ISCA Archive
http://www.isca-speech.org/archive

Speech and Language Technology in Education (SLaTE2009)
Wroxall Abbey Estate, Warwickshire, UK
September 3-5, 2009

Responses to Ville: A virtual language teacher for Swedish

Preben Wik[1], *Rebecca Hincks*[2], *Julia Hirschberg*[3]

Centre for Speech Technology, CSC, KTH, Sweden[1]
The Unit for Language and Communication, CSC, KTH, Sweden[2]
Department of Computer Science, Columbia University, USA [3]
preben@speech.kth.se, hincks@speech.kth.se, julia@cs.columbia.edu

Abstract

A series of novel capabilities have been designed to extend the repertoire of Ville, a virtual language teacher for Swedish, created at the Centre for Speech technology at KTH. These capabilities were tested by twenty-seven language students at KTH. This paper reports on qualitative surveys and quantitative performance from these sessions which suggest some general lessons for automated language training.

1. Introduction

Although the use of computers seems almost ideally suited to the practice of pronunciation skills in a new language, computer assisted pronunciation training (CAPT) remains in its infancy in many ways [1]. The literature points to several reasons why CAPT has not lived up to its expectations. Some are pedagogical, some are technological, and some are related to teacher preparedness. The lack of correct, appropriate, individualized and motivational feedback is one of the central issues that has been raised from both the pedagogical and technological points of view. Design decisions regarding appropriate exercises for language-specific pronunciation difficulties inherent in every language is another.

Ville is a virtual language teacher for Swedish, developed at The Centre for Speech Technology (CTT), at KTH [2]. The use of embodied conversational agents (ECAs) in computer assisted language learning (CALL) is seen as one way to address feedback issues [3]. Ville guides, encourages and gives corrective feedback to students who wish to develop or improve their Swedish language skills.

A first version of Ville was offered in the fall of 2008 to all foreign students at KTH who wanted to learn Swedish. The first version focused on helping students with vocabulary training, providing a model pronunciation of new words and drilling students in memorization exercises. Recent research has focused on developing pronunciation and perception exercises designed to raise the awareness of specific aspects of the language that are known to be difficult for many L2 learners to master. Rather than giving a numerical score for how native-like a student's pronunciation is, a common strategy in many current state-of-the-art CAPT systems, Ville has been designed to pinpoint the type of pronunciation error the student makes in linguistic/phonetic terms. We are trying to address both design issues regarding appropriate exercises and feedback issues, by first identifying, in phonetic terms, which pronunciation errors are most important to practice from an intelligibility point of view, and then building specific detectors able to identify and give feedback on such errors. Many difficulties L2 learners have are predictable, and often based on the influence of their native language. More specifically, difficulties are likely to occur for the learner of a new language (L2) if a distinction that carries meaning in L2 does not exist in the learner's native language (L1). For example, L2 features not used to signal phonological contrast in L1 will be difficult to produce and perceive for the learner.

Bannert [4] investigated pronunciation difficulties in L2 learners from 25 L1 languages, with Swedish as target language. Some of the most serious errors with respect to intelligibility were found to be: lexical stress (insufficient stress marking, or stress on the wrong syllable), consonant deletion in a cluster before a stressed vowel, vowel insertion (epenthesis) in, or before a consonant cluster, vowel and consonant duration errors, vowel quality (difficulties with Swedish vowels not present in L1), and prosodic errors.

Work on expanding the repertoire of Ville has resulted in a series of new capabilities addressing these errors. In this paper we investigate how these new capabilities were received, and how difficult and useful students find them. At this stage we do not investigate the long-term effectiveness of new exercises. Instead, this paper reports on an initial test of the system by a group of second-term Swedish learners, in which qualitative feedback was collected and performance monitored, before the new version of Ville is released to a larger audience.

2. Experiment

There were 8 exercises in all, implementing 8 Ville capabilities: Three perception exercises, and 5 production exercises.

2.1. Perception exercises

The first step in learning a new sound contrast is to be able to perceive it. If learners are unable to perceive a linguistic contrast, they are not likely to be able to produce it correctly.

