COMPUTER-ASSISTED SECOND-LANGUAGE SPEECH LEARNING: GENERALIZATION OF PROSODY-FOCUSED TRAINING

Debra M. Hardison

Department of Linguistics & Languages
Michigan State University
hardiso2@pilot.msu.edu

ABSTRACT

In a pretest-posttest design, native English-speaking learners of French were given three weeks of training focused on prosody using a real-time computerized visual pitch contour display. Native speaker (NS) exemplars stored on hard disk provided feedback including overlay of the NS pitch contour on the subjects’ productions. These were recorded and later presented to NSs in two conditions to determine the influence of segmental information on their perception of prosody: filtered (unintelligible segmental information) and unfiltered. Ratings were given using 7-point scales for the prosody of filtered samples and both the prosody and segmental accuracy of unfiltered samples. Results revealed a) significant improvement in prosody (filtered and unfiltered samples) after training, b) transfer to segmental improvement, c) generalization to novel sentences, and d) some segmental influence on the NS perception of prosody. In a subsequent memory recall task using a subset of filtered NS training exemplars with reduced intelligibility, subjects identified the exact content of more than 80% of the sentences. Subjects also reported a greater awareness of the various aspects of speech and increased confidence in producing another language.

1. INTRODUCTION

Many factors influence a native speakers’ judgment of second-language (L2) learners’ accent including suprasegmental features [1]. The current study focused on the acquisition of French prosody by native speakers (NSs) of English using computer-assisted training that permits visual display of pitch contours in real time, and the extent to which such training generalizes to improvement in segmental production. This is reported as Experiment 1. Experiment 2 explored the relationship between prosodic information and lexical content in long-term memory using prosody as a lexical access cue.

Previous studies have been conducted on the effectiveness of visual pitch feedback with Dutch university students learning English [2]. In a pretest-posttest design, auditory-visual (AV) versus auditory-only (A-only) feedback, and the amount of practice (one session of 45 minutes vs. 2 sessions) were investigated. After a sentence was presented auditorily, the pitch contour was displayed and subjects imitated the sentence. Their pitch contours then appeared on the display below the target. Subjects were allowed to repeat the process a nonspecified number of times. Ratings of these productions were done by three teachers of English using a 5-point scale. Results indicated that AV feedback was significantly better than A-only, but the amount of practice time was not significant. A subsequent study with a similar population demonstrated that the duration of feedback delay, that is, the period of time between the speech signal and pitch contour display (40 ms., 250 ms., or at the end) did not significantly affect improvement [3]. These studies involved brief training periods and focused on the subjects’ imitation of NS sentence models.

Experiment 1 had the following primary objectives:

• Evaluate the use of computerized feedback of subjects’ productions in the form of visual pitch contour displays in real time through a pretest-posttest experimental design and three weeks of training.
• Limit the role of NS pitch contour displays to feedback for comparison with subjects’ attempts in contrast to the use of NS data as models for imitation [2].
• Determine if training focused on prosody with feedback only on prosody would transfer to improved segmental accuracy.
• Determine if improvement would generalize to the prosodic and segmental features of novel sentences in a generalization task.
• Compare native speaker (NS) ratings of subjects’ productions with filtered versions (rendering segmental information unintelligible) to investigate the rating procedure; that is, are NSs able to avoid being influenced by segmental quality when asked to focus their attention only on prosody?
• Compile trainer observations on the use of speech technology in language learning.
• Gather information from anonymous questionnaires completed by the subjects on the effectiveness of such training.

The acquisition of prosody in French by NSs of American English presents an interesting contrast that often poses problems for L2 learners. The traditional syllable-timed/stress-timed distinction used to characterize the difference between French and English has not been supported by experimental findings. Instead, researchers have suggested the rhythm of French is unlike that of English in that it is “trailer-timed” [4]; that is, the salient element is a right-boundary phrase-final prominence at the end of a word or sense group in contrast to English described as leader-timed (left-boundary foot-initial prominence). In French, the right boundary stress is generally associated with a perceptible fluctuation in pitch as well as some degree of lengthening.

2. EXPERIMENT ONE

A pretest-posttest design was used to measure the effects of training. There were 20 testing sentences. The posttest was followed by a test of generalization involving 20 novel sentences. There were 13 training sessions each about 40 minutes. Each
testing and training period took three weeks; a total of three were conducted. A control group completed the testing but received no training.

2.1. Method

2.1.1. Subjects

Subjects were 16 NSs of American English who were undergraduate students at an American university enrolled in the first semester of the second year of French study. None were language majors and none had studied abroad. Contact with French was limited to classroom instruction. All were motivated to improve their production of French. Some expressed an interest in studying abroad while others wanted to travel, and all wanted to improve classroom performance. A control group included 8 subjects who had been similarly motivated to participate but could not attend the daily training sessions because of their schedules. Those in the experimental group were paid $15.00 for their participation.

