Speech synthesis for teaching lexical stress

Hincks, R.

KTH Computer Science and Communication

Speech synthesis for teaching lexical stress

Rebecca Hincks
Department of Speech, Music and Hearing, KTH

Abstract

This paper reports on a study carried out on a group of KTH students studying Technical English. Students used WaveSurfer audiovisual synthesis to experiment with differences in pitch and duration in Swedish-English cognates. The exercise helped students achieve long-term acquisition of correct lexical stress for the particular words.

Introduction

Speech analysis and speech recognition are currently in use for computer assistance in teaching foreign languages. Speech analysis has been used for the past quarter century to give visual representation of pitch and duration, illustrating how a learner's intonation differs from a model (e.g Bannert et al. 1981, deBot 1983, Anderson-Hsieh 1992, Germain-Rutherford et al. 2000). In the past decade, speech recognition has been used in language teaching software as a way of creating communicative situations and giving limited feedback on production (e.g. Eskenazi 1999, Delmonte 2000). Speech recognition is also used to score and evaluate L2 production of English (de Jong et al. 2001, Hincks 2001). In the future, animated agents will possibly be able to act as personal automatic language tutors. These 'tutors' will however not necessarily be 'native speakers' of the language they are teaching. Instead, they may speak as is appropriate for artificial beings, that is to say, with synthetic speech. Text to speech synthesis as a learning tool would empower learners to generate the pronunciation of utterances in the absence of authoritative speakers of the language. However, one can ask whether it is appropriate for learners to model their production on an artificial voice.

In the exercise described here, synthesis was used not strictly as a teaching model but as an interactive tool for gaining an understanding of the concepts of pitch and duration and how they can differ between cognates. The pedagogical goal was to achieve long-term acquisition of correct lexical stress in two words that present difficulties for Swedish speakers of English. These words were *component* /kəmˈpoʊnənt/ and *parameter* /pəˈræmɪtər/. A typical Swedish-speaker realization of these words would put primary stress on the first syllable of *component* and the first or third syllable of *parameter,* rather than correctly on the second syllable of both words. These words were chosen because they present problems for Swedes, appear often in technical contexts, and, importantly, are cognates that differ primarily in the placement of lexical stress. The consonants are equivalent and the vowels are similar between the two languages. This allowed us to use Swedish synthesis as a starting point. The segments could not be replaced, but they could be altered in terms of duration and F0 so that the production sounded more like English than like Swedish.

Stress placement in English

The rule governing placement of primary stress of polysyllabic nouns in English is described as follows (Giegerich 1992): "The penultimate syllable is stressed if it is heavy; otherwise stress falls on the antepenultimate syllable." By 'heavy' Giegerich means that the syllable consists of CVC or C + dipthong or tense vowel. It is interesting to note that students' mispronunciations of *parameter* and *component* for the most part adhere to this rule. If the nucleus of the second syllable of *component* is reduced to schwa (as it is in *composition* and *company*) then the antepenultimate, or first, syllable of the word must receive the stress. The pronunciation */ˈkʌmpənənt/ thus follows the stress rules for English nouns. The same holds true of the pronunciation */pærəˈmiːtər/: if the L2 user assumes that the stem of the word is to be pronounced as it is when it appears as an unbound morpheme, /ˈmiːtər/, then the third or penultimate syllable should receive the stress. This presents a chicken-and-egg type of problem to L2 learners. Without knowing the correct vowel quality, which is not obvious from the spelling, the placement of stress cannot be determined,

even if the student has a subconscious awareness of the appropriate pattern. The reverse is of course also true; without knowing the stress pattern, vowel quality cannot be determined. Dickerson and Finney (1978) proposed strategies for helping learners decipher vowel quality and stress placement; their suggestion for a complicated system of rules (three stress rules and seven vowel quality patterns) to be learned by the foreign language student seems however unrealistic and impractical.

The hypothesis behind the exercise described here was that working interactively with the synthetic realization of the words would be an improvement over simply illustrating the correct pronunciation in class. A control study is to be carried out in April and May of this year.

Method

A short text containing at least three appearances of each word to be tested was selected by searching the electronic archives of *New Scientist* magazine. The texts are presented in the Appendix. They are difficult texts but typical of the reading done for the course.

Subjects

Thirteen students of Technical English, eight male and five female, participated in the study. They attended the KTH College of Engineering, a program located on various campuses in the Stockholm area and leading to a Bachelor's degree in Engineering. Their ages ranged from 22 to 39, with a mean of 29. The students' scores on a 100-pt placement test ranged from 32 to 71, with a mean of 54, in line with the average score of KTH students studying English at the intermediate level. Eleven students were native speakers of Swedish, one was a native speaker of Serbo-Croatian , and one had grown up in a bilingual Swahili/English environment.

Procedure

The exercise was carried out as part of a regular course session. Students were divided into three groups, and worked with other assignments when they were not doing the exercise.

Students were given a chance to silently read the texts and were then individually recorded reading them aloud. The equipment used for recording was both a DAT recorder and a mini-disk recorder as a backup.

