. To some extent, evidence of participant 'perceptions of effectiveness' has been documented in earlier survey findings. However the range of the final questions allowed for a broader and deeper analysis of responses by providing the opportunity for comparison and triangulation to earlier participant descriptions and reports.

Survey question 9 consisted of two parts and was organized as follows: How helpful is the ASR software feedback (scores, native model, visuals) compared to other feedback you have received on your French pronunciation and speaking? Could you please comment about how this software's feedback has been more or less helpful to you? The first part was a quantified statement of degree, and asked for comments on how helpful the ASR feedback was compared to other feedback. The second part was a request for a descriptive comment regarding the specific feature of the ASR software feedback that was or was not helpful. The findings for Q9, 'ASR compared to other feedback,' are
listed in Table 8a. Although not a specific part of Q9, a second (Table 8b), was designed in order to compare the “perceived helpfulness” according to length of time studying French.

Table 8a  Survey Q 9 - ASR Compared to Other Feedback

<table>
<thead>
<tr>
<th>Q 9 Responses - helpfulness of ASR compared to other feedback</th>
<th>N = 30</th>
<th>%</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. Much more helpful</td>
<td>14</td>
<td>47%</td>
</tr>
<tr>
<td>2. Somewhat more helpful</td>
<td>13</td>
<td>43%</td>
</tr>
<tr>
<td>3. About the same</td>
<td>3</td>
<td>10%</td>
</tr>
<tr>
<td>4. Not as helpful</td>
<td>0</td>
<td>0%</td>
</tr>
</tbody>
</table>

The rationale for the comparison of 'Perceived Helpfulness according to length of French study' (Table 8b) is based on the supposition that longer length of study would result in more experiences with feedback, or a broader variety or type of feedback. Students who have studied French longer may have more feedback experiences available for comparison which raises several interesting questions. Does the ASR software
feedback benefit beginning learners more than students with more or longer learning experiences? Or are students who have studied longer more experienced and more qualified to compare their feedback experiences?

The literature provides some support for the use of ASR software with beginning learners (Hincks, 2003). However, the general lack of FL pronunciation and speaking practice, particularly within a classroom setting (Eskenazi, 1999), is a common and ongoing concern in the field. One participant commented that they felt the teacher had limited time for interaction.

In survey Q9 the helpfulness of the ASR compared to other feedback resulted in two feedback types or themes: 1) feedback from a person and 2) feedback from the ASR software. The comparisons resulted in the ASR software always being compared to a person, or evaluated based on a person giving feedback. The person listed as giving feedback was the professor, French teacher, 'real' teacher and native speaker. Other qualifiers such as in-person, one-on-one, someone and 'your voice' were used to signify a human giving feedback vs a machine or as one participant commented a 'sound device'. This finding triangulates with the concept of human raters being the “gold standard” (Eskenazi, 2009) for FL pronunciation and speaking evaluation, particularly for students within a classroom. Students expect the human, teacher-rater to be available and expect that they are qualified to assist them, if only during class periods.
Table 8b  “Perceived Helpfulness” According to Length of French Study

<table>
<thead>
<tr>
<th>Q9 Responses</th>
<th>&gt;1s</th>
<th>1-2s</th>
<th>1-2yr</th>
<th>&lt;2yr</th>
<th>Totals</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. Much more helpful</td>
<td>2</td>
<td>5</td>
<td>4</td>
<td>3</td>
<td>14</td>
</tr>
<tr>
<td>2. Somewhat more helpful</td>
<td>3</td>
<td>5</td>
<td>4</td>
<td>1</td>
<td>13</td>
</tr>
<tr>
<td>3. About the same</td>
<td>1</td>
<td>2</td>
<td>0</td>
<td>0</td>
<td>3</td>
</tr>
<tr>
<td>4. Not as helpful</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
</tbody>
</table>

The feedback from the ASR software was described based on comparative advantages of using a computer and headset for pronunciation recording and practice. These advantages included it being easier to 'click' and faster feedback. Specific advantages related to the ASR software included several components such as: features that can compare more aspects of pronunciation, red highlighting as indicators for sounds or words that need attention, ability to pronounce unfamiliar words when presented by the NS, as well as, the ability to reproduce pronounced words more clearly.
Pronunciation practice, or the ability to repeat individual pronunciations, was seen as a comparative advantage relative to a teacher. One can listen over and over, repeat, and can go back and hear yourself and these were qualities of the ASR software that were seen as helpful and were valued. “One on one feedback doesn't allow you to compare as many aspects of pronunciation, also, its easier to just click a button and have the phrase repeated to you than it is to have your professor repeat it over and over” (P25).

Being able to repeat French sentence pronunciations was associated with not only practice but improvement, as well. Participants 31 and 32 stated they felt being able to repeat and practice meant they could work harder and the work would result in a better outcome or pronunciation. “The play back of what I recorded against a native speaker helps me see what words I need improvement on” (P32). “I think it is more helpful because it allows you to listen over and over constantly try to better improve you better score” (P31).

The concept of practice, reported descriptively as being able to repeat pronunciations, was seen as an advantage. One participant creatively equated the ability to record and listen with “speaking the language”: “It is nice to be able to go back and hear myself speak the language” (P30). Another participant felt that the software was unbiased “The feedback is unbiased to me, which I value” (P16), and another was perhaps overly trusting of the software! “It actually shows you where you are. This software can tell you exactly where you are and it isn't just the opinion of a person. This is actually physical proof of how you are speaking and how you should be speaking. It's incredible.” (P23).
Of the thirteen participants who viewed the software as “somewhat more helpful” the speed of the NS pronunciations was the focus of concern and the participants saw this as a setback for their pronunciation. For example, “It depended on the sentence, there were times where the sentence was said to fast, so I had a hard time making out a word” (P20). There were three participants who rated the software as similar to or about the same as other forms of feedback. Three rated the software “about the same” where two of these three comments contained reference to a person: “in person you get a lot more feedback and more quickly” (P42) and another commented, “I think it does almost as well as a teacher since it is basically the same concept as listening to and repeating after her. The advantage though is I can practice more with this than I can with the teacher” (P38). The concept of using the ASR software for practice, despite uncertainty about the comparison to other forms of feedback in the above cited cases, was seen as better for pronunciation practice.

Overall, twenty-seven of the thirty participants reported that they perceived the ASR was more helpful than a teacher rater for providing feedback. The analysis indicated that participants were qualified to differentiate between the relative advantages of the various forms of rater, teacher-based feedback based on their experiences. Where a FL teacher-rater has provided input and feedback to students, they valued and trusted the teacher and at the same time employed their technological savvy to interpret the contribution of technologically-enhanced features such as ASR. The participant responses provide insight into, and support for, students’ ability to perceive and distinguish variable feedback methods and still assign value to their own learning experiences.
Survey question 10 addressed the issue of teacher ratings of French pronunciation and speech and was organized as follows: *Could you describe a time when you had your French speaking or pronunciation rated or evaluated by your French teacher? a) How do you feel about having your French speech or pronunciation rated by an expert human rater (either known or unknown to you)? b) Would you want the rater to be a native French speaker?* The responses and assigned codes represent the most commonly used form of FL evaluation for speaking and pronunciation. The coded responses are presented in Table 9.

The responses to teacher-rated speaking or pronunciation were indicative of the expectations students have of teachers, generally. Students expect teachers to rate FL speaking and to correct their pronunciation. Not only do students want to be rated, they want to be corrected and given feedback on how to improve. Participants expressed that, although aware of limited class time for learning, they view the teacher as ‘rater’ each class and especially when they are asked, during class, to speak or pronounce words or sentences. Although this is a common FL classroom practice for pronunciation exercises, students did not report feeling uncomfortable and want to be corrected. “During class if we pronounce a word wrong our teacher will correct us. After I try to repeat the word correctly” (P32). In another example, a participant spoke of having a conversation with the teacher: “Well we had a conversation in French and she would correct my pronunciation as I went. It was helpful but a little distracting from the conversation” (P19).
Table 9  Survey Question 10 - Coded Responses for French Teacher Rating

<table>
<thead>
<tr>
<th>Coded Responses</th>
<th>Informal Comfortable</th>
<th>Formal Uncomfortable</th>
</tr>
</thead>
<tbody>
<tr>
<td>Teacher-rated experiences</td>
<td>Informal = French classes Projects and presentations</td>
<td>Formal = Exams Oral interviews</td>
</tr>
<tr>
<td>a) Expert rater</td>
<td>Comfortable = 25</td>
<td>Uncomfortable = 5</td>
</tr>
<tr>
<td>b) Native Speaker rater</td>
<td>Yes = 27</td>
<td>No = 3</td>
</tr>
</tbody>
</table>

Students expect to be evaluated and they want and accept specific feedback, both on informal and formal speech ratings. In one case a participant reported that the teacher was responsive to pronunciation at the level of syllables: “They pointed out a couple of instances where I had pronounced a final syllable. That is usually the only commentary I get in oral interviews” (P34) and “I was pretty relaxed because the instructor helped me through any hard pronunciations” (P30).

Originally, the oral proficiency interview (OPI) was developed by ACTFL to evaluate FL conversational ability. Modified and adapted versions of oral interviews have been incorporated into FL curricula as a way to quickly evaluate spoken language. The oral interview students referred to in response to Q10, and especially at the beginner
level, is relatively short - less than five minutes. A French I student reported that they had not been rated orally yet: “I don't believe my teacher has listened to just me yet” (P38).