2.1.2. Materials and Procedure

Selection of stimuli for testing and training followed these guidelines: a) familiar vocabulary (determined by examining instructional materials for the first and second years of study), b) functional value, c) several exemplars within semantic domains (e.g., food and wine, student life, travel), d) sustained phonation (to provide the best possible continuous display of pitch contour), e) relatively short length (to facilitate production from memory rather than reading), f) structural variety (i.e. declaratives, interrogatives and imperatives), and g) range of segments (including those that are often difficult for learners, e.g. nasal vowels, liaison, and the French /r/).

Training sentences were recorded by three female NSs of French (non-Meridional) direct to computer hard disk using Kay Elemetrics Computerized Speech Lab (Model 4300B) with a Shure unidirectional microphone and JBL studio speaker. Each NS recorded a different set of 30 sentences. Each was instructed to look at a sentence printed on a card, look up, and then produce it at a conversational rate of speech. Sentences were played back to check intelligibility and naturalness of expression.

Subjects were tested and trained individually. For the pretest, posttest, and test of generalization, subjects were shown sentences printed on a card. They were allowed to practice the sentence aloud before recording (primarily to deal with any anxiety especially during the pretest). After looking at each sentence, they were instructed to look up and produce the sentence at a conversational rate into the microphone. Sentences were stored as separate files on hard disk. During testing, there was no auditory or visual feedback.

For the training sessions, subjects were shown each sentence printed on a card. As with testing, they were then instructed to look away from the card, pause and produce the sentence into the microphone to avoid ‘read’ speech. Each training session focused on one set of 30 sentences. The pitch contour of a subject’s utterance was displayed in real time in View Screen B on the monitor and played out through the speaker. The utterance was then replayed and redrawn. In View Screen A, the NS’s version was displayed providing audio and video feedback. In this way, the NS’s pitch contour serves as feedback but not as a model for imitation. It was then overlaid on the subject’s in View Screen B and displayed in a contrasting color. The screen was then cleared and subjects practiced the sentence again. Only a subset of subjects’ training productions could be saved because of limited disk space.

Following the posttest, subjects were given a test of generalization involving 20 novel sentences with no visual or auditory feedback.

2.2. Results and Discussion

Subjects’ recorded productions were evaluated on a 7-point rating scale by a total of three NSs of French. Rating sessions were done individually. Each subject’s pretest, posttest and generalization sentences were randomized. Raters were not told which productions by the experimental group preceded or followed training. The productions were blocked by subject for presentation to raters in filtered and then unfiltered versions.

A digital filter (low pass, Blackman, 100th order) was created using the CSL to render the segmental content of speech unrecognizable leaving only the prosodic information. The cutoff frequency was set at 300 Hz (all subjects were female). To determine this value, a spectrogram with overlay pitch extraction was generated for each speech sample to determine the point above which there were few, if any, F0 components. Each filtered speech sample was saved as a file on hard disk and later evaluated by NSs of French who were unable to identify any of the words. These two versions (filtered and unfiltered) were used in order to determine if raters’ judgments of prosody were being influenced by segmental quality (either in a positive or negative direction). In addition, in the filtered version, some NS samples were included to ensure that raters were able to appropriately rate native-like prosody as none had had any experience listening to filtered speech. These samples received ratings of 7.

For each version, raters were given the sentence on a response sheet along with a 7-point scale ranging from “1” (definitely not native) to “7” (definitely native-like) for the prosody rating. In the unfiltered version, there were two scales for each sentence: one for a rating of prosody and one for segmental accuracy. Providing the sentence was necessary for the rating of filtered speech; therefore, it was given for all ratings. All speech samples were played directly from the computer through the studio speaker for ratings. Ratings for the filtered speech samples were obtained first. Raters were presented with a series of filtered versions of sentences recorded by the author to familiarize them with filtered speech.

The tabulated ratings were analyzed with matched t-tests. Interrater reliability was .83. Results are shown in Figure 1. Comparison of pre- and posttest prosody ratings revealed a significant improvement as a result of training [df=15, t=31.67, p<.001]. Some of the subjects’ posttest tokens received ratings of 7; this occurred more often for prosody than segmental accuracy but there were a few in that area as well.

A comparison of pre- and posttest segmental accuracy ratings revealed significant improvement over the period of training even though training had focused on prosodic features and subjects had not been told that segmental accuracy would be assessed [df=15, t=12.15, p<.001].

To determine if training improvement had generalized to novel sentences, ratings of the novel sentences were compared to those of the pretest. A significant difference again was found for both prosody [df=15, t=7.4, p<.001] and segmental accuracy [df=15, t=8.09, p<.001].
unfiltered and filtered speech samples were compared. Analysis influenced NS assessment of prosody, the ratings for the pretest revealed a significant difference [\(df=15, t=4.14, p<.001\)]. Ratings comparison showed a greater difference between the ratings for prosody and segmental accuracy had improved to the point where there was a greater difference between the ratings for prosody and segmental accuracy compared to the pretest as shown in Table 1.

