CALL:
Linguistic reality and technological limitations
- brains vs chips

Thomas Hansen
University of Southern Denmark
Natural Interactive Systems Lab
thomas@nis.sdu.dk

Abstract

Behind the somewhat pompous name of this paper hides a much more modest aim. It is an attempt at clarifying one of the major problem areas found in the field of Computer Assisted Language Learning (CALL), specifically the use of Automatic Speech Recognition (ASR). The aim of CALL is to produce tools aiding the acquisition of foreign languages and, long-term, to produce full-fledged, virtual, multi-modal language tutors. In essence, *the teacher in the machine.*

Focusing specifically on Computer Assisted Pronunciation Training (CAPT) I intend to present a, hopefully, pedagogic picture of how the mind works in relation to adopting the sounds of a foreign language (L2), both perception- and productionwise, and illustrate the shortcomings of ASR in comparison.

I also intend to present you with some rather unsubstantiated ideas on language learning which can hopefully generate a wholesome debate.

1. Introduction - and some fairly unsubstantiated claims open for discussion

Learning a foreign language is by no means a task which is accomplished easily. In recent years a trend has become visible which, within school systems, exposes children to foreign languages earlier than ever seen before and the immigration politics of different countries focus much more on adult immigrants acquiring the native language as a prerequisite for permanent residence.

As the demand for faster, better and more autonomous language study increases, scientists have begun looking at the computer as a potential aid in language acquisition. Some even speculate in the idea of an autonomous interactive, multi-modal language teacher/tutor. Although I find myself convinced that this will someday be reality, there are, as yet, multiple areas in which technology needs to improve in order to qualify for such a label, ASR not the least.

In my opinion the area in which computers can, presently and long-term, benefit learners the most is within the area of pronunciation. Very often in classes consisting of 20-40 people teachers will not have the available time to practice with and correct individual student pronunciation. Areas such as vocabulary and grammar tutoring lend themselves much more easily to a (blackboard) class-based teaching environment. In my opinion languages are not learned in class. Only building blocks aiding the learning of the language are provided within such an environment. One does not fully learn a language until one has to use the language in an actual situation with native speakers. Furthermore, using a language entails possessing some rudimentary communication skills. It is my firm belief that three different priorities for language acquisition can be established[1]:

1. Vocabulary
2. Pronunciation
3. Grammar and all the rest.

[1] This claim was partly reached through a conversation with a Chinese test subject who said, “I know everything about the grammar, I just can’t for the life of me pronounce it.”

In order to be able to speak a language a learner must know some basic words. Without a vocabulary nothing else will work. Again, in order to use that vocabulary a learner needs to be able to pronounce the contents. Subsequently learners can be taught to arrange words in their proper position. Whether a student utters the sentence: 1) “Me beer want,” or 2) “I would like a beer,” he or she will still be understood. This is comparable to tourists visiting foreign countries and availing themselves of standard phrases which have been picked up in one way or the other.

Learning a language in class is much like learning to drive a car. Initially you are presented with all the rules and interpretations of the signs, but it is not until you find yourself alone, driving the car on your own in actual traffic that you really learn how to drive.

Computer Assisted Language Learning, in time, carries the potential to provide you with your own car, but there are many hurdles which need to be overcome before this becomes reality.

2. Computer Assisted Language Learning

Computer Assisted Language Learning (CALL) and Computer Assisted Pronunciation Training (CAPT) are areas in the academic field that explore the role of information and communication technologies in language learning and teaching. This includes: materials development, pedagogical practice and research. Both are interrelated areas that have experienced rapid growth in recent years.[2] Especially the introduction of the internet and the computer as a household item has spawned the possibility of online learning and computer applications in language learning [Levy, 1997].

Today CALL/CAPT have established themselves as prolific areas whose advantages are well known to language educators. Regardless of learner age these include:[3]

[2] As a matter of fact I believe that this area should be viewed as a branching tree with a headline and several subdivisions, depending on your focus. CALL seems to me to be initial node, CAPT comes directly beneath it on its own branch. Other branches could be areas of grammar learning and vocabulary learning.

[3] See for instance (Neri, Cucchiarini, Strik & Strik 2002).