In reference to survey Q10 where participants were asked to comment on their feelings about having an expert rater rate their sentences, the twenty-five participants were comfortable with the idea. The responses that were indifferent (“It doesn't bother me” (P20)) were also coded = comfortable, and in some cases participants stated they were glad not to know the person. Only five were coded as being uncomfortable: “It freaks me out a little” (P29) and “I would feel anxious” (P30) and one participant commented they were more concerned about their performance on the sentences than about the expert rater: “It made me a little self-conscious but I soon forgot about it” (P36).

The responses were positive regarding a native speaker rater for rating their sentences. Twenty-seven of thirty participants reported that would want the rater to be a native speaker. The NS model is used by the ASR and participants had the model available during the recording period. Participants perceived the functionality of the ASR to be a comparison of their pronunciation to the model where the NS pronunciation was the best “score”. Three participants responded “no” that is, these three would not want the rater to be NS. However they were not uncomfortable with an expert rater.

While coding survey question 10 several interesting questions about FL rating emerged. For example, how are FL rating and scoring related, for students? It appears that students (or at least these participants) are less concerned about an expert rater and had very little affective response (indifference, don't care, don't mind, doesn't bother me etc.) but had a much more definitive response to a NS rater. It might be possible to
suggest that a NS is an expert, yet do students see the NS and expert as the same person? What is the relationship between rating and scoring? Does the question of how these two evaluations are related have a basis in how students are impacted, that is how it affects them? These questions arose when looking at the “no” responses to 10 (a) and (b) and, although rhetorical, may provide a basis for future analysis.

An interesting aspect of survey question 10 was the reported positive impact FL teachers have had on the participants. There was a clear understanding and acceptance of teacher feedback, especially when referring to correction. Participants referred to the teacher being 'not too strict' or 'more lenient' on pronunciation. One comment sums up the sentiment: “I would like to know what I can do to sound comfortable speaking conversational French in a relaxed atmosphere” (P45).

Final Survey Comments

At the end of the survey, a final request for “comments” on the experience at the lab or the ASR software resulted in 20 comments (10 = no answers). The responses were grouped according to comments relative to the experience: 1) at the lab and 2) responses about the ASR software. The lab experience was described as: enjoyable, fun, a good or great experience. Three comments included participants being distracted by others present while recording in the lab. However, in each of the three cases, the participants stated clearly that the overall experience was positive.

As regards the ASR software, participant final comments were positive and related that the software was easy to use. Participants suggested they would like to see the software integrated into the FL curriculum, into their language classes and expressed the
hope they could use the software again. One comment was particularly informative and summarizes one participant's overall experience: “I wish the sentences had been recorded slower. I also wish there was a way for the software to give more specific feedback on a particular recording...I do think the software has the potential to be very useful in the future, but at its present state, I would hope that this program would not be used as a method of evaluation for a course grade. Just a useful tool for practice!” (P40).

The results of the survey questionnaire meaningfully contributed to the participant experience recording French sentences using ASR software. The participants’ provided positive suggestions for their learning and useful insights into the features of the ASR software. Their interpretations of what this means for their French learning included how and in what ways they perceived the ASR could effectively contribute to their mastery of French.

Through a triangulation of methods, itemized survey questions and a final analysis of the qualitative findings, participants provided consistent support for their views and perceptions concerning the actual and potential use of ASR to contribute to their French learning. Student perceptions of the usefulness of the ASR software for assisting with pronunciation and speaking practice, as reported and analyzed, are meaningful and contributory qualitative ‘evidence’ for the usefulness of ASR software.

The insights that emerged in the qualitative findings suggest, definitively, that participants perceive the ASR software as effective for their mastery of French. There was a clear sense that participants perceived their participation in the study as a “confirming statement” of their support for the use of ASR software for future FL
courses. In a final comment, one participant expressed this view: “I think that this program has good potential and it should be integrated in French classrooms” (P44).

*Researcher Bias*

Researchers are aware of their relationship to their data, whether participants or numbers. Empirical researchers seek to have the data ‘speak’ in numbers, qualitative researchers to examine data in words. As a researcher my relationship with the software began over a six week period the summer of 2004 at the University of Perpignan (south of France). During several afternoon sessions in the multimedia center I was introduced to the TeLL me More software and the ASR features. As an FL researcher my impression of the software, its ability to produce and repeat speech, was a fascinating adventure. Upon returning to the US, I learned the FL community was unaware of the tool. Barbara Lafford’s (2004) review was a refreshing contribution as she too perceived similar qualities in the software and its potential to assist students.

A positive initial experience with the software, leading to research and final reporting of results, given these were the experiences of the researcher, may have contributed to researcher bias. However, knowing the impact the experiences with the software had imparted, an awareness of, and sensitivity to, the potential for bias was consciously present during analysis. An effort was made to maintain a researcher’s distance from the data. In this way the nature of the participant experiences were reported ‘at a distance’ by the researcher, especially because the researcher’s experiences may have been similar to the participants’ experiences using the software. Feedback from
colleagues and further reflection suggests that the expressed perceptions of the participant voices have been analyzed and reported.

*Other Emergent Survey Data and Researcher Reflections*

During the early coding of all the surveys several interesting concepts and ideas emerged and while they are not directly applicable to the issues addressed by the research questions there are important connections that were identified. Three of these connections will be outlined, as they are perceived to be supportive of specific topics related to 1) FL interaction and speaking, 2) student learning experiences, and 3) FL teaching.

*FL interaction and speaking.* Following the initial coding, the first run-through of all the surveys, and after arriving at the final survey question 10, the process and duration of time emerged as a significant contribution to the concept of ‘interaction’ (see coding example Appendix E). Participants had used the word “during” to qualify and quantify an experience, either an earlier learning experience or the experience with the software. The word “during” implies process, movement, start/end, motion, that is a process of something happening in time, while time is passing. The question that emerged was: “How does this concept of time as having a flow, a movement, as a beginning and ending describe a student's view or perception of FL speaking? Or does it?

There were very few mentions of computers or software. It appeared that students might be expressing the concept of FL speaking as 'during interaction as process'. Students describe teacher input and expect, desire and want teacher feedback. Participants then describe their responses where they correct themselves, take in the feedback, plan or
produce a spoken thought or message. In this view of interaction as process, interaction occurs with the student, teachers or materials, language, and the action, end, outcome or situation is produced by the student.

**FL learning and speaking.** This concept of time, in FL terminology, is considered an interaction process. Students expressed the distinct feeling that they accept, for the most part, the process, that is they expect to interact and to experience an outcome. Even if they are not happy with their input (often the case with FL pronunciation), they expect to gain, to improve, to learn something about themselves.

What emerged was a clear perception that students want to learn to speak, to pronounce, to learn to use the language. Is it possible they conceive of speaking as using the language? Are speaking and ‘using’ the language, synonymous, in their learning experiences? This idea might explain why computers and software are both important and irrelevant to the process. Important because they are a feature of the interaction process and may be an important one for pronunciation learning; irrelevant because students ultimately feel responsible, towards themselves, for their production.

**FL teaching and speaking.** Teachers were, overwhelmingly, seen as the source of input for FL speaking and pronunciation. Students viewed their teachers as resources, contributors and evaluators. Students are generally uncomfortable in speaking situations and often this relates directly to their feeling of a lack of control over what they want to produce or need to produce in the language. This is especially difficult in the area of FL speaking when students want to do well and desire to be prepared. Pronunciation and
speaking are areas where a teacher can have considerable influence on the students’ level of comfort, even in perceived high stakes situations. The ability of teachers to accept a student's discomfort or fear, whether the fear is about not being understood or not being heard, appears to override the teacher's actual ability to produce fluent, native or near-native language. Again, the emerged theme - that it is the student's expectation of the self that drives their learning. This theme was expressed repeatedly in comments such as: “to hear and be able to respond”, “to say what is meant or what one is thinking”, “to not offend a native speaker”, “to have conversations”, and finally “to have a good accent”.

Numerous ideas and researcher reflections emerged throughout the entire coding, categorizing, analyzing and survey questionnaire documenting process. Many appeared at first to be unrelated. However, the underlying thread, given such a short period of data collection, is that students are perceptive, they desire to learn and understand. FL pronunciation and speaking presents a considerable challenge for their learning. Yet, this is a challenge they appear willing to undertake given helpful resources, materials, and practice situations with reliable and understanding teachers.
CALL studies, given the rapid advances within the domain (advances from CD-ROM to DVD in terms of speed and storage are two simple examples) are technically challenging. Studies reported less than ten years ago may not be feasible, depending on the software design. For example, TMM Education (2003), the version used for this study was designed for a computer lab, and services networked computers from within the lab, by a server computer. The ASR software saves the produced audio recordings on the networked server. So each time a participant records or pronounces an audio file, it is saved for the 'tutor' or teacher to listen to, or score, later. However, the internet version, TMM Campus (2006) does not save audio from the ASR once the student has left the screen.

In a recent text designed for FL teachers, Blake (2008) evaluated two tutorial CALL products (LeLoup and Ponterio, 2005) that included FL vocabulary glosses for reading. Both products provide glosses in a large number of languages, yet only a limited number include audio recordings of pronunciations. As Blake notes: “...but not every language database has extensive audio recordings to accompany the word definitions (p. 56)”. Although FL audio is useful for pronunciation, especially when native examples are provided, dealing with audio files, from a technological perspective, presents distinct challenges.
Chun (1998) in her seminal article suggested research designs that although impossible at the time, would become possible given continued technological developments in speech analysis, synthesis, recognition and foreign language. Chun envisioned that more and varied research was necessary. Chun's vision and research inspired the design the TMM pilot studies and of the current study.

Overcoming technological challenges to conduct CALL research, sometimes the result of inherent software design, can be daunting or even untenable. Nowhere has this been more the case than in the area of Automatic Speech Recognition (ASR) where CALL, foreign language, natural language processing (NLP), software design and basic computer science converge.