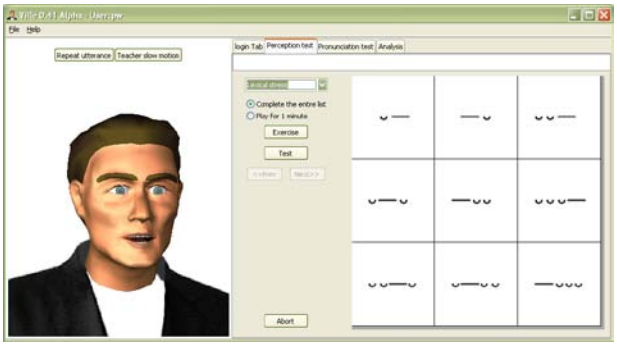

Figure 1: The animated agent Ville, and a three-by-three grid of symbols representing lexical stress patterns. (Underlined syllables are stressed.)

Perception exercises were presented that deal with three common difficulties in Swedish: Lexical stress, Quantity, and Vowel Quality. Minimal pairs are very useful for intuitively exposing learners to contrasts which exist in L2 but not in L1. In minimal pair exercises for vowel quality for example, a pair such as */bita/-/byta/* ('bite' vs. 'swap') is presented on the screen, and Ville randomly says one of the words. The student's task is to select which word has been uttered by clicking on it. Ville then gives verbal feedback on the student's choice. Lexical stress exercises are performed in a similar fashion, by Ville saying a word and the student selecting the word's stress pattern. The classes in this case are not binary, but a three-by-three grid with symbols representing different stress patterns and number of syllables, as shown in Figure 1.

2.2. Production exercises

In the first three production exercises, individual words are targeted. Each word is placed on a 'card', and the cards are stacked on top of each other (so called *flashcards*). Ville says what is on the top card when the student clicks on the card. Carefully selected words that contained specific pronunciation difficulties are grouped together in separate stacks, corresponding to the targeted exercises. There is an underlying XML-structure associated with each card, indicating which mispronunciation detectors should be used to analyze the student recording on each word. The detectors are built on top of Snack, an open-source sound processing tool developed at KTH, in conjunction with N-align, the CTT aligner tool [5].

Quantity -Duration of vowel and following consonant. The Swedish language has what is known as complementary distribution, i.e. a long vowel is followed by a short consonant and vice versa. This causes major difficulties for many students who do not have such a quantity contrast in their L1. A common error is that the duration of the stressed vowel and the following consonant is neither long nor short. Students practice by recording words in which duration changes the meaning of the word.

Lexical stress errors. Lexical stress (making one of the syllables in a word more prominent than the rest) is difficult for students whose L1 has a fixed stress pattern (e.g., Finnish, Polish, French). Commonly measured acoustic correlates to stress are pitch, intensity and duration. Some languages do not have a duration correlate, whereas in Swedish, duration is considered the most important correlate of stress. The same symbolic representation used in the lexical stress perception exercise (Fig. 1) is given on each card in addition to the word

Insertion and deletion errors. The phonological constraints on what sounds can appear in what positions in a student's L1 will often make the student add or omit sounds in L2 words. For example, many native Spanish speakers will produce a consonant cluster with an initial /s/ in Swedish by inserting a vowel in front of the /s/: 'Stockholm' thus becomes 'Estockholm'. Insertion and deletion errors are predictable in the sense that a mispronunciation hypothesis can be created in conjunction with certain consonant clusters. Words that contain such consonant clusters are included for practice.

Mispronunciation feedback. We are experimenting with a layered type of feedback, where red- or green-light icons appear after a student recording to indicate whether the student has performed correctly or not. Since the exercise type in itself has narrowed down the interpretation of the feedback to a question of a phonetic contrast, a binary right/wrong can be informative enough. If the student wishes to know more about why one of the icons is red, he or she can click on the icon, and a new page will appear with more detailed information such as graphs or spectrograms. If, on the other hand, this is a recurring error, and students feel that they have already understood the information, they can simply make a note of the visual feedback and move on. This adheres to the observations of [6], that "Interventions can appear to users as being either timely or irritating. Bothersome interventions tend to be caused by either recognition errors or by a system that intervenes too frequently and is too verbose."

