AUTOMATIC SPEECH RECOGNITION: COMPUTER AIDED TOOL FOR TEACHING AND LEARNING OF ENGLISH LANGUAGE IN NIGERIA SCHOOLS

REBECCA ISAAC UMARU, Ph.D.
Faculty of Education,
Nasarawa State University,
Keffi.

and

BARAU NUHU BEST
Department of Computer Science,
Nasarawa State College of Education,
Akwanga.

Abstract

Good communication skills continue to be the foundations of learning, emotional development and socialising throughout a young person's schooling and onward into the workplace. Young people need effective communication skills especially speech and language in order to have a wide range of life choices. Automatic speech recognition (ASR) is special computerized means whereby voices are converted into text. The writers see this as a potential opportunity on the education threshold especially in the teaching of English language. The terms involved were duly explained, the technicalities and associated challenges with this new trend were not left out. The writers are of the view that proper use of ASR will go a long way in assisting students and teachers by easing the teaching and learning of some aspects of communication skills in English language. Strong recommendations were outlined and the need for improvement in ASR technology was duly acknowledged.

Speech syntheses widely referred to as Automatic Speech Recognition (ASR) is the artificial production of human speech. A computer system used for this purpose is called a speech synthesizer and can be implemented in software or hardware products. Speech synthesis could be in text to speech where written words are converted to speech it could also be speech to text, here words are converted into text. This is termed or called online synchronous. In educational system it could be referred to as online synchronous teaching and learning. This paper is based on speech to text recognition (STR) system and its usage in teaching and learning of English.

Speech to text technology is described as a process which involves a teacher speaking into a microphone, wherein the speech is recognized and shown synchronously in the form of text for students to read (Way, Kheiret and Bevilacqua, 2008). Automatic speech recognition (ASR) can be defined as the independent, computer-driven transcription of spoken language into readable text in real time (Stuckless,1994). In a nutshell, ASR is technology that allows a computer to identify the words that a person speaks into a microphone or telephone and convert it to written text. Having a machine to understand fluently spoken speech, has driven speech research for more than 50 years.

The goal of an ASR system is to accurately and efficiently convert a speech signal into a text message transcription of the spoken words independent of the speaker, environment or the device used to record the speech (i.e. the microphone). This process begins when a speaker decides what to say and actually speaks a sentence. (This is a sequence of words possibly with pauses, uh's, and um's.) The software then produces a speech wave form, which embodies the words of the sentence as well as the extraneous sounds and pauses in the spoken input. Next, the software attempts to decode the speech into the best estimate of the sentence. First it converts the speech signal into a sequence of vectors which are measured throughout the duration of the speech signal. Then, using a syntactic decoder it generates a valid sequence of representations (Rabiner andJuang, 2004).

There are fundamental reasons why so much research and effort has gone into the problem of trying to teach machines to recognize and understand speech. On general terms, these reasons are;

1. accessibility for the deaf and hard of hearinghelps students of all abilities to better exploit their full academic potential by enabling them to:
   i. command and control virtually any Windows application.
   ii. send email and IMs to collaborate on group or classroom projects—entirely by voice.
   iii. search for information on the Web or on their computer.
   iv. edit and format documents by voice.(www.nuance.com).

2. cost reduction through automation

3. searchable text capability

ASR Technicalities and Associated Challenges

ASR is a cutting edge technology that allows a computer or even a hand-held Personal Digital Assistants(PDA)(Myers,2000) to identify words that are read aloud or spoken into any sound-recording device. The ultimate purpose of ASR technology isto allow 100% accuracy with all words that are intelligibly spoken by any person regardless of vocabulary size, background noise, or speaker variables (Center for Spoken Language Understanding (CSLU), 2002). However, most ASR engineers admit that the current accuracy level for a large vocabulary unit of speech (e.g., the sentence) remains less than 90%. A good example of ASR like Dragon's Naturally Speaking or IBM's Via Voice, show a baseline recognition accuracy of only 60% to 80%,depending upon accent, background noise, type of utterance, etc(Ehsani andKnodt, 1998). More expensive systems that are reported to outperform these two are *Subarashi i*(Bernstein, Najmi and Ehsani, 1999), *EduSpeak* (Franco, Neumeyer, Kim and Ronen, 2001), *Phonepass*Hinks, (2001), *Interactive Spoken Language Education ISLE* Project (Menzel, Herron, Morton, Bomaventura and Howarth2001) and *Rapid Application Developer RAD*(CSLU, 2003). ASR accuracy is expected to improve. This still remain a question whether we might use this "Automatic Speech Recognition" (ASR) technology to replace the tedious process of manually transcribing oral spoken words. As with many new technologies, the answer turns out to be both yes and no.