Neri et al. (2008) and others (O'Brien, 2006) comment on the rapidly emerging interest in speech recognition technology used for computer-assisted pronunciation training (CAPT). ETS researchers are working on speech recognition software that could potentially be used for analyzing and rating speech produced by examinees in the speaking section of the TOEFL (Xi, 2008). Rodman (1999) in his excellent work: *Computer Speech Technology*, is one of the early applied linguists to bridge the divide between the branches of FL speech, phonology, phonetics and an understanding of the technology and speech science. Chapelle (2007), an early and eminent CALL researcher has suggested a more expansive view of CALL is on the horizon, one needed in order to incorporate developments in natural language processing. Chapelle comments that a broader view is necessary for understanding the type of research needed to further CALL designs, especially designs that incorporate speech features. More is being asked of CALL researchers today, than ever before, for many reasons. Automatic speech
recognition and related areas of study provide great promise for collaborative research and can encourage developments that will enhance foreign language speaking and pronunciation learning resources for our students.

Numerous CALL researchers (Eskenazi, 1999; Hincks, 2003; Barr et al., 2005; Levis & Pickering 2004) have explored the use of FL speech-interactive and ASR features used in both commercial and research-oriented products. Several researchers (Chapelle, 2001) have made meaningful contributions to our theoretical and practical understanding of how these features and products can contribute to CALL software designs and student learning with multimedia. Likewise, the limitation of current technologies and of the CALL and ASR studies have provided a solid framework for future investigations.

Mixed Method Design and Issues of Significance and Meaningfulness

The purpose of a mixed method study is to provide a broader view of a topic and to shed light on findings and the issues investigated from a variety of perspectives. Importantly, a mixed method was designed for this study to investigate a quantity and quality related to automatic speech recognition software (ASR). The quantitative findings provided meaningful information, although no statistically significant relationships were determined. The qualitative results indicated significant positive feedback from participants. Both methods and the related findings contributed to our understanding of ASR and its position within the larger domain of CALL and foreign language teaching and learning.
Bachman (2005) suggests that investigating relationships among two or more variables is “fundamental to two of the qualities of test usefulness: reliability and construct validity (p. 113)”. In this regard, using correlations as a means of analysis for ASR scores, raters' scores and relationships among these variables is practical and informative. However, correlations and tests of significance are often difficult measures to employ in foreign language research and linear relationships as representations of FL learning can be incomplete. Investigating the ASR “score” construct by comparison to the relationship of the rater's scores was an attempt to establish the validity of the undefined ASR “score” construct. Examining the reliability of the rater's scores also contributes to our understanding of the ASR “score” value and its potential or practical usefulness as an evaluation tool.

Mackey and Gass (2005) discuss the use of correlation in FL studies and comment that “in second language research we generally deal with much smaller sample sizes, making it difficult to get statistical significance (p. 267)”. FL qualitative studies, particularly longitudinal and case studies, are difficult to design and following participants, even if there are only a few, can be time intensive and costly. Mackey and Gass (2005) argue for meaningfulness in foreign language studies and suggest:

In considering the difference between meaningfulness and significance (in the statistical sense), we need to recognize that second language learning is a slow and complex process often involving a period of production of correct forms only to be followed by a later period of production of incorrect forms. Therefore we need longer periods of observation, but
the exigencies of research do not often allow long periods of time. Thus, it may be that meaningful trends are worthy of discussion, independent of statistical significance. (p.268)

A balanced FL study, such as the investigation into the assigned meaning of the ASR score, has contributed to meaningfulness, if not significance in the statistical sense, and has shed light on meaningful trends, as suggested by Mackey and Gass (2005).

Several CALL researchers (Chapelle, 2001; Weir, 2005) have suggested that research involving several methods are lacking within our domain. Where empirical analyses are conducted, complete reports of findings are neglected or are unreported because the results are considered statistically insignificant. Others (Bachman, 2005) suggest that complete reporting is necessary in order to advance our profession and responsibly assist other researchers with more complete study designs or replication. With an interest in responding to Bachman's suggestions, and a desire to appeal to a larger audience of CALL and FL researchers, a complete report of the results have been provided.

By conducting both quantitative/empirical and qualitative data collection and analyses a more complete picture has been presented of an area of CALL - multimedia with automatic speech recognition (ASR) software. ASR is often misunderstood or misinterpreted by the FL community (Davies, 2006) and clarifying or simplifying are necessary for furthering our understanding. This study presented a view which employed not only more inclusive measures for a broader investigation, but a view presented from
multiple perspectives: 1) ASR software design and multimedia; 2) FL professionals and
FL raters as evaluators; 3) and the French learners and students users. This investigation
and reported findings have contributed to a more complete understanding of ASR and its
usefulness as a tool for FL practice, performance and learning.

ASR Sentence Pronunciations and Score Validity

One of the first questions posed by FL users and teachers is: What does the ASR
“Score” mean and what do the “green bars” = score represent? Most important, what do
they indicate to the student users? However, if there is no definitive explanation, as is the
case with the ASR “Score,” then how is the score interpreted both empirically and
qualitatively? What is the interpreted meaning? As instantiations of the interpreted
meaning construct, how valid are the ASR and rater scores?

The ASR scores assigned to the French sentence pronunciations were generated
by the ASR software through a complex process using a speech recognizer and related
components, with French as the language component. The TMM French version of the
ASR software uses a core beginner-level vocabulary of approximately 2,000 words.
Through an algorithm designed to compare incoming speech with a language model, a
sound comparison-type score is generated. A simplistic explanation for the very complex
process of independent speech recognition can explain, generally, the basis for an
assigned score. The score is recorded in the software on a point scale 1 (low) through 7
(high), yet in the user interface and on the viewed screen the score is recorded as “gray
bars” (below a threshold of 2) and “green bars” for scores 3 = 3 green bars, through 7.
Thus students using the software do not actually see the number scores, only “green bars.”

As reported in chapter 4, participants viewed more green bars as better and worked to improve, or increase the number of green bars attained as they progressed through the ten sentences. It appears that the score assigned by the ASR as a number and the green bars viewed by the participants, on average, were interpreted as having similar, if not the same, meanings or values. As the qualitative findings indicated, participants expressed that the “green bars” resulted in a positive or 'need-to-improve' self-evaluation of their pronunciation depending on the number of “green bars”.

Although the ASR-generated scores cannot be considered actual “test scores”, inferences and generalizations, cornerstones of validity research, about a participant’s pronunciation have been made by both the ASR software and human raters. Reliability, or repeatability of the inferences made about scores presents problems when tests involve FL productive skills, such as speaking and pronunciation. McNamara (2006) emphasized, “scoring involves human judgment, and issues of repeatability here involve estimating the extent of agreement between judges – in other words, would the candidate get the same score next time with a different judge”? (p. 33). Further, McNamara (2006) has encouraged researchers to consider new perspectives when engaging in validation studies and comments that “input from non-measurement traditions is leading to the exploration of new insights into the limitations of such inferences (p. 27).”
Possible Confounding Variability

The correlation analysis of the ASR scores relative to the rater scores did not reveal a relationship between the ASR and rater scores. There was an undefined and indiscriminate amount of variability which does not allow, empirically, an identifiable relationship. The sources of confounded variability, despite controlled variables, are assumed to be the result of several factors. One factor includes the rubric used by the raters. The rubric was a numerical scale adapted from the ACTFL Speaking Guidelines – 1999. The guidelines for speaking are descriptive in nature and have been criticized (Tatton, 2007) for being difficult to reliably interpret, especially for evaluating pronunciation components. The adapted numerical scale used for the pilot (April 2008) was initially confusing for the raters and was subsequently revised for the study. For the scale revisions, the ILR (Interlanguage Roundtable) scale was reviewed as it was the precursor to the ACTFL guidelines and included numerical values. Generally, the ACTFL - Speaking Guidelines – 1999 have been difficult for FL raters to apply and interpret, especially when evaluating the pronunciation components of speaking skills.

Another factor contributing to score variability relates to the problems the raters encountered using the adapted scale to interpret a value for such a short sentence pronunciation sample. As well, perhaps the varied background of each rater and their rater training experiences also contributed to variations in the scores. For example one rater related they felt they were “stricter” early in their rating and became more lenient as they continued rating. Another felt the opposite and stated feeling more lenient and less experienced early on in the rating process. Examining rater scoring trends and patterns
would help identify specific problems with both the rating scale and sentence length and clarity.

The backgrounds of the raters were varied and diverse in many aspects: diverse ages, length and type of teaching experience and they had taught in various educational systems. Even though there were two native French speaker raters, their individual experiences within the US and working with French language learners were different. How much did the diverse backgrounds and experiences contribute to the “human factors”, and thus possible variability, during the rating period? The adapted numerical scale and the identified problems the raters encountered and the human rater factors were sources of variability for the study.

Neri et al. (2008) used a research-based automatic speech recognition (ASR) system for their study. Researchers modeled the ASR system after TeLL me More KIDS (Parling) and they found similar difficulties with their ASR scale and scores. The participants were 11-year old Italian native speakers learning English and raters were unable to consistently rate the young children's word pronunciations. Eventually, the ASR evaluation component was adapted for a simple “accept/reject”, as the short length of the recording and young speaker/voice variations were consistent problems in the early test studies.

Another factor that may have contributed to variability is the fact that the French ASR sentences were not presented to the participants as sentences 1, 2, 3, ...10. The software, at the beginner level, contained the corpus of 540 sentences assigned to learning units. Where the selected sentence (of the ten) came from the same unit, they could only be grouped together. If the sentence was the only one chosen from the unit
then it was presented alone (for example, sentence 4). The screenshot (Figure 6) presents
the screen as seen by the participants where the sentences are grouped accordingly. It
would be interesting to see how the grouped sentences versus the sentences presented
alone were scored, according to participant. Preliminary analysis indicates there is
variability in the ASR scores of the sentences presented alone. However, the individual
sentence pronunciation features would need further analysis as well before any practical
concerns regarding sentence groupings could be determined.