1) Training available at any time the learner wishes – (install on laptop and transport)
2) More and faster individualized feedback for the learner. – (only the learner uses it)
3) Stress free environment for the learner to practice in. – (nobody listens, no time limit)
4) More practice time than in a classroom setting. – (available anytime)

The above four points should be viewed in terms of a class setting versus a non-class setting.
But, as pointed out by Neri [Neri, 2001], many are concerned with the missing link between technological advances and Second Language Acquisition (SLA) research. One task at hand lies in bridging the gap between proceduralists and formalists, those who wish to build applications versus those with knowledge of language [Chambers, 2001]. In essence a method for pronunciation training is needed which focuses on both the linguistic perception and production facets of learning a language as well as taking into consideration the shortcomings and possibilities of present day technology.

3 Perception and production

Nearly all people are born with the capacity of speaking one language with native competence (L1), but children and adults differ greatly in their ability to learn a second language (L2). To find bilingual children with the competence of speaking both languages with native qualities is not unusual and even throughout their initial years of education they may acquire a third or fourth language where they attain near-native qualities. Research shows that around the age of thirteen the capacity for learning a language seems to diminish, otherwise known as the Critical Period Hypothesis (CPH) [Lenneberg, 1967; Flege, 1987]. Upon reaching adolescence our capacity to interpret new phonetic input (i.e. sounds not indigenous to our native language) seems to fossilize.

This is evidenced in many adults attempting to acquire a second language but producing it with heavy accent coloring.

In 1957 Noam Chomsky coined the term Universal Grammar (UG) [Chomsky, 1957] arguing that all people possessed an underlying language module which acted as a blueprint for acquiring language.[4] Once all the sounds of the native language have been 'put into place,' they are grouped in acoustic categories appropriate for the L1.[5] Heavy accent coloring in L2 acquisition has then hitherto been explained via Equivalence Classification [Flege, 1987]: Upon attempting to acquire the sounds of the L2, these pass through the 'phonological filter' of the L1 and are assimilated to the sounds already known. In other words, the learner will attempt to compare the new sound with an already existing one in his or her native language, and produce the already known sound rather than the new one. This is also known as Acoustic Assimilation. Hence, a necessary component of speaking a foreign language with native quality is the ability to perceive the phonetic differences between the mother tongue and the target language. Additionally the ability to produce the sound is dependent on the ability to perceive it. Polivanov [Polivanov, 1931] and Flege [Flege, 1987] claim that a learner must be able to perceive the difference in acoustic quality before he/she can produce it and thereby make room for new acoustic categories for the language in question.[6]

There are, of course, numerous other processes involved in the acquisition of a foreign language besides the ones that relate specifically to speech viewed as a simple acoustic signal. These relate to extra-linguistic factors such as body movement, but are not treated in this paper.

[4] It should be noted that the theory of Universal Grammar is highly controversial and opposing viewpoints can be found in most of the linguistic literature by J.R. Firth and M.A.K. Halliday. I am not an avid proponent of UG theory, but it does seem to provide a pedagogically sound picture of the elements involved in language acquisition.

[5] Acoustic categories can here be thought of as those 'vowels' and 'consonants' that we view as a fixed inventory for a particular language. Other languages will posses different acoustic qualities.

[6] There are opposing views on the matter such as Borrel [Borrel, 1990] who claims that production precedes perception. Mastering the motor mechanics (in relation to the tongue), should facilitate the perception factor. My personal view, based mostly on simple intuition, disagrees with that statement.

4 CALL – ASR and all the problems we encounter

Today the most widely used piece of technology in CAPT applications are Automatic Speech Recognition (ASR) systems. ASR allows the L2 learner to interact with the constructed software as well as provides the possibility of the learner's pronunciation to be understood and evaluated immediately, by analyzing the acoustic signal provided by the learner. In an ideal CAPT application the level of feedback should emulate that provided by human teachers in terms of, for instance, segmental correction. A teacher standing in front of a student is able to offer corrective feedback in respect to motor movements of the mouth, articulatory advice, guide the student on the phonetic as well as the sentence level. Hence, a non-native speaker pronouncing the word <hello> as <hillo> should be informed that the second segment should be pronounced differently, followed by exemplification. Currently the best available option in terms of computer feedback is provided through the use of ASR systems. However, numerous problems are encountered when employing ASR technology in CAPT systems. These problems are manifold and can be exemplified on, at least, two separate, yet interdependent, levels:

1. Single word pronunciation
2. Sentence level pronunciation

ASR systems, unlike the human auditory system, are currently incapable of distinguishing between speech and non-speech sounds. Hence a cough, a sigh, a hiccup, the slamming of a door, the chirp of a bird, all these sounds will be interpreted by the ASR system as constituting a speech signal and the system will attempt to match this signal with an already incorporated vocabulary. Even the possibility of having a faint echo in the room when producing an utterance will potentially be misinterpreted by the ASR.