<table>
<thead>
<tr>
<th></th>
<th>Unfiltered</th>
<th>Filtered</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Prosody</td>
<td>Segments</td>
</tr>
<tr>
<td>Pretest</td>
<td>4.40</td>
<td>4.23</td>
</tr>
<tr>
<td>Posttest</td>
<td>5.35</td>
<td>4.75</td>
</tr>
<tr>
<td>Generalization</td>
<td>5.14</td>
<td>5.12</td>
</tr>
</tbody>
</table>

Table 1: Effects of Prosody Training: Mean accuracy ratings (7-point scale where 7 is most native-like)

To address the question of whether segmental quality influences NS assessment of prosody, the ratings for the pretest unfiltered and filtered speech samples were compared. Analysis revealed a significant difference [\(df=15, t=4.14, p<.001\)]. Ratings for the unfiltered version were higher. However, the difference in ratings between the posttest unfiltered and filtered versions was not significant [\(t=2.00, n.s.\)]. At the time of the posttest, prosody had improved to the point where there was a greater difference between the ratings for prosody and segmental accuracy compared to the pretest as shown in Table 1.

Note that the three ratings for the generalization sentences are quite similar (5.14, 5.12 and 5.13). Segmental accuracy ratings were actually higher for the generalization sentences than for the posttest ones. I discount a mere practice effect here as the prosody ratings (the focus of the training) were not similarly influenced. The prosody ratings for filtered and unfiltered generalization samples, while good and significantly higher than those for the pretest, were a bit below those for the posttest sentences. The high segmental accuracy ratings could be the influence of the content of the sentences despite efforts to make them as equivalent as possible. No attempt was made to count up the number of /r/ sounds, nasal vowels, etc. and, as noted below, subjects did begin to notice these segmental features in the later stages of training.

Throughout the training sessions, subjects’ comments and the author’s observations as the trainer were noted. There appeared to be a hierarchy of noticing in terms of the various features of speech. Initially, all subjects focused their attention on the “flatness” of their pitch contours in contrast to the “peaks and valleys” evident in the contours produced by the NSs and displayed as feedback. The salient convex shape of final intonation in French drew the attention of most subjects. By the middle of training, other features of their speech were then noticed such as rate of speech, general fluency, mispronunciations of words, and the absence of obligatory liaison in some contexts. Some subjects would describe their own early pitch contours as just “dots on the screen” in contrast to the more “flowing” contours of the NSs.

At the conclusion of each training session, subjects were given a questionnaire to complete and return anonymously to me in order to assess the pedagogical value of this type of speech technology. Of the 16 subjects in the experimental group, 13 returned the questionnaires. Briefly, subjects reported that prior to the training program, the most difficult elements of French pronunciation had been rhythm, intonation, the /r/ sound and liaison. The training had raised their awareness of the rise and fall pattern of NS intonation, the shape of the pitch contour at the end of sentences, and the fluidity of the speech often described as being like “humming a tune”. They reported noticing more features of their own pronunciation in French and becoming more comfortable in speaking the language as a result of the training. Subjects reported improved pitch variation, smoother rhythm, and fluency.

3. EXPERIMENT TWO

Following the test of generalization in Experiment 1, subjects were given a separate memory recall task to explore the relationship between prosody and the lexical content of sentences in long-term memory. Specifically the research question was whether subjects could recall the content of a sentence used in training if prompted by its prosody (i.e. by using filtered speech).

Results of a related experiment dealing with recognition memory of English sentences by advanced learners (native speakers of Cantonese) had shown a high level of lexical memory in recognizing sentences they had heard before (“old” sentences) but performance generally was poor in terms of recognizing when an old sentence’s prosody had changed causing a different interpretation suggesting that unanalyzed prosodic patterns had been stored in memory and subjects had failed to generalize the information to new examples [5].

3.1. Method

3.1.1. Subjects

All 16 subjects from the experimental group and 8 subjects from the control group in Experiment 1 also completed Experiment 2. Subjects had not been told of this task in advance.

3.1.2. Materials and Procedure

From the set of sentences from training, 20 were selected representing a range of structural types and all training talkers. This subset was filtered using the CSL filtering program; however, as there is not a unique correspondence between the suprasegmental and segmental features of sentences, the objective for this experiment was to reduce substantially but not eliminate the intelligibility of the lexical content of the sentences so that prosodic information was the principal lexical access cue. NS judgments were used to determine the appropriate level of reduction.

The filtered sentences were played through the studio speaker to the subjects who were asked to try to recall as many of
the words of the sentence as they could. Each sentence was played three times. Subjects’ responses were oral.

3.2. Results and Discussion

Results revealed that the training group from Experiment 1 was able to recall the exact content of an average of 80% of the filtered sentences, and the content words only of an additional 10% of the sentences. Recall performance was quite consistent across subjects in this task. Subjects from the control group of Experiment 1 who had not been exposed to these training sentences were unable to identify any lexical content.

4. GENERAL DISCUSSION

This study has demonstrated significant effects of computer-assisted training in the acquisition of L2 