Future research into the actual ASR scores assigned for sentences 1 through 10
and across participants may indicate whether participant ASR scores improved as they
progressed through the sentence pronunciations and the ASR recording experience. An
examination of specific sentence scores could reveal more information about each
sentence and how it functioned as a French pronunciation sample. A cursory glance
revealed that several sentences presented pronunciation problems and one participant
commented that they were unfamiliar with the vocabulary. Another commented that the
sentence pronunciations were more difficult than they had originally anticipated.
Although relatively short, and few, the French sentences were common and standard
beginner level vocabulary and pronunciations.
FL speaking and pronunciation are one aspect of language skill that is generally unanalyzed separately from other skills, such as listening for example. In fact FL tests of different skills are often given together. Researchers are required to analyze test scores accordingly and a considerable degree of covariance (or overlapping of skills) in the scores can be expected. Although not a test, the ASR software provides a form of performance evaluation of pronunciation. Clearly, other aspects of FL language skills are implicated. For example, participant feedback indicated that they listened to their pronunciations and the native speaker (NS) model. Listening is an FL skill that is usually
associated with measures of comprehension. One participant mentioned having a problem with vocabulary and how the NS example allowed them to listen to (hear) how to pronounce the words in the sentence. In this case, listening to the NS assisted the participant with producing the sentence pronunciation but perhaps without comprehension of the individual meaning of the French words.

The French sentences were written on several screens for reading as participants entered the software and recorded. As well, each individual sentence was written on the ASR screen. During the pilot (2008), the question arose as to whether the sentences should be written and available for reading during the ASR recording. Consultation with French professors resulted in comments regarding the level of understanding of French pronunciation and speech by French I and II students. It was decided that not only should the written sentences be available but that they were necessary for their pronunciation production. In this case, reading, as a skill, supported pronunciation and again, perhaps comprehension.

It is important to emphasize that speaking and pronunciation practice, for which ASR seems especially well suited, is not necessarily concerned with a confounding of skills for evaluative purposes. The ability of the task to identify and selectively target the skill for practice is generally the primary consideration. ASR as a component of CALL for speaking and pronunciation task practice fills a need for both the student and the teacher. The student can practice speaking and pronunciation on their own or during a lab session without evaluation by the teacher. The teacher has a limited amount of class time with students to practice speaking or pronunciation, yet participants value the teacher's input. CALL-multimedia with embedded ASR can complement FL teachers’ lack of time.
during class for pronunciation practice where students need to practice in order to
improve their pronunciation and speaking.

Initially CALL software was designed for FL practice as a result of the nature of
technological limitations. The designs resulted in software that would allow only certain
actions, and interaction was almost impossible. With advances, multimedia software has
grown to incorporate multiple, interactive processes. In one sense, ASR software is where
early CALL software began, as a practice tool for FL pronunciation. Yet pronunciation
and speaking practice remain the most difficult skill areas to incorporate in the classroom.
For various reasons teachers lack the classroom time and sometimes the FL language
culture. At the university level, teachers are under pressure to introduce (what are often
viewed as) more difficult aspects of the FL curriculum.

FL pronunciation is also one of the most difficult skills to acquire. There is some
research to indicate that after a critical period (Krashen, 1985), before the age of fifteen,
generally the L1 or native language will continue to exert an influence on the
pronunciation of the second language. FL speaking and pronunciation practice is of value
at any age and may be more important for young adults who have had earlier FL
experience, and want to advance or improve upon their abilities. Study participants
related that they had studied French in high school or earlier, sometimes for two or three
years, but felt they had learned very little and expressed frustration that they did not learn
to speak the language. The study demonstrated that using ASR software for pronunciation
practice was perceived, even in the short recording period, as a valuable pronunciation
practice tool.
ASR and Pronunciation Performance

Foreign language performance, usually evaluated within a course or classroom setting is relatively low stakes compared to other tests, such as an oral proficiency interview (OPI). For the student who is asked to model learning and produce language - the FL evaluation experience - especially as relates to FL speaking and pronunciation, can be troubling. When a course grade and a required level of performance are expected, the stakes, for the student, can be relatively high.

FL performance as expressed in ASR scores collected on the French sentences was productively received by the participants. The majority related a positive overall experience. Many expressed insightful and self-reflective comments regarding their interest, motivation and desire to learn. One participant described in a final comment their overall experience: “It was a great experience to hear my own voice because it is something I have never done before” (P43). Another participant summed up the experience with: “Wish there were more sentences. It was a fun experience” (P37). Regardless of the ASR scores obtained, participant's expressed an interest in continuing to learn, described feeling as though they could learn and wanted to improve their French pronunciation.

Teachers and professors were cited as key in promoting participants' interest in learning a foreign language and participants accepted teacher critique, correction and feedback. One participant commented regarding the ASR feedback, “However sometimes there is nothing better than having a real teacher there with you, though this is a great substitute.”
Interestingly, Andrews (1991) in her dissertation investigated whether the organization of French language in textbooks mirrored the way in which FL learners learn to speak. Not only did she find that it did not, she also found no difference between French I and French II for speaking as measured by the production of morphological features. Students were at the same level for speaking French after one year and again after two years of study at the university level. Within the academic course framework, we need to make changes for student French learning and use, especially in the areas of pronunciation and speaking. McNamara (2006) extends this view of needed change: “We urgently need to explore ways in which assessment concepts can be better used in classrooms to benefit learning and teaching (p. 39)”.

Participant feedback on all aspects of the interactive features of the ASR software was positive and definitive. Students perceived the value of being able to use technology to assist them with their pronunciation practice. Participants understood the limitations of the ASR software for demonstrating performance. When speaking and pronunciation are the goals they know they need more resources, more time to work to improve, and continued motivation to learn. We are responsible as CALL researchers and FL teachers for contributing to our students’ resources for learning.

**FL Raters and Rating Online**

Rating foreign language speaking and pronunciation, especially recorded speech has been a time-consuming, physically demanding and often an exhausting task. When teachers or professors are rating high stakes tests, training and monitoring are necessary
to insure reliable and consistent ratings. The rubric or scale used for rating also needs to undergo testing to insure that a valid scale or measure is used.

The Center for Applied Linguistics (CAL) has developed a multimedia rater training package (CD-ROM) for French (2007) and much like the TOEFL speaking section rater training (2005) various speech samples at the required levels are provided for training. Following from suggested foreign language rater training protocols, the study involved training raters using participant ASR-produced French speech samples. Further, a separate site had been prepared that included fourteen training participants and several complete 1-10 sentence sets. Raters underwent the training, then proceeded to rate the thirty participants and afterwards they re-rated ten random participant sentences. The entire rating period was less than two hours, for a grand total of approximately 360 audio samples per rater.

In essence, and similar to the CAL rater training model, raters for the study were trained on actual samples and then rated in the same multimedia environment. When cassette tapes were the standard, rating could take hours or days. Earlier when recording was untenable, students would provide written responses, usually in the form of dictation or translation (Valette, 1965). Today raters are trained and then sent speech samples to rate (often through a secured internet link) in their own environment and on their own time. In this study the purpose of the rater training was to insure consistency and reliability among raters. No only has the study added the vital quality of time efficiency, but the rating environment prepared for each rater was consistent (for each rater). Each rater provided invaluable input to the entire rating process. The raters have expanded the
view of how and in what ways raters can significantly contribute to rating FL speaking and pronunciation in the future.

The raters expressed that the website, created using a university sponsored survey creation tool, significantly reduced the amount of time needed to rate a large number of audio files. Using the site, rater fatigue was also reduced and raters commented on the efficiency and ease of rating, compared to other forms they had experienced. The raters commented that the design and layout significantly contributed to their concentration and focus when rating. The website was designed to make the stressful job of rating easier. Raters were presented with each participant and ten sentences on one screen. The ratings were input by the rater simply “clicking” on a numbered circle after listening to a sentence. The rater could re-listen to any sentence, at any time. The screen shot (Figure 7) illustrates the rater's view of the rating screen where Participant 16 – Sentences are presented. For P16 French Sentence 1 is written and underneath the speaker icon with the right pointed arrow indicates where the rater must click the mouse in order to hear P16’s audio recording of sentence 1. The 1-7 “buttons” located under the sentence are spaces where the rater inputs their assigned score.

Raters appeared to need the print copy of the rating scale early on, but later used it for troubling or confusing recordings only. They also rarely needed to re-listen. From the feedback provided by the raters, there is support for not only consistency, but also efficiency in rating, at least from a practical point of view.

A modest empirical example of the rater’s consistency is the outcome of the inter-rater correlations for rater b, and where rater a, b and c assigned similar ratings to the participant sentences. The fact that there was an observable relationship between what
rater a, rater b and rater c were rating, signifies that the rater training, the rater site and
the rating scale worked well enough across raters to be meaningful. The intra-rater
reliability was based on a very small random sample. In retrospect, it would have been
advisable to increase the intra-rater sample. However, at the time there was concern for
the raters' time constraints, given that the training, rating and repeat rating were to occur
within the same time period.
Positive Impact and Usability

“Impact” as a concept can be traced to Bachman and Palmer’s (1996) adaptation of a framework for language test validation. Language testing and validation have been complex issues for FL and test developers. FL researchers (Bachman, 1990) felt a need to define and interpret construct validity in order to evaluate language competence. With the development of concepts of communicative competence and FL production, the tasks used for language test development and validation became even more challenging to evaluate. Bachman et al. designed a framework based on ‘test usefulness' that included aspects representative of communicative competence and ability such as authenticity,
interactiveness and impact. In the context of ‘test usefulness’, impact can be considered a quality of validity.