On the workings of ASR, Oard,(2012) used a very skilful illustration to explain it. Taking a peek under the hood to see how ASR works, Oard said it is a bit like the board game Scrabble in which you have some tiles with letters on them with which you want to make a word that fits with what's already on the board.

In the case of ASR, tiles contain the sounds associated with a word, and one will want to fit them together in a way that matches the sounds that were said. Imagine, for example, four tiles with the sounds for "cream," "I," "ice," and "scream." Then one will hear something that sounds like "iscreem" one might either put down the tiles for "ice cream" or "I scream." To decide which of those would be the best choice, is necessary to fit the tiles by putting down into what is already there. For example, if the tiles already laid down correspond to "my favorite desert is …" then "ice cream" would be the better choice. Of course, ASR systems do not really move tiles around, and they consider many more than two choices.

Thinking about ASR as if it were Scrabble helps to explain the kinds of mistakes those systems make. These could lead into three kinds of problems. First, the ASR system simply might not have a tile for the word that was actually spoken. The system, not knowing that, simply does the best it can with the tiles it has. So if interviewee says "anaconda snake," an ASR system that lacks a tile for "anaconda" might produce some nonsense like "anna con the snake." Second, if the person

interviewed has an unusual accent, the ASR system might not have any of the right tiles. This can result in what might charitably call "word salad," producing long runs of nonsense that have little in common with what was actually said. Third, if the interviewee uses words in ways that the ASR system is not prepared for, it might make the wrong choice.

In furtherance, Kim (2006) asserted that ASR has been commonly used for such purposes as business dictation and special needs accessibility, its market presence for language learning has increased dramatically in recent years (Aist, 1999; Eskenazi, 1999;Hinks, 2003). Kim's assertion shows that much research and development on ASR is still on-going. Presently, most ASR-based software programs adopted template-based recognition systems which perform 'pattern matching' using dynamic programming or other time normalization techniques (Dalby and Kewley-Port, 1999). These programs include *Talk to Me* Auralog, (1995), *the Tell Me More* Series Auralog, (2000), *Triple-Play Plus* Mackey and Choi, (1998), *New Dynamic English* DynEd, (1997), *English Discoverie s*Edusoft, (1998), and *See it, Hear It, SAY IT!*CPI, (1997). Most of these programs do not provide any feedback on pronunciation accuracy beyond simply indicating which written dialogue choice the user has made, based on the closest pattern match.

Despite challenges encountered, improvements are on-going in the design process of ASR. The obvious alternative to such a specialized system would be to build an ASR system that works reasonably well for a broad range of oral collections and then doing the best with the higher word error rate that would result. John Hansen at the University of Texas at Dallas has done this with a system called SpeechFind (http://speechfind.utdallas.edu). Despite the higher word error rate, a little experience showed that Speech Find suffices for some searches that one might want to do. SpeechFind also illustrates a second key idea: by providing speech indexing and search as a centralized service, it becomes possible for a wide range of cultural heritage institutions to participate easily by uploading their collections.

Substantial investments in ASR technology continue to be made, both through government-funded research seeking to advance the state of the art and, increasingly, by companies seeking to leverage research results to create profitable applications such as search engines for news broadcasts, podcasts, or personal photograph collections (if your camera records your voice when you take a photo). So it seems quite reasonable to expect continued improvement for some time in ASR accuracy under difficult conditions. Indeed, no one is aware of any theoretical limits that would prevent ASR systems from achieving accuracies (Oards, 2012).

To paraphrase Church and Hovy, (2007), perhaps the problem is not that so much that ASR systems cannot yet do all that is wished, but rather making the best use of the systems that is now build is not attain. This is what actually necessitates the writing of this paper.

Using ASR in Teaching of English Language

In the meantime, several computer-based reading tutors relying on speech recognition and oriented to promote reading in children have been developed (Gruenenberg, Katriel, Lai and Feng, 2008) and (Williams,2002). A recently completed evaluation of several off-the-shelf reading programs that rely on speech technology (Campuzano, Dynasrky, Agodini, Rall, 2009) showed that for at least one such program, statistically significant effects with respect to standardized test scores were found. By far, the most extensive research effort is *Project Listen* (http://www.cs.cmu.edu/~listen/) at Carnegie Mellon University. Among the project's recent findings are improvements in comprehension over a four-month period (Mostow, Aist, Huang, Junker, Kennedy and Lan, 2008), and improvements in reading fluency for elementary students whose first language is not English (Poulsen, Wiemer-Hastings andAllbritton2007).