Limited vocabulary, or simply out-of-vocabulary, constitutes another problem for ASR systems. Most often the training material available to the student is fairly limited since not all words, counting inflections and similar, are part of the trained or inbuilt vocabulary.

This in turn brings up a related problem, namely that of dialectal variation and speaking style. Dialects can in some cases vary to such a degree that they can almost be considered languages within languages. Even the human auditory system can run into trouble at this level. Differences are also visible in terms of physiological differences in the human vocal tract.

In addition to dialects, we also encounter differences in speaking styles as well as differences between the acoustic signals of men, women and also children. In many cases ASR systems are not capable of handling these differences unless specifically designed to cater to a particular group.

Furthermore ASR systems often run into trouble when the vocabulary becomes too similar. Hence, minimal pairs such as <beep> and <peep> are very difficult to distinguish. Most ASR systems seem happier when the phonetic material provided is polysyllabic.

Even more difficult, if not entirely impossible is the aspect of getting ASR systems to provide segmental feedback. Segmental feedback requires the isolation and identification of single-phones in a context of other phones. Hence the system should ideally be able to recognize whether a [k] is pronounced with aspiration or without aspiration, or even as non-released. ASR systems are composed, mostly, of diphones or triphones, hence eliminating the possibility of single-phone identification. Creating an acoustic model which would be able to distinguish between all the different phonetic realizations of sounds does indeed provide a daunting task.

The already mentioned problems are then even further magnified when examining continuous speech. Continuous speech has no apparent boundaries from which the ASR system can solidly deduce single words. This can be seen from perhaps the most widely used example listed below. How is an ASR system to know whether a speaker says:

1. Recognize speech
2. Wreck a nice beach

As human listeners we have the advantage of context and reasoning, an ASR system does not. Further examples can be found in the area of homophones, which are words that sound the same but are spelled differently i.e. sale versus sail.

All in all speech recognition still suffer from severe shortcomings in terms of functioning as substitute and autonomous language teachers.

5 Conclusion

In this paper I have attempted to outline two very different processes, which aim at achieving the same goal, namely that of creating new speech sounds in learners of a foreign language, perception and production of the mind versus perception in ASR.

As we have seen, acquiring new speech sounds is by no means an easy task once a native language has become fossilized. Human teachers have the ability to tutor and monitor individuals as well as correcting minuscule differences between the mother tongue and the L2, but to some degree they suffer from not having enough time to devote to each individual students. ASR, if viewed as part of a larger package, does contain the possibility of providing students with more time for learning and individual practice and correction, but are simply not able to generate feedback which is fine-grained enough to highlight small differences in language sounds. Furthermore, ASR systems as standalone applications are entirely unable to provide the learner with the perception part of language learning.

ASR does currently provide the best option in terms of generating feedback, but we are a long way away from creating alternatives to real-life teachers.

Borrell, H. Perception et 'crible phonologique in LANDERCY, A. (Ed) *Melanges de phonétique et didactique des langues. Hommage au professeur Renard*. Mons: Presses Universitaires de Mons / Didier Erudition pp. 31-42, 1990.

Chambers, A. Davies, G (Eds) ICT and Language Learning: A European Perspective. Swets & Zeitlinger, 2001.

Chomsky, N. Syntactic Structures. The Hague: Mouton, 1957

Flege, J. E. A critical period for learning to pronounce foreign languages. Applied Linguistics 8 (2) pp. 162-77, 1987.

Lenneberg, E. Biological Foundations of language. John Wiley, New York, 1967.

Levy, Michael. Computer Assisted Language Learning: context and conceptualization. Clarendon Press, 1997.

Neri, Ambra. Cucchiarini, C & Strik, H. Effective feedback on L2 pronunciation in ASR-based CALL. Proceedings of the workshop on Computer Assisted Language Learning, Artificial Intelligence in Education Conference, San Antonio, Texas, pp. 40-48, 2001.

A. Neri, C. Cucchiarini, H. Strik & L. Boves, The pedagogy-technology interface in Computer Assisted Pronunciation Training Computer Assisted Language Learning, 15:5, pp. 441-467, 2002.

Polivanov, E. La perception des sons d'une langue étrangère. Travaux de cercle linguistique de Prague 4: 79-96, 1931.