Chapelle (2001) has interpreted Bachman’s framework for a number of earlier studies and later applied the concept, successfully, to the design of a CALL evaluation framework. Chapelle suggests that both empirical and judgmental evaluations of CALL tasks must be evaluated for ‘usefulness’ and impact. She outlines criteria for CALL designs and suggests qualities necessary for a term she refers to as “positive impact” on learning. Where CALL multimedia includes features that are designed with sound theoretical foundations of SLA and methodological rigor, evidence of FL learning may be possible. Student response in terms of their views on designs and in what ways they perceive learning can occur are valued. These types of responses, when measured either empirically and qualitatively (Jamieson, Chapelle & Preiss, 2004), or using both methods, can be interpreted as having “positive impact” on student learning and include factors such as interest, motivation, ability to improve skills and to provide practice.

Usability, a technological term employed at various stages of software design and development (Neilsen, 2004, 2006) and as formative evaluation of a final product, defines many of the qualities of “positive impact” from a technology viewpoint. Usability refers to the quality of the design as regards users and all factors relating to users such as ease of use, intuitiveness, clarity or transparency and feedback, among others. All of the factors inherent in FL or CALL software that includes user-oriented features are, theoretically, expected to contribute to or enhance student learning or practice.

The ASR software, a feature of the multimedia TMM product, can be interpreted, based on the findings from this study, both empirical and qualitative, as meeting a test of
usefulness and usability. Despite the fact that the study was conducted on second
generation computers, the software was determined, by student participant users, to
provide a learning environment that did meet the criteria for positive impact. Students
were responsive to the software and perceived that it could be useful as a FL resource and
further that it could provide incentive and support for their French language learning,
pronunciation and speaking abilities.

Limitations and Directions for Future Research

ASR Score Validity

The main focus of the inquiry into the ASR scores generated by the software was
to investigate whether it is possible to make comparisons between rater scoring and the
ASR software scoring and thus to determine if the ASR scores are, in any way, valid
measures or statements of French sentence pronunciation. The significant problem with
the comparison is that while raters can be correlated with each other and themselves, the
ASR can only truly be correlated with the rater scoring. The ASR scores, in this study,
could not be crossed checked from time 1 to time 2 to establish a correlation for the ASR
as a measure of the internal consistency of the ASR software. In order to check this
consistency the ASR audio participant sentences (or a random selection) would have to
be re-entered and re-processed through the ASR software for re-ASR scoring.

This idea, while technically possible, was abandoned based on practical
considerations as it is unlikely that the same conditions, with the same audio files could
be re-entered for consistent measures. For example, the audio files were converted from
.ogg to mp3 format for insertion into the rater website. Any further conversion would
certainly result in reduction in the quality of the audio and could thus only create more unexplained variability.

One way used to establish a consistent and reliable comparative measure for the ASR scores was to have the expert raters produce the ten French sentences and record the ASR scores assigned to their pronunciations. Table 10 is a record of the rater scores for each of the ten sentences. While not a direct one-to-one relationship, the consistency of the expert rater ASR scores does lend credibility to the high ASR score = 7, for a near-native pronunciation.

*Table 10* Rater ASR Scores for Sentences s1- s10

<table>
<thead>
<tr>
<th>Rater ASR Scores</th>
<th>s1</th>
<th>s2</th>
<th>s3</th>
<th>s4</th>
<th>s5</th>
<th>s6</th>
<th>s7</th>
<th>s8</th>
<th>s9</th>
<th>s10</th>
</tr>
</thead>
<tbody>
<tr>
<td>ra = NNS</td>
<td>7</td>
<td>7</td>
<td>5</td>
<td>6</td>
<td>7</td>
<td>7</td>
<td>7</td>
<td>6</td>
<td>7</td>
<td>7</td>
</tr>
<tr>
<td>rb = NS</td>
<td>7</td>
<td>7</td>
<td>7</td>
<td>7</td>
<td>7</td>
<td>7</td>
<td>7</td>
<td>7</td>
<td>7</td>
<td>7</td>
</tr>
<tr>
<td>rc = NS</td>
<td>7</td>
<td>7</td>
<td>7</td>
<td>7</td>
<td>7</td>
<td>7</td>
<td>7</td>
<td>7</td>
<td>7</td>
<td>7</td>
</tr>
</tbody>
</table>

Note: NNS = Non-native speaker

Several factors are important to note relative to the ASR software and are factors that may play a part in or impact the validity of the scores, strictly from the software perspective. First, the ASR software consists, generally, of a three part system: a speech recognizer, an acoustic model and a language model. The speech recognizer's judged effectiveness is based on an error rate, when working together with the other two
separately designed features, in the case of foreign language: the acoustic and language models. In response to a question regarding error rate: “What is the speech recognition error rate for the recognizer used for the ASR in TeLL me More?” - Auralog developers responded with the following information:

The error rates that are usually used in speech recognition engines are related to native speakers speaking in their own language. The speech recognition engine that Auralog is using has an accuracy rate of 99% (error rate of 1%). However this does not reflect the reality of the language learning domain where the speaker is not a native. If a sentence that is extremely poorly pronounced is misrecognized, is it an error from the speech recognition engine (that should account in the error rate) or the consequence of the bad pronunciation? We do not have a simple error rate that would show our performance in our specific domain. We use multiple different indicators that measure the various possible problems: an average pronunciation that gets a good score, a good pronunciation that gets a poor score, etc. We try to reduce them all knowing that they are contradictory and that the final tuning is based on a good balance between all of them. (Personal communication: Auralog Developers and Tim O’Hagan, May 2009).
As described by *Auralog*, error rate and accuracy rate are relative and technical terms when applied to ASR and interpretation involves variable and complex factors. Accuracy rate is, at best, based on a combination of diverse factors: programmed recognizer algorithms, modeling, human voices and languages used for testing. Foreign languages and component language models have continually created obstacles for speech recognition researchers (Rodman, 1999). Researchers are consistently working to test and reduce the error rates and in some products the reported error rate is very low. Thus, error rate is generally a given feature of ASR software depending on numerous and complex variables. The recognizer error rate is usually reported, and determined, by the developer and distributor, in the case of commercial software products (Rodman, 1999).

The question remains, how is the ASR score generated from a participants’ sentence pronunciation to the interface screen? Context is an important factor, as the commercial software has the ASR embedded within, or as a feature of the product. Contextual factors such as type of computer, sound card efficiency, noise, headset with microphone and location are all factors affecting the recording environment simultaneously, and apparently the ASR scoring. There is a reported two second “delay” from the actual time the sentence is pronounced to the visual on-screen representation of the score. Is this the time needed for the ASR algorithm to produce an evaluation for the sentence pronunciation score? Realistically, and based on the feedback from *Auralog*, it was reasonable to hypothesize a correlation for the ASR scores of $r = .99$ within the context of the study.
In an early overview of speech-interactive CALL, Wachowicz and Scott (1999) evaluate one of Auralang's (Auralog, 1995) ASR software activity features. The researchers found that the “listen and imitate” style allowed for word boundaries and pauses in sentences and in their estimation eliminated the possibility of ASR errors. To their credit, Wachowicz et al. (1999) caution reviewers: “However, we did not test the Auralang activity with students, nor did we find teachers familiar with the software who could give an impressionistic assessment (p. 268)”.

Other important factors that can affect ASR scoring are: ambient or background noise and microphone quality. Microphone quality is such a significant factor that companies suggest the type of microphone to use or even send an actual microphone with the product. Although to some extent these features can be controlled, they still impart variability. For this reason when testing any equipment related to ASR software the same exact conditions must apply for measurement purposes (Rodman, 1999). Another factor, posited by Wachowicz and Scott (1999) that realistically and practically may make the most sense is as follows: “To put recognizer errors in perspective, we would note that even human teachers cannot guarantee 100% accurate recognition of students' utterances—particularly in noisy, crowded classrooms, as anyone with language teaching experience knows (p. 270)”.

Finally, and particularly for this study, the second generation computers and sound cards used for collecting the audio files were clearly technologically inefficient and undoubtedly contributed to cases of poor quality audio. It can be assumed that re-scoring
audio files using an already inferior system would not have added to the consistency of the ASR scores.

On the other hand, it seems logical to assume that a computer-based system, internally programmed (algorithmically), similar to the commercially produced TMM-ASR software, given the same audio file, recording circumstances and optimally reproduced conditions would produce a highly, if not completely, correlated pronunciation score. This was an assumption made for this study and was based on the ASR score support of the expert raters and the explained details relative to the ASR and related technologies.

According to a report by Balough, et al. for Harcourt Assessment (2006) on their Versant for English test, an ASR-scored speaking test, the ASR score validation process used by Harcourt employed three metrics, one of which included a correlation to human scores. The correlations reported for fluency and pronunciation (two of four sub-scores) were strong (r = .89 for both score types) and according to Harcourt, “suggest that machine ASR-generated scores for Versant for English systematically correspond to human-generated scores (p. 8)”.

In order to increase the validity analysis for evidence relating to the ASR scores, it would be necessary to also increase the statistical power by increasing both the total number of audio files produced and the number produced by each participant. For example, increasing the total participants to 50 with 20 sentences each would result in (50 x 20) = 1,000 audio samples or 20 audio samples per participant. In the Versant English Harcourt Assessment (2006) study the total sample number reported was over 1,000,
although it was not clear whether this number represented total participants, audio files, sub-scores or overall scores.