Prosodic errors. In the final two production exercises, students produce whole sentences. An analysis is made of the student's ability to mimic target sentences on four prosodic aspects: timing, length, melody and psyllabicity (pseudo syllabic units). If the student's performance is above a certain threshold, Ville moves on to the next sentence; otherwise, the same sentence is presented again, until the student has repeated it successfully. Two versions of this exercise were tested: Say-after, where Ville says the sentence first, and the students repeat it, and Shadow, where Ville and the student speak at the same time.

2.3. Subjects

Twenty-seven students (13 men and 14 women) were recruited for the study and were compensated for participating. They completed a pre-experiment questionnaire with demographic and language experience questions. Twenty-three were between the ages of 20 and 30. They had been in Sweden for an average of just over a year, and all reported using computers every day. The largest single native language represented was French (6), followed by German (4) and Chinese (3). There were two speakers each of Italian, Turkish and Russian, and individual speakers of eight other languages. All spoke English as a second language, most of them (reported) fluently, and eight spoke a third language as well – excluding Swedish. All were taking their second Swedish course at KTH, classified as 'Advanced Beginners.' They reported using little Swedish outside of the classroom: only about 50% reported using Swedish with friends or watching Swedish TV, while about 20% heard lectures in Swedish or listened to Swedish music, and just over 10% spoke Swedish at home or went to Swedish films. Most of them found speaking Swedish to be the hardest linguistic skill to master, closely followed by understanding spoken Swedish, and then writing and reading in that order. Likewise, they found learning pronunciation and word recognition to be more difficult than learning vocabulary and grammar.

2.4. Experimental design

All the instructions for each exercise were given by Ville. Pre-recorded utterances, switching of panes in the program, and highlighting of buttons or areas of the screen were collected into scenarios, and a collection of these, tailor-made for the experiment, was prepared. In addition to exercising Ville's ability to display scenarios to the students, this approach had the advantage of ensuring that all students were given the same, unbiased information before and after each exercise.

3. Analysis of Subject Performance and Feedback

3.1. Questionnaires: closed questions

Students completed an identical questionnaire after each of the eight exercises of the experiment. Each questionnaire asked five closed and two open questions. The closed questions elicited a response on a five-point Likert Scale, where 1 was the most positive response and 5 the most negative. Written descriptors (i.e. 'very good'—'very bad') accompanied the numbers. The five closed questions were:

Q1. How easy was it to understand how to use Ville for these exercises?
Q2. How useful do you think these exercises were?
Q3. How good were the examples in the exercises?
Q4. How useful was the presence of the animated agent (Ville) in these exercises?
Q5. How likely would you be to do more exercises like this if they were available?

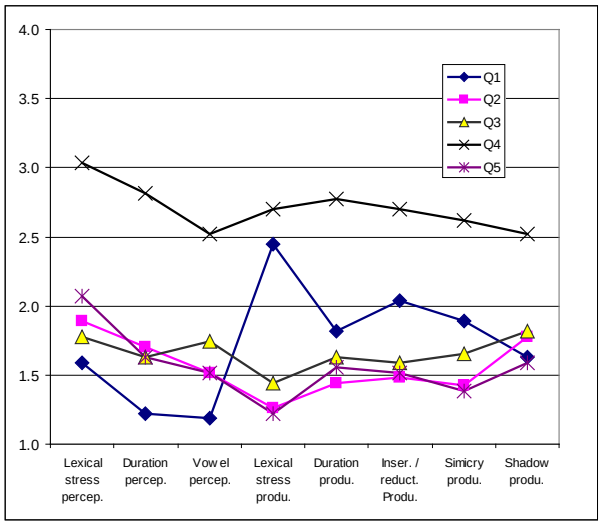

Figure 2: Mean responses to the five closed questions in each of the eight exercises, where 1= most positive and 5= most negative

Responses overall were very positive: the means to all eight Q2s and Q5s regarding usefulness and desire to use again were both 1.56. Figure 2 plots responses to the five closed questions. The first three exercise sections of the experiment tested perception only, while from the fourth section on students were also producing speech. Responses to Questions 2, 3, and 5 are similar to each other across the eight sections, where subjects were most satisfied with the section testing lexical stress production, although they at the same time found it difficult to understand how to do the exercise (Q1). This is possibly because this was the first exercise to test production, and therefore students were simultaneously excited about vocalizing instead of only listening, and unsure of how to follow prompts and interpret feedback. In general, they found it harder to understand how to perform the production than the perception exercises. Subjects were less convinced of the usefulness of the presence of the animated agent. Responses to Q4 deviated strongly from the other very positive answers, though they are better than the neutral point of three.