An ASR related research was carried out using English pronunciation on EFL (English as a Foreign Language) students by Kim (2006). Kim unveiled that teaching pronunciation to EFL students at a low level can be a labor-intensive task for EFL instructors, especially when their classes have 30 to 40 students with a diverse range of proficiency levels. However, ASR pronunciation software such as *FluSpeak* can be used effectively in conjunction with live classroom teaching to develop oral skills. An excerpt from Kim's research in which steps to blend live teaching with self-training in pronunciation for students enrolled in an intermediate English conversation course were outlined thus;

Step 1: Choral Repetition of Each Sentence After the Speaker In Fluspeak Software

This step is a tuning-up session where instructors let students know what they are going to learn. If necessary, instructors explain the meanings of key words and useful expressions that need special attention. Students repeat after the model speaker on the software. During this time, they are allowed to look at the sentences in the book.

Step 2: Self-training Initiated by Students

Once students have established some degree of familiarity with the target sentences in class, they can spend more time with the software in the lab, working sentences that are problematic for them individually. When students see the score of their own pronunciation on the screen, they have good reason to try again to reach a

higher score. This motivation makes students stick to self-training and use the software for a longer period of time. One teaching requirement is that adequate lab time should be allocated for students' self-training with the software.

Their practice recordings are, of course, kept in the program file for the instructor's review.

Step 3: Instructor's Question& Answer Session

An instructor takes up a whole class session to practice the dialogue student by student or in chorus. By this time, students should feel somewhat confident with speaking the sentences since they have self-trained with them on the ASR software. The instructor asks the question in the dialogue and students respond individually or in group.

During this time various other skills such as reading aloud can be practiced.

Step 4: Student Pair Practice

Once students are ready to use the sentences in Steps 1 through 3, pairs of students sit near each other and take turns reading the dialogue sentences to their partners.

Step 5: Students' Simultaneous Repetition with the Model Pronunciation on the Software

In a subsequent lab session, students try to repeat the sentences almost simultaneously with the model speaker. At this step, students are encouraged not to look at the script of the sentence they are repeating. This is the point where the fluency they worked on during the previous steps becomes evident.

Step 6: Role-play Session and Other Creative Skill-Using Activities

Students are given the opportunity to role-play the dialogues in front of the class without looking at the script. Other creative skill-using activities include making up new dialogues based on the one they learned and pair or group presentation in front of the class.

The lesson plan exemplifies the case where ASR pronunciation software leads to communicative skills. In Kim’s experience, students feel more confident with speaking in class when they have practiced pronouncing the sentences privately. Also, were instructors to spend much time drilling students with pronunciation of the basic sentences in the dialogue, which is often the case, they would not have a reason able amount of time to provide the opportunity for communicative practice. Furthermore, instructors tend to agree that this type of pronunciation drill is not always as successful as it should be and rarely can be adequately individualized. Thus, ASR use has two advantages:

1. Students feel more confident in their speaking skill with individualized ASR-based training, and
2. Human instructors can plan on more motivating communicative activities if they leave the low-level basic pronunciation drills to the ASR software.

At the end of the class that the author taught during this study, he took a survey to determine students' reactions to this ASR-based pronunciation class. The survey showed that an overwhelming number of students (90%)reacted positively to the question, "Do you think that *FluSpeak* helps you improve your English in general?"

Their response to the question "Do you think that *FluSpeak* helps you improve your pronunciation?" was slightly lower (86%), and yet still highly favorable, indicating that students perceived an educational benefit from using the software. However, only 30 % of students answered favorably to the question, "Do you think that the pronunciation and intonation ratings of *FluSpeak*are accurate?" This indicates that students tend to discredit the *FluSpeak* software as a reliable tool for evaluating their pronunciation. One of the reasons for their apparent discontent with this aspect of the software might have something to do with the low pronunciation scores the software gave them. Several students complained about low pronunciation scores from *FluSpeak*, even though their pronunciation seemed to be above average as far as the author could judge them. Their idiosyncratic pitch may have been the culprit.

In using the analysis and results of Kim, is evidently clear and acceptable that ASR will certainly play a very good role in improving teaching and learning of English language. Although not all aspects of English language but subsequently more areas will significantly be influenced in the near future.