*ASR and Rater Correlations*

The correlations conducted for the study are summarized in Table 11 and are organized in ‘descending magnitude’ order. The summary indicates that the ASR could be assumed to be, for practical purposes, almost completely correlated. Rater b moderately correlated on an intra-rater correlation and in this study rater b was the most consistent compared to all other measures. While all of the correlations are significant (*p* value) the degree of relationship as evidenced by the consistency of raters’ scores and the ASR scores indicated that the raters were more consistent amongst themselves as compared to evidence of their scoring consistency relative to the ASR scores. From a quantitative perspective, given the parameters of correlation, the ASR scores and the rater scores would not appear to be related. Although statistically significant at a *p* < .0001 value, the correlations indicate that the ASR scores with the rater scores are below *r* = .50.

For this study it is important to contextualize these coefficient measures, especially because no previous measures, comparing a commercial ASR automated scoring- product - with expert raters, has appeared in the literature. However, numerous studies are in press and many more are yet to be reported. What the studies (Eskenazi, 2009; Zechnar et al. 2009) are reporting are correlations, and in some cases correlations with expert or human raters. For example, the Versant – Harcourt Assessment automated scoring and speech recognition system, using speech collected through a telephone and
directed to the ASR system, report a correlation of $r = .89$ between human raters and the automated scores. The reported correlation was arrived at through a comparison to a combination of sub-scores. Would the correlations have been less had only the individual pronunciation sub-scores been correlated? It is unclear as to the actual sample used for the reporting.

The reporting of correlations for FL studies is standard, yet how can these correlations be evaluated within the context of ASR and FL human rating? In a recent study, conducted by the Educational Testing Service (ETS) (in press), researchers were interested in testing the ASR-automated scoring system developed to evaluate the TOEFL-iBT speaking test data. ETS’s speech recognition system, SpeechRater v1.0 with automated scoring, generated scores for the TOEFL-iBT speaking practice test that were then compared to human raters. A correlation of $r = .57$ was considered acceptable and reasonable for a practice test and practice environment.

ETS had thousands of speech data samples available for testing their system and for the study. ETS has an interest in developing the automated system and providing valid test measures for the high stakes testing environment and users. ETS employs and contracts renowned and expert FL researchers who work collaboratively on developing and testing the automated scoring systems. Beginning with the release of the new TOEFL-iBT in 2005, the inclusion of the speaking test section requires ETS to train and employ expert raters for rating the speaking test samples. The expected rating score turn around of seven days - to score users - is more than acceptable given the time and expense needed to employ and train a large pool of expert raters.
If ETS researchers are reporting correlations for ASR and automated scoring of speech or speaking of $r = .57$, how will the profession interpret the correlations? Perhaps more important, for our profession, how will ASR studies be replicated? For the current study, the parameters of the correlations have been defined and are replicable. The correlations reflect the exact ASR-produced raw scores, produced by beginner students from within a lab networked commercial system, correlated or compared to expert rater raw scores, without averaging or excluding scores on any dimension. Given these considerations and these examples, the correlations for this study can be considered sufficient and acceptable for a pronunciation practice environment. Where research-based, and in-house designed, ASR systems are used, an advantage is that researchers have large amounts of readily available speech data, to both model and test systems.

The purpose of ordering the correlations (Table 11) is to demonstrate a meaningful conclusion while pointing to the current limitations of using correlation as a determinant of score relationships. The ASR scores were not compared from time 1 to time 2 either functionally or practically across participants. The rater assigned scores across participants, when compared to the ASR score evidence, demonstrated that there was no identifiable relationship. Raters, when compared to each other and themselves, were likewise found not to correlate on the ASR assigned scores, rater b presenting an exception.

Clearly, rater b was the most reliable and perhaps experienced rater across measures. Many situations where raters are used involve using only two raters where a third rater is used as an arbitrator for questionable ratings, depending on the study design. Where more than two raters are used (10 raters) the “third rater” is usually designated for
the same purpose. Given that three raters were used for the study, it might be of interest for future research to work with the current rater (ra, rb and rc) scores and identify a “third rater” arbitrator, then correlate ratings of adjusted ra and rb scores with ASR scores. Another consideration would be to train two or three “new” French expert raters (NS only or NNS only) and have them rate the 300 audio samples.

Ezkenazi (in press, 2009) in an article entitled, “An Overview of Automated (ASR) Scoring in Education” suggests that communication and the development of domain understanding among diverse groups of researchers is necessary. A clear, concise presentation of findings can assist professionals with interpreting results and designing practical applications.
However, are there other contextual factors impinging on the actual numerical correlations? Mackey and Gass (2005) argue for interpreting meaningfulness in foreign language studies where language may be in a specific developmental stage associated more with a “regressive period” rather than identifiable progressive learning stages. Even

<table>
<thead>
<tr>
<th>Magnitude-ASR and Rater Correlations</th>
<th>Pearson $r$</th>
<th>$P$ - value</th>
</tr>
</thead>
<tbody>
<tr>
<td>ASR and ASR</td>
<td>$r = .99$</td>
<td>$p &lt; .0001$</td>
</tr>
<tr>
<td>rb and rb</td>
<td>$r = .71$</td>
<td>$p &lt; .0193$</td>
</tr>
<tr>
<td>rb and ra</td>
<td>$r = .57$</td>
<td>$p &lt; .0001$</td>
</tr>
<tr>
<td>rb and rc</td>
<td>$r = .49$</td>
<td>$p &lt; .0001$</td>
</tr>
<tr>
<td>rc and rc</td>
<td>$r = .46$</td>
<td>$p &lt; .0001$</td>
</tr>
<tr>
<td>ASR and rb</td>
<td>$r = .45$</td>
<td>$p &lt; .2027$</td>
</tr>
<tr>
<td>ra and ra</td>
<td>$r = .42$</td>
<td>$p &lt; .2260$</td>
</tr>
<tr>
<td>rc and ra</td>
<td>$r = .40$</td>
<td>$p &lt; .0001$</td>
</tr>
<tr>
<td>ASR and ra</td>
<td>$r = .36$</td>
<td>$p &lt; .0001$</td>
</tr>
<tr>
<td>ASR and rc</td>
<td>$r = .30$</td>
<td>$P &lt; .0001$</td>
</tr>
</tbody>
</table>
before actual FL learning theories were espoused, linguists were grappling with the mysteries of language learning. Bloomfield (1933) an early American linguist, in his seminal work: “Language”, described simple qualities of language such as word stress that can affect not only pronunciation, but comprehension as well. For example, in English contrasts can used to express stress at the beginning of a word, for example, ‘a name and an aim’. “French uses no stress phonemes, and cannot in this way mark its word units (Bloomfield, p. 182)”. Could something as simple as word stress used by English learners have been applied to sentence pronunciations by beginning French learners? In this case would the misplaced stress have affected the ASR scores students received? Would expert raters, having had experiences with French learners, accommodated, intuitively, for these factors in their rating? Would these factors, human factors, be accounted for in a correlation?

The human factors in this study, as expressed through the rater scores and correlations of the scores among the raters, point to the fact that raters were more consistent among themselves than they were with the machine-generated scores. Generally the problems for foreign language raters are: 1) it is difficult to use scales for rating fluency, 2) it is difficult to train raters to use the scales, 3) where raters are trained effectively, it is difficult to ensure consistent use of the scales for rating FL pronunciation, 4) raters are conscientious and want to rate effectively and judiciously, but are often fatigued, 5) over an extended time raters may become distracted or report rating more strictly or more leniently. Given what we know to be true of the problems for FL raters, especially for rating pronunciation - coupled with what we have learned about ASR scoring - added to what students have reported about their use of the ASR software, it is
reasonable to conclude that ASR scoring was, on average, comparable to human raters, at scoring participant’s French sentence pronunciations.

What is needed is a more robust study as regards statistical power, that is a larger $N$, and a statistical tool that can account for more of the variability observed in the study. McNamara (1996) suggested the use of RASCH measurement tools where the individual aspects of each dimension investigated can be independently measured and compared. For example, the raters could be evaluated for degree of leniency or harshness in scoring participants. Individual French sentences could be analyzed by participant for high (7) or low (2-3) ASR scores. With the computing power available for large numbers of operations and comparison measures, RASCH is a tool suggested by numerous researchers (Weir, 2005; Bachman, 2005). However, there is a learning curve for a study designed using RASCH methods and special statistical tools are necessary. Likewise the numerical parameters of each measurement are important and even trained measurement researchers collaborate on RASCH designed studies. Depending on the research orientation there are other tools that can concretely address the limitations existing in an ASR study.

Questions Related to FL Raters and Training

Using expert raters to evaluate FL speech, referred to as the “gold standard” (Eskenazi, 2009), (where human judgments are used) has been a consistently difficult task to validate, given human factors. However, it is the one that has and continues to be used for FL evaluation of pronunciation and speech, especially within the classroom
setting. In some cases, tests have been designed around the use of expert raters. The ACTFL-Oral Proficiency Interview (OPI) is one example.

As noted earlier in chapter 5, there were questions about the adapted scale used for rating. In a future study or replication, the adapted rating scale needs revision based on suggestions made by the raters, especially if sentence length pronunciations are to be rated. The short length of the audio recording contributed to the difficulty of listening quickly and with extreme concentration. Another challenge for the raters included using a numerical scale and deciding whether the sentence included the sound or phoneme variables heard in the pronunciation.

The findings indicate that future studies using human raters would benefit from having the raters trained together to encourage interaction among the raters. Raters interacting with each other during a training session by providing feedback regarding the technology, the rating scale and practice rating sample audio files together would provide a collaborative and perhaps more consistent training environment across raters. Each of these features were included in the rater training provided to raters a, b and c, but they were done individually with each rater. Determining the time factors and constraints from real-time rater experiences were meaningful and can provide useful insights for future research.