The subjects completed a final questionnaire when they had finished working with Ville. Here they answered questions about the usefulness of the system overall. Mean responses to questions regarding usefulness and desire to use again were now even slightly improved: both a mean of 1.37. Responses to the question regarding the usefulness of the agent were also least positive: 2.77. Subjects were asked to rank the usefulness of the eight different exercise sections, a task which may have been difficult since they had been introduced to a large amount of new material at once. The resulting ranking favored SayAfter production (1), followed by Shadowing production (2), Duration production (3), Lexical stress production (4), Vowel perception (5), Duration perception (6), and Lexical stress perception (7). Subjects thus preferred the production to the perception exercises.

3.2. Animated agent

The presence of the animated agent is a unique feature for a CAPT system, and therefore subject response to the agent is of interest. As mentioned, subjects gave the agent poorer ratings than the other parts of the system. However, ratings of the agent improved over the course of the experiment. A Pearson correlation of .68 was found between answers to Q4 and time. It may be that, like a person or a human teacher, the agent's presence grew on the students and they came to appreciate him more as they became more familiar with him. One student explained her score of 3 (moderately useful) in the final questionnaire with the added comment that "but I liked that he was there." Another female wrote "I never looked on the face before but in this section, I recognized that I didn't make so many mistakes when I had watched how the animation pronounced the words." In fact, female subjects rated the animated agent significantly more positively than male subjects over all ($t(7) = 6.89$, $p<.001$, two tailed).

3.3. Performance scores

We computed performance scores for each of the exercise sections individually by calculating percentage of tasks performed correctly vs. all attempted tasks, and computed overall means for the whole study, and means for the perception and production sections from these. The overall mean for the study was .65 accuracy, with a standard deviation of .07. There was a significant difference between the perception and production scores, with the former significantly higher than the latter (.77 vs. .59; tstat=7.17, p=.001). Overall score showed main effects for several demographic factors, based on one-way ANOVAs. Younger subjects (18-30) did significantly better than older ones ($F(1,25)=9.57$, $p<.005$). There was also an effect for native language type ($F(1,25)=7.91$, $p=.009$), with students whose native language was Germanic or Romance performing better than students from other language backgrounds. Curiously, we found that subjects who reported speaking Swedish at home did more poorly than others ($F(1,25)=8.04$, $p=.009$). The number of non-native languages (excluding Swedish) that subjects reported knowing showed a tendency to influence overall score but was not significant ($F(1,25)=3.29$, $p=.081$).

Mean scores for each exercise set allow us to rank exercises from least to most difficult: Insertion (.92), Minimal Pair Vowels (.81), Minimal Pair Duration (.81), Lexical Stress Perception (.67) Lexical Stress Production (.62), Shadowing (.61), Reduction (.58), SayAfter (.50) and Duration Production

(.30). Performance and post-exercise responses showed some weak correlations (using Pearson's correlation coefficient); recall that since '1' is the most positive rating in each case, a negative correlation is a positive ranking: Subjects who performed better rated Ville easier to understand how to use (Q1) for the Lexical Stress Perception exercises (r=-.56) but were less likely to do more similar exercises on their own (Q5) (r=.33). Subjects who performed better on the Minimal Pairs Duration perception study rated the presence of the agent (Q4) as less important (r.=.31) although those performing well on the Duration Production exercises said that they would be more likely to do similar exercises on their own (r=-40). Finally, there was a correlation between rating the agent as less important and performance on the SayAfter exercises, with higher performance correlating with worse scores (r=.45). In general, there is a weak correlation (r=.29) between overall performance and agent ratings for the post-questionnaire, i.e., subjects who performed better rated the agent as less useful. None of the other questions show a correlation with the performance scores.