Conclusively it is true that ASR is of enormous benefit to educational classroom teaching and learning. The advantages are numerous and cannot be over emphasized. Some of the advantages enumerated (www.nuance.com) are listed thus;

1. It is three times faster than typing - Most people speak over 120 words per minute, but the average keyboard user types less than 40 words a minute. By delivering the ability to create documents and emails about three times faster than typing.
2. It is up to 99% accurate – ASR like Dragon learns to recognize the student’s voice with minimal training and delivers increasingly accurate recognition results the more it is used. Also, it never makes a spelling mistake—a real advantage for students with certain learning disabilities. Dragon can easily learn new terms, which can be added individually or quickly imported from existing text.

3. It is a proven tool for improving core reading and writing skills - Speech recognition can serve as a remedial function for improving core reading and writing abilities for students of all abilities. Researchers attribute these benefits to the heightened, strategic engagement with print and language that users experience while dictating and correcting errors.
4. It is the interface of the future - Computing industry pundits like Bill Gates have said that future technology will increasingly leverage speech and touch interfaces. Learning to use ASR products will better prepare students to take full advantage of emerging technology in the classroom, at home and in the workplace.

Summary and Conclusions

In summary, the available research suggests computer-based reading tutors that use speech technology, specifically speech recognition, can lead to increases in reading performance over a relatively short period of time (Bejar, 2010). Also, with continuing hardware and software improvements, speech recognition technology is becoming an increasingly cost-effective educational and productivity tool for a wide range of students. It lets a student talk to a computer and watch his spoken words quickly appear in documents, emails or instant messages is indeed faster than typing.

ASR tools can help students quickly and easily transfer ideas from their minds onto paper; a basic task that is often painful or even impossible for some students. When students can dictate their papers and examination responses to a computer, they are more likely to exploit their full language capabilities.

Therefore, is worth concluding that development of speech technology is a major scientific achievement. The technology has reached a level of maturity that suggests the time may be right to apply it to the acquisition and assessment of reading and speaking skills, including assisting students and adults in developing literacy, and as a tool for the acquisition of English. However, technology should seldom be viewed as a solution to educational challenges in its own right rather, it can at best, support and enable learning. A deep understanding of how students learn and a supportive learning environment are essential for technology to be an asset in the learning process.

Recommendations

Encouraged by some innovative models, developments in ASR appear to be accelerating. The outlook is optimistic that future applications of automatic speech recognition will contribute substantially to the quality of life among students, and others who share their lives, as well as public and private sectors of the business community who will benefit from this technology. The writers wish to recommend thus;

1. Computerization of classroom teaching and learning should no longer be an option rather a must for both students and stakeholders.
2. Acquisition of such sophisticated systems could be expensive thus government should aid the educational sector to realize this 'near dream'.
3. Would be teachers should be encouraged to be computer oriented bearing in mind that their pupils will no longer be of the same aspiration since everyday everywhere today is new inventions.
4. Research into ASR should be on the rise knowing that the need for speech engine is no longer a question and the aspect needs more improvement towards its accuracy.

References

Aist, G. (1999).Speech recognition in computer-assisted language learning. In Cameron, K. (Ed.), *CALL: media, design & applications*, Germany: Swets & Zeitlinger, 165-181.

Auralog (1995).*Talk to me: User manual*, Voisins le Bretonneux, France: Author.

Auralog (2000). *AURALANG user manual*, VoisinsIe Bretonneux, France: Author.

Bejar Isaac I. (2010) Can Speech Technology Improve Assessment and Learning? New Capabilities May Facilitate Assessment Innovations. *R&D Connections* No. 15 August 2010. www.ets.org

Bernstein, J. (1997). Automatic spoken language assessment by telephone (*Technical Report* No. 5-97), Menlo Park, CA: Entropic.

Bernstein, L., & Christian, B. (1996). For speech perceptions by humans or machines, three senses are better than one. Paper presented at the International Conference on Spoken Language Processing, October 3-6, 1996, Philadelphia, PA, USA.

Bernstein. J., Najmi. A. & Ehsani. F. (1999). Subarashii: Encounters in Japanese spoken language education, CALICO, 16 (3), 361-384.

Campuzano, L., Dynasrky, M., Agodini, R. & Rall, K. (2009).*Effectiveness of reading and mathematics software products: Findings from two student cohorts*. Retrieved from http://ies.ed.gov/ncee/pdf/20074005.pdf

Church K. & Hovy E. (2007). Good Applications for Crummy Machine Translation, http://dx.doi.org/10.1007/BF00981759.

CPI (Courseware Publishing International) (1997).*See It, Hear It, SAY IT!* retrieved October 25, 2005 from http://www.usecpi.com.

CSLU (2002).Retrieved October 25, 2005 from http://cslu.cse.ogi.edu/learnss.

CSLU (2003).Retrieved October 25, 2005 from http://cslu.cse.ogi.edu/toolkit.