It is a question also whether raters could be trained separately from the rating experience where two or three separate sessions would be scheduled: 1) a training session, 2) a rating session and 3) a re-rating session. As suggested, rater a and b may have had an advantage as they rated for a pilot study (2008). In cases where more than two raters are used, pilot study training would be useful for each rater.
The rating scale adapted for the study could also be further scrutinized and tested by the raters. Although the scale was adapted after the pilot, more work needs to be done to validate adapted scales, especially where FL pronunciation and speaking are concerned. High-stakes testing and the resources available to large testing bodies (ETS and IELTS) will hopefully contribute to rating scales for use by researchers and teachers, in the future. The International English Language Testing System (IELTS, 2008) has developed a scale for pronunciation that is being tested and validated which could significantly contribute to high-stakes FL pronunciation evaluation.

**Conclusion**

If participant perceptions and feedback from users are considered as “evidence”, for positive impact (Chapelle, 2001) and usability (Neilsen, 2006), then a strong case for using CALL designs that include ASR for student pronunciation and speaking practice has been made. Participant voices and perceptions were expressed clearly through their audio productions and their descriptive comments, regarding their own learning experiences. Not only participant perceptions but also statements of their needs for pronunciation and speech learning support the need to include FL language learning - ASR software as integral component of speaking and pronunciation skill practice. Students need to have tools available for FL learning in the classroom, lab or at home. This becomes even more critical for students engaged in distance, or hybrid learning where students need immediate access to FL native models and examples.

In the interest of addressing the larger FL community and CALL researchers in particular, it is reasonable to conclude that much work still needs to be done in the area of
speech-interactive products using ASR. While the findings of the study indicated that there are inherent limitations to CALL designs using ASR features, the results point to the fact that the long-standing FL tradition of using trained expert human raters for rating speech or pronunciation may, in this case, be comparable to the ASR-produced scores. Using machine-scored ASR software to assist FL learners with pronunciation, given the problems experienced by expert raters, would: 1) eliminate the need for creating, testing and training raters to use scales, 2) ensure that a consistent measure was used for each student, 3) eliminate worries about rater fatigue, 4) reduce distractions and human factors that could interfere with reliable rating.

What we know to be important in any form of feedback and especially feedback for beginner French learners applies to the ASR-generated feedback. Students want to engage in interaction, to practice and have consistent feedback for improvement. From this perspective the machine can undertake these tasks much more reliably and consistently. Based on conclusive evidence, it is almost inexcusable not to make ASR available to FL learners. ASR can only add to the speech-interactive CALL resources, when they can be made accessible to learners - a technologically sophisticated group of foreign language students
References


Ducate, L. & Arnold, N. (eds.). (2006)/ Calling on CALL: From Theory and Research to New Directions in Foreign Language Teaching, CALICO Monograph Series, Volume 5. San Marcos, TX: CALICO.


153


Appendices
Appendix A – Pilot Projects Using TeLL me More French

Design Considerations and Features

Chun (1998) made several recommendations for future software designs for teaching intonation and many of the features suggested are technically possible with current versions of ASR. Her suggestion to “provide learners with visualization of their intonation patterns and specific contrastive feedback” (p. 81) is available with pitch-tracking software. With consideration of the suggestions for research provided by Chun (1998) and by using TeLL me More with ASR software, many of the research tools are available and current software “tools can provide audio and visual feedback to the learner, can serve as a data collectors, and utterances produced can be saved, ordered and compared over time” (Chun, p. 87). However, analysis of natural, spontaneous speech, specifically with ASR, is still limited.

A research design suggested by Chun (1998) addresses the question of type of ASR feedback provided. The ASR software of TMM can provide corrective feedback for pronunciation errors, through native speaker audio models and visual models of waveforms describing pitch curves. Student performance can be recorded aurally and visually. As well learner performance can be scored on a scale (3-7), and students can compare their pronunciation to the model provided.

With the permission of Auralog’s educational division, a preliminary investigation was conducted to examine the usefulness of the TeLL me More (French) speech recognition software for teaching French pronunciation. The oral productions of French students and an examination of the features of the ASR were evaluated by several
Appendix A (continued)

volunteers during a one semester period Fall 2005. A short pilot questionnaire was
designed to evaluate student perceptions and satisfaction with the TMM learning
activities and ASR features. The preliminary findings and feedback from the volunteers
was used to inform the procedures and design of the Pilot Study – Spring 2006.

Pilot Project - Spring 2006

In the Spring 2006, using Tell me More (TMM) French software for research
purposes, a project was conducted in order to examine the Automatic Speech Recognition
(ASR) component. The software was installed on fifteen of the twenty-five computers in
the foreign language computer lab. University-enrolled French students were oriented to
the ASR and the software was available during lab hours. Students were free to use the
software to practice their French pronunciation, or to engage with other TMM multimedia
activities.

A research group was established and members cut across disciplines and
contributed a variety of background knowledge in French language, teaching and
instructional technology, including research in multimedia CALL. A research study was
designed to investigate the effectiveness of the speech recognition features of the
software, with a specific focus on the efficacy of the ASR for providing students with
French pronunciation and speaking practice. Also of interest were student satisfaction and
perceptions, as well as language learning, and software design and usability.
Appendix A (continued)

Three learning paths: sounds, words, sentences. The tutor tools available with the TMM allows for the design of learning paths for practicing specific speech and language features. Pre-designed paths can be directed to specific student accounts. In this way participants were only given access to pre-designed learning paths.

Three learning paths were designed to provide students with samples of French speech for pronunciation practice. Learning path 1 provided pronunciation practice with distinct sounds. For example, if the sound practiced was /l/ then path 1 provided the French word, belle, which includes the sound. Path 2 included words, including sounds practiced in path 1. Path 3 included sentences that provided practice using combinations of previous sounds and words as well as punctuation. Sentences provide a more diverse speech sample because with word/sound orders and punctuation students can practice and demonstrate discourse features of speech such as intonation. Intonation is expressed through the voiced, affective dynamics of a sentence. For example, an exclamation in French can result in a rising, assertive tone at the end of the sentence.

Speech data and ASR scoring. The scores produced by the TeLL me More ASR and assigned to the recorded speech are internally generated by algorithms. The ASR scores are part of the computer-programmed design of the ASR software. Auralog has proprietary rights to the ASR design algorithms and has developed a speech recognition system entitled: S.E.T.S or Speech Error-Tracking System. The ASR –S.E.T.S system designed and owned by Auralog is now only part of the multimedia software, an
important feature, but originally (1993) the French company designed and sold, only, the speech recognition software.

The speech recognition features of the software are able to produce scores (seen on the screen as “bars”) based on the following scale: 1-2 not-rated (considered a threshold level for recording and recognition); 3-4-5-6-7. The ASR scores are produced as a result of an internal, algorithmic comparison of the French sounds to an ASR corpus-base of sounds. Students can also choose to listen to a native speaker example (recorded in the software). A score of 7 represents the most complete score. Figure 8 is a screen shot of the scores or bars in the ASR software as seen on the Speech Recognition activity screen.

Using the three paths outlined above, produced speech samples were collected and recorded for forty participants. Two separate data collection sessions, approximately three weeks apart were conducted for each participant. For each session, a minimum of five ASR-scored recordings, per example, per path were collected. For example, if for path 1 the sound was /l/ and used in the word belle, then five repetitions of the sound resulted in five ASR scores for belle.
Appendix A (continued)

Figure 8. ASR Software “Score” Box

Participant survey. After each of the two data collection sessions an online survey was completed by each participant. The date and participant number were used as identifiers for both the recordings and the survey. The two online surveys provided a qualitative component and an informative and direct link between individual participant speech recordings and related ASR scores. The survey included several categories of questions designed to elicit participant responses on experiences with learning French and the TMM software. Participant perceptions of the speech recognition features and experience using the ASR were of particular interest.

Data collection and analysis. The data have been collected from the FL lab computer that serves as the “server” for the site licenses. Data cannot be exported or
Appendix A (continued)
electronically transferred and is an acknowledged limitation of the software. A significant drawback for research purposes has been the manually intensive data collection process. As of December 2007, all of the data have been manually transferred, recorded and organized in an Excel data file.

In June 2006 a case analysis of three participants, representing three levels of French language based on years of learning, was conducted. In the three cases the participant’s scores and survey data were analyzed and examined to address the following questions: 1) Did a participant’s produced speech scores improve across path 1, path 2, and path 3? From session 1 to session 2? 2) What were the identifiable and observed features and interesting characteristics of the scores by learning path? 3) How does the survey submitted for session 1 correspond to and differ from session 2? Are there relationships between the survey reports and the produced speech for each session?

Analysis of the Spring 2006 performance data continues to be empirically and qualitatively analyzed. The design features of the pilot study have informed the current study design and method. The methodological considerations that were used to examine features of the ASR, speech production and oral performance have been thoroughly considered for the method and corresponding questions.

Pilot Project – April 2008

In April 2008 a pilot project was conducted for the purpose of pilot testing the rater site instrument to be used in the current study. Using the sound files of sentences
Appendix A (continued)

extracted from the software and collected from participants in a pilot study conducted August 2007, the rater site was pilot tested. The sound files were converted from the open .oggs to MP3 format and the survey was designed using the audio recordings of 12 participants x 5 sentences x 2 recordings each sentence for a total of 120 sound files for rating.

**Human raters.** Two French speaking raters (one native speaker and one French Professor) were trained using a Rater Site (Pilot). Three participants were rated for the training. Raters were instructed in the use of the rubric, the rubric designed and adapted from the ACTFL Guidelines-Speaking 1999 and practiced rating the three participants. Following the rater training, the raters entered the Rater Site (Beta) and proceeded to rate the 120 audio recordings of sentences.