3.4. Open response questions

The open responses to the questionnaires provided useful and varied suggestions regarding Ville's interface design, content, and feedback. The most common comment expressed some frustration at practicing words whose meanings subjects did not know, and requested that words and phrases be presented in writing and translated into English so students could learn new vocabulary. This would be easy to provide in the system, but it may be that, by leaving out the semantics, students can more easily place the cognitive focus on the intended phonetic aspects of these exercises. One of the most successful studies regarding the acquisition of L2 pronunciation deliberately refrained from letting learners know the meaning of the sentences they were learning to say [7].

The feedback on production of lexical stress and duration production was met with some skepticism by some students, who made comments such as "sometimes I feel me and Ville are pronouncing the word the same way, but it's red=wrong". This raises the question of whether the students were unable to themselves perceive the differences in the pronunciations, or whether the feedback was inaccurate in some way. Duration distinctions are not binary in reality, so that a student could have produced a long vowel that was almost long enough to receive a green light, but still have received a red light. A design option to adjust for this could be to add a third feedback alternative such as a yellow light for borderline cases. The analysis tool in the system, with its visual representation of vowel length, was there to help students diagnose their problems, but while some students found it useful, others complained that they did not understand the scoring system used in the analysis. Feedback for other sections was sometimes seen as too lenient; some students felt they had not done a good job but were still rewarded by green lights or high scores. Suggestions were made for summative feedback, showing how well one had done in a section, and adaptive exercises, where subjects were given more examples to practice items they had gotten wrong.

We had asked the students for constructive criticism, and that is to a large extent what we received; however, we also received many enthusiastic comments expressing appreciation of the interface and feedback, and gratitude for the opportunity to improve oral and aural skills. The few comments regarding the agent suggested that he be friendlier and more encouraging, and perhaps female. Only a few subjects realized the usefulness of the visual information the agent provided regarding the articulation of Swedish. It is possible that students need to be explicitly guided to look at the agent; it is also possible that if they were to use the system for more than an hour, they would be able to do so naturally because they would be more familiar with the interface. Students did not hesitate to personify the agent: it is consistently referred to as 'he', 'him', 'the character' or 'Mr. Ville' and ascribed abilities such as 'thinking', 'liking' 'approving' or 'disapproving.' We plan to further explore the effect of the presence of the agent in future work.

4. Conclusion

We have reported on results of a laboratory test of new capabilities for Ville, a virtual tutor for Swedish language learners that uses knowledge of phonetics/phonology to help students learn pronunciation. We found that there is a huge demand for oral training that can be provided outside the classroom to adult language learners at introductory levels. Our findings confirm our expectation that perception exercises are easier to perform than production exercises, and constructing good feedback mechanisms for production studies is harder than for perception. However, our studies have provided useful information on how to improve such feedback. We have also seen that users rate the animated agent in our system as less useful than other features, although students who performed more poorly appreciated him more, and there was a general tendency to think of him as human. This encourages us to believe that the agent is useful for those most in need of help.

5. Acknowledgements

This work was partly supported by NSF IIS-HLC 0534568.

6. References

[1] J. Levis, "Computer Technology in Teaching and Researching Pronunciation," *Annual Review of Applied Linguistics*, vol. 27, pp. 184-202, 2008.

[2] P. Wik and A. Hjalmarson, "Embodied conversational agents in computer-assisted language learning," *Speech Communication*, in press.

[3] O. Engwall and O. Balter, "Pronunciation feedback from real and virtual language teachers," *Computer Assisted Language Learning*, vol. 20, pp. 235-262, 2007.

[4] R. Bannert, *På väg mot svenskt uttal*. Lund: Studentlitteratur, 2004.

[5] K. Sjölander, "An HMM-based system for automatic segmentation and alignment of speech," presented at Fonetik 2003, Umeå, 2003.

[6] M. Eskenazi, "Using Automatic Speech Processing for Foreign Language Pronunciation Tutoring: Some Issues and a Prototype," *Language Learning and Technology*, vol. 2, pp. 62-76, 1999.

[7] G. Neufeld, "On the acquisition of prosodic and articulatory features in adult language learning," in *Interlanguage Phonology* G. Ioup and S. Weinberger, Eds. Cambridge MA: Newbury House, 1987, pp. 321-332.