Dalby. J., &Kewley-Port, D.(1999). Explicit pronunciation training using automatic speech recognition. *CALICO, 16* (3), 425-445.

DynEd (1997).*New Dynamic English CD-ROM Series*, CA, USA.

Edusoft (1998).*English Discoveries CD-ROM Series*, Hebrew, Israel.

Ehsani, F., & Knodt, E. (1998). Speech technology in computer-assisted language learning: Strengths and limitations of a new CALL paradigm. Language Learning &Technology, 2 (1), 46-60.

Eskenazi, M. (1999).Using automatic speech processing for foreign language pronunciation tutoring: Some issues and a prototype. *Language Learning & Technology, 2* (2), 62-76.

Franco, H., Neumeyer, L., Kim, Y., & Ronen, O. (2001).Automatic pronunciation scoring for language instruction. Proceedings of International Conference on Acoustics, speech, and Signal Processing, 1471-1474, retrieved October 25, 2005 from http://www.speech.sri.com/people/hef/papers/icassp97_ pronunci ation.pdf.

Gruenenberg, K., Katriel, A., Lai, J., & Feng, J. (2008). Reading companion: A interactive web-based tutor for increasing literacy skills. In C. Baranauskas, P. Palanque, J. Abascal, & S. D. J. Barbosa (Eds.), *Human-computer interaction – INTERACT 2007* (pp. 345–348). Berlin: Springer.

Hinks, R. (2001). Using speech recognition to evaluate skills in spoken English.Working Papers, 49. Lund University, Department of Linguistics, 58-61.

Hinks, R. (2003). Speech technologies for pronunciation feedback and evaluation. ReCALL, 15 (1), 3-20.

http://dx.doi.org/10.1145/642611.642628. A paper on the “Books with Voices”

http://speechfind.utdallas.edu .SpeechFind was used to index oral history collections as part of a 2004 IMLS National Leadership Grant to the Colorado Digitization Program. See to try out the system. http://www.cs.cmu.edu/~listen/

Kim, I.-S. (2006). Automatic Speech Recognition: Reliability and Pedagogical Implications for Teaching Pronunciation. Educational Technology & Society, 9 (1), 322-334.

Mackey A., & Choi, J.-Y.(1998). Review of Tripleplay Plus! English. *Language Learning and Technology, 12* (1).19-21.

Menzel, M., Herron, D., Morton, R., Bomaventura, P., & Howarth, P. (2001).Interactive pronunciation training. ReCALL, 13 (1), 67-78.

Mostow, J., Aist, G., Huang, C., Junker, B., Kennedy, R., & Lan, H., (2008). 4-month evaluation of a learner-controlled reading tutor that listens. In V. M. Holland & F. P. Fisher (Eds.), *The path of speech technologies in computer assisted language learning: From research toward practice* (pp. 201–219). New York: Rutledge.

Myers, M. (2000).Voice recognition software and a hand-held translation machine for second-language learning. Computer-Assisted Language Learning, 13 (1), 29-41.

Oard, D. (2012). Can automatic speech recognition replace manual transcription? In D. Boyd, S. Cohen, B. Rakerd, & D. Rehberger (Eds.), *Oral history in the digital age.* Institute of Library and Museum Services. Retrieved from http://ohda.matrix.msu.edu/2012/06/automatic-speech-recognition/.

Poulsen, R., Wiemer-Hastings, P. & Allbritton, D. (2007).Tutoring bilingual students with an automated reading tutor that listens .*Journal of Educational Computing Research, 36*(2), 191–221.

Rabiner, Lawrence R. & Juang, B.H. (2004).Statistical Methods for the Recognition and Understanding of Speech. Rutgers University and the University of California, Santa Barbara; Georgia Institute of Technology, Atlanta

Stuckless, R. (1994). Developments in real-time speech-to-text communication for people with impaired hearing. In M. Ross(Ed.), Communication access for people with hearing loss (pp.197-226). Baltimore, MD: York Press.

Way, T., Kheir, R. & Bevilacqua, L. (2008).Achieving Acceptable Accuracy in a Low-Cost, Assistive Note-Taking, Speech Transcription System. Proceedings of the IASTED International Conference on Tele health and Assistive Technologies (pp. 72– 77), ACTA Press.

Williams, S. M. (2002). *Speech recognition technology and the assessment of beginning readers*. Paper presented at the National Research Council Workshop on Technology and Assessment: Thinking Ahead, Washington, DC.

www.nuance.com. Dragon Naturally Speaking Helping All Students Reach Their Full Potential. A White Paper for the Education Industry from Nuance Communications March 2009