Students were then introduced to the audio-visual synthesis component of the program WaveSurfer (Beskow et al. 2000). The Infovox 330 diphone Swedish male MBROLA voice was chosen as the model. Students synthesized the word *parameter* or *komponent,* and listened to it a number of times. Their attention was drawn to the part of the window that showed the orthographic representation of the phonetic segments of the word. WaveSurfer places each segment in a box whose size indicates the relative duration of the segment. By adjusting the size of the boxes (via clicking and dragging) the duration of the sounds can be altered. Figure 1 shows the Wavesurfer interface for *komponent*, after alteration.

Students were asked to listen and look for the longest sound of the synthesized word, and then reduce that sound by about half. They were then asked to think about the longest sound in the English version of the word (using their teacher's pronunciation as a model), and add a corresponding amount to the equivalent sound in the synthesized word.

With the durations of the sounds adjusted, the resulting word sounded a bit strange, since the pitch movement indicating stress placement was still in its original position, within the now shortened vowel. Students were asked to move the points indicating the beginning of the rise and the beginning of the fall to positions over the new long vowel.
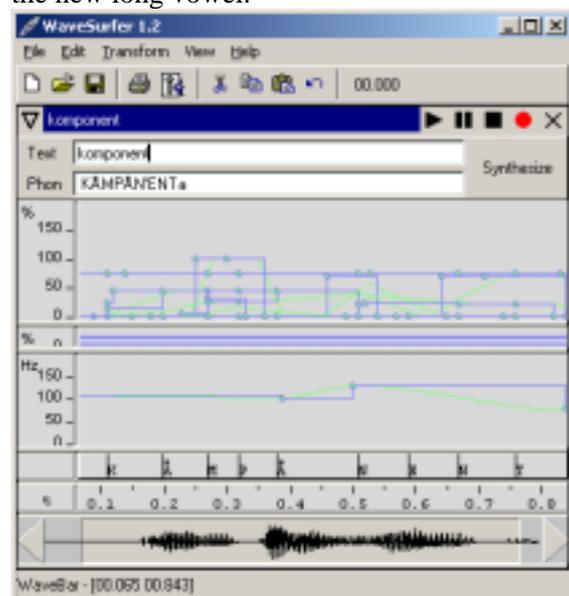

Figure 1. Swedish 'komponent' re-synthesized to sound like English 'component'.

Next, small adjustments needed to be made to some of the other segments, for example

shortening the final /n/ in komponent, to make the word sound more like English.

Finally, the students turned their attention to the visual component of the synthesis. They were shown how to make the talking head make a slight nod on the stressed portion of the word. The entire process was repeated for the other word.

Immediately after the exercise, the students listened individually to the recordings that had been made of their own reading. They were post-recorded reading the same texts four weeks later, in the last week of the course.

Results

In the first reading of the text, students mispronounced both words more often than they correctly pronounced them. In the second reading, most students had mastered the placement of stress within the word. Figure 2 shows the improvement in pronunciation from the pre-test to the post-test.
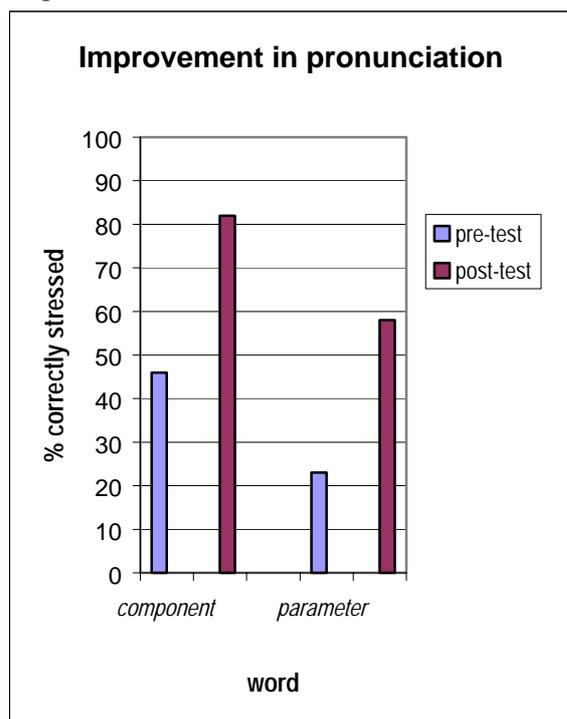

Improvement in pronunciation

Figure 2. Pre- and post-test results, showing improvement. The second bar in each pair represents the post-test.

All students were consistent in their pronunciation of the words in the first recording. In the second recording, however, two students pronounced the word both correctly and incorrectly within the same text, indicating an incomplete acquisition of the new form. An unexpected result was that much of the improvement in the group as a whole could be attributed to the five female students.

Error analysis

All mis-pronounced versions of *component* were stressed on the first syllable: */ˈkʌmpənənt/. Mispronounced versions of *parameter* were stressed on either the first or the third syllable: */ˈpærəmiːtər/ or */pærəˈmiːtər/. In the pre-test, mispronunciations favoured /pærəˈmiːtər/ by about 2:1. In the post-test there was an equal distribution of both types of error. One student attempted to stress both the second and the third syllables of *parameter* in the post-test.