Following the rating session, raters were asked to enter a final site to answer five questions regarding the rating experience. The training, rating session and question period lasted approximately forty-five minutes.

**Analyses.** The Rater Site (Beta) data was put into an Excel file for analysis. Using SAS code for a correlation calculation, an analysis was conducted to compare the consistency of the raters-ratings (Rater A and Rater B) as well as the ASR scores provided for the sentences. Preliminary findings indicated that the raters correlated well with each other, that is their ratings using the rubric were consistent and that the rubric can be expected to assist the raters in providing an evaluation of the participant audio.
Appendix A (continued)

recordings. The ratings of the ASR-scored audio, when compared to the ASR-generated
scores, did not show a correlation.

The data revealed interesting patterns that could be considered for the future
analysis. For example, it was noted that the ASR scores for the first recording were, in
many cases, the same for the second recording. As well, the increase in participants and
the number of data points, that is an increase of about 180 data points (audio files)
contributed to the effect size and the numerical significance of the analyses for the
current study.

April 2008 Pilot Instruments

Rater Site (Pilot) location:

http://survey.acomp.usf.edu/survey/entry.jsp?id=1207155217939

Rater Site (Beta) location:

http://survey.acomp.usf.edu/survey/entry.jsp?id=1204735687244

Rater Questions location:

http://survey.acomp.usf.edu/survey/entry.jsp?id=1207772730008
Appendix B – French Sentences for Pronunciation

Produced using TeLL me More French-ASR software

Beginner Level

10 French Sentences

1) Le « R » français est tellement charmant !
2) Je ne sais pas comment le dire en français.
3) J’ai encore des progrès à faire.
4) Je fais de gros efforts en ce moment, tu sais !
5) Je ne parle pas très bien le français.
6) J’espère qu’ils parlent anglais !
7) Pouvez-vous répéter ?
8) Oh non, mon vocabulaire est très réduit !
9) Je viens de terminer un cours intensif.
10) En réalité, je voulais améliorer mon français.

Data Set: 10 Sentences

Data Points:

Totals: 300 audio files and 300 ASR Scores

10 sentences repeated 1x = 300 ASR scores per participant

1 sound file per sentence = 10 ASR sound files per participant

10 sound files x 30 participants = 300 sound files/sentence pronunciations for rating for Rater A, Rater B and Rater C.
Appendix B (continued)

English sentence translations:

1) The French “R” is so delightful!

2) I don’t know how to say that in French.

3) I need to make more progress.

4) You know, I am really working hard!

5) I don’t speak French very well.

6) I hope they speak English!

7) Could you repeat that please?

8) Oh no, my vocabulary is very limited!

9) I just finished a very difficult course.

10) Honestly, I would like to improve my French.
Appendix C – Numerical/Descriptive Rating Scale

(Adapted from: ACTFL Proficiency Guidelines-Speaking 1999 level descriptors and the Interagency Language Roundtable (ILR) skill level descriptions for Speaking.)

Novice: Ability to speak minimally with only learned material.

1 – 2 Low/Mid: Undecipherable, unintelligible speech, difficult production and/or poor recording (recording may be chopped off). Difficult to rate the sentence pronunciation sample.

3 – Mid/High: Oral production of the sentence is limited and includes frequent errors and faulty sound pronunciation and word stress.

Intermediate: Combining learned elements primarily in a reactive mode.

4 – Low – Elementary sentence pronunciation, stress and intonation are generally poor and may be influenced by a L1. Fluency is strained and sentence flow is difficult to maintain.

5 – Mid – Pronunciation may be foreign but individual sounds are accurate. Stress and intonation may lack control and exhibit irregular flow. Longer sentences may be marked by hesitations.

6 – High – Able to pronounce sentences with a degree of precision. Sound, word and sentence pronunciation and stress are obvious, and effective.
Appendix C (continued)

Advanced: Ease of speech, expressed fluency and sentence pronunciation is marked by competent sound and word pronunciation; complete sentence stress and intonation.

7 – Low/Mid – Sentence is comprehensible with no discernable errors in sound or word pronunciation. Speech range is “native-like” and accurate.

Please note: In some cases the sentence audio recording is incomplete. This may be due to other factors than the participant’s pronunciation. For example, if there was a longer hesitation (more than 2 seconds) the Automatic Speech Recognition (ASR) may have not recorded the sample completely. It is assumed that the rater training and the “good” recordings, along with your French expertise and pronunciation ability will give you a gauge for identifying where these factors are the result of pronunciation errors or audio recording errors.
Appendix D – Survey Questionnaire

Question 1

If you have studied French, at the university, for only one semester please choose: French
I. If you have studied French, at the university, for 2 semesters choose: French II. If
neither of these categories apply choose: Other

Question 2

How many years have you studied French? [One semester = .5 years]
Is there anything about your French studies that you would like to explain further?

Question 3

Do you usually feel comfortable using computers and new software?

Question 4

Have you used language learning software or speech recognition software before?

Questions 5a and 5b

Many people are anxious when speaking aloud in any situation. Could you please
describe two situations:

a) One situation where you are or have felt comfortable speaking French.
b) A second situation where you feel or have felt anxious, self-conscious or
uncomfortable speaking French.

Question 6

How would you describe your feelings about the experience recording your
pronunciation of the French sentences with the ASR software today?
Question 7 a, b, c, d

Based on the French pronunciations that you recorded today did you find it helpful:

a) To listen to your own pronunciations?

b) To listen to the native speaker model?

c) To look at the visual graphs?

d) Please tell us more about your impressions and your experience with these software features.

Question 8 a, b, c

TeLL me More ASR provides a score (represented by 'green bars') for each of your sentence pronunciations:

a) Were the ASR scores helpful for your pronunciation?

b) Do you think the ASR scores reflect your French pronunciation ability?

c) We are most interested in the ASR score as it is presented to you. Could you please comment on exactly what this score meant to you?

Question 9

How helpful is the ASR software feedback (scores, native model, visuals) compared to other feedback you have received on your French pronunciation and speaking?

Could you please comment about how this software's feedback has been more or less helpful to you?
Appendix D (continued)

Question 10a and 10b

Could you describe a time when you had your French speaking or pronunciation rated or evaluated by your French teacher?

a) How do you feel about having your French speech or pronunciation rated by an expert human rater (either known or unknown to you)?

b) Would you want the rater to be a native French speaker?

Final Comment:

Please take a minute or two to add anything more you want to say about your experience at the lab today or about the ASR software.
Appendix E – Survey Question 10 Qualitative Coding Example

The notes for the coding of survey Q10 were handwritten and scanned to a PDF document. The PDF is attached to the electronic document. The printed documents are attached with this document.

a) Coding – memo notes: page 64

b) Memo – draft diagram page 65

c) Diagram: page 66

d) Theoretical note: page 67
Appendix E (continued)

a) Coding – Memo notes: page 64

Memo: 5/10/09  re: C10 part 1

After reading through one time and beginning initial coding
to become clear that there were specific "coding" words
used to describe time. This first became apparent
when after reading I started to notice I actually had
to write down the word during and some subcategories
that began to emerge – for example –

"during class"

"during oral interviews" etc.

It is interesting to me that this is the word that
popped out because during as a word – implies
process, movement, start and, motion – a process
happening while something else is happening – is
this how students feel about speaking – pronunciation
that it is something that happens during speaking?
or during "trying" (or "speaking") a foreign language?

It is interesting as well because often they are
referring to during class – so does this mean that
they feel they are being evaluated anytime there
is the chance (during class being a good chance –
depending on how they describe the teacher) they might
have to speak the FL (or French)...

How does this concept of time as having a
flow, a movement, as beginning and ending describe
students' view of perceptions of speaking? Does it?

There were very few mentions of the computer or software
and only 1 or 2 mentions of recording?!
b) Memo – draft diagram: page 65

It appears that students might even be expressing the concept of "interaction" as process? The teacher describes the teacher input, and even seem to expect desire, want teacher feedback, and then describe their response — by either correcting themselves, taking it in planning — something doing something — in response — so time is the process of interaction — with the teacher or material or language or both, and the outcome — they perceive — receive.

Draft Diagram (idea) for Time Concept:

- Time
  - Input
  - Teacher/prof/instructor
  - Competition
  - Test/exam
  - Self
  - Speaking/communication
  - Outcome
  - Grades
  - Corrections
  - Specific word
c) Diagram: page 66
d) Theoretical note: page 67

---

Theoretical note for Q10 part 1  S1/10/09

So exciting to stay with it and see the diagram emerge — it is almost as if you are writing and then... etc. Something emerges and you had just had to start drawing + writing. A saw this process of time clearly — or seemingly so — a concept of time — and in EL terminology as "interaction" and the distinct feeling that students accept for the most part this process — they expect to interact and to experience an outcome, even if they are not happy with their input — they expect to gain enhance learn? Something about themselves.

Again, this emergence of students wanting, wanting to speak, pronounce, learn the language — to use the language is it possible that they conceive of speaking as using the language? Are speaking + using synonymous in their thinking? In their learning experience? This whole process — feels so exciting right now.
About the Author

Deborah J. Cordier, Ph.D. in Second Language Acquisition/ Instructional Technology (SLAIT), University of South Florida, Tampa (2009). Research interests include CALL and interactive technologies for foreign language speaking, pronunciation and FL testing; Automatic speech recognition (ASR), developing research and related trends in NLP and human-computer interaction (HCI). Online community involvement (2005) working worldwide with researchers and academics to develop awareness and materials for Open Education Resources (OER). Previous experience includes French, eleven years in K-12 international education and work with IB programs; Translation studies (Fr-Eng) certificate and RSA-CTEFLA certificate.