Discussion

Component was an easier word to master than *parameter*, though both words follow regular phonological rules for stress placement. *Component* follows a pattern shared by a number of other well-known words, such as *composite, compulsion* and *companion*. This could have contributed to the relative ease of acquisition. *Parameter* is a word in which phonological rules conflict with learner predictions based on morphological divisions. It is natural to pronounce the word following analogy to words such as *parachute* or *paratrooper*, instead of in analogy to words like *paralysis*.

Another problem with the acquisition of these words could be the negative models presented by many fluent L2 speakers in the KTH environment, who regularly use them incorrectly. One can speculate whether words such as *parameter,* with such unpredictable pronunciation, will be able to survive the increasing use of English by non-native speakers without undergoing a shift in what is considered an acceptable pronunciation in international settings. It would be interesting to look at non-native speakers' perception of incorrectly vs. correctly stressed *parameter.* It could be that incorrectly stressed productions are easier to process for some quite competent L2 users, making */ˈpærəmiːtər/ easier to understand than /pəˈræmɪtər/.

English is the world's lingua franca and as it grows, educators have realized the inadequacies of the existing teaching models for pronunciation, received pronunciation (RP) and general American (GA). Jenkins (2000) proposes the creation of a new phonological model for inter-

national English, incorporating a 'lingua franca core' in which features from RP, GA and L2 varieties of English have been selected for their practicality in functioning as features that can easily be taught and learnt, perceived and produced. As an artificial construction, there are no native speakers of this variety of English. Synthetic speech could be an appropriate vehicle for its dissemination.

Conclusion

This teaching exercise is not generalizable to a large amount of vocabulary and is clearly a time-consuming way of teaching the pronunciation of one or two words. Still, it was gratifying to see the enthusiasm with which the students worked with WaveSurfer and the success with which many of them attained long-term acquisition of particularly difficult words.

Swedish speakers are clearly proficient users of English. One potential area of improvement for many of them is in the placement of stress in polysyllabic words. The multi-tiered nature of the English lexicon usually provides synonyms to which an L2 user can turn when in doubt as to how a word should be stressed, even when the user is sure of the meaning. It is safer to say 'sneaky' than 'clandestine'—for how can one predict the non-reduction of the second syllable in the more sophisticated word? It is our hope to find methods to use speech technology to help English L2 users bring more words from their passive to their active vocabularies

Acknowledgements

References

Anderson-Hsieh J (1992). Interpreting visual feedback on suprasegmentals in computer assisted pronunciation instruction. *Calico Journal* 11, 4, 5-21.

Bannert R and Hyltenstam K (1981). Swedish immigrants' communication: Problems of understanding and being understood. *Working Papers* 21, Lund University Department of Linguistics.

Beskow J & Sjölander K (2000). WaveSurfer -- a public domain speech tool. In *Proceedings of ICSLP 2000,* 4, 464-467, Beijing, China.

Bot K de (1983). Visual feedback of intonation: Effectiveness and induced practice behavior. *Language and Speech* 26, 4, 331-350.

Delmonte R (2000). SLIM prosodic automatic tools for self-learning instruction. *Speech Communication* 30, 145-166.

Dickerson W & Finney R (1978). Spelling in TESL: Stress Cues to Vowel Quality. *TESOL Quarterly* 12:2, 163-175.

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

Germain-Rutherford A & Martin P (2000). Présentation d'un logiciel de visualisation pour l'apprentissage de l'oral en langue seconde. *Apprentissage des Langues et Systèmes d'Information et Communication* 3, 1.

Giegerich H (1992). *English Phonology: An introduction.* Cambridge University Press.

Hincks R (2001). Using speech recognition to evaluate skills in spoken English. *Working Papers from Fonetik 2001*, Lund University Department of Linguistics.

Jenkins J (2000). *The Phonology of English as an International Language*. Oxford University Press.

Jong J de & Bernstein J (2001). Relating PhonePass™ Overall Scores to the Council of Europe Framework Level Descriptors. In Proceedings of Eurospeech 2001, Aalborg, Denmark., 2803-2807.

Appendix

The shell of an egg has the same composition as bone. It has a mineral **component** and an organic **component** (the protein collagen). The mineral **component** makes the shell hard but fragile, while the flexible collagen acts like a glue that maintains the integrity of the whole structure. Acids such as vinegar dissolve the calcium-based mineral **component**, which is alkaline, leaving the flexible collagen intact. Hence you get a rubbery eggshell.
(Pedro Gonzales, *New Scientist*, 10/04/1999)

The density fluctuation **parameter** determined the clumpiness of the Universe when the first structures started to congeal out of the cooling gas of the big bang, about 300 000 years after the Universe's birth. The bigger the **parameter**, the more pronounced would have been the dense regions compared with the average, making them grow faster by pulling in matter through gravity. … [Since] no physics predicts their size at the time of galaxy formation, [scientists have investigated] what kind of Universe might have arisen if the density fluctuation **parameter** had been slightly different.
(Marcus Chown, *New Scientist* 29/11/1997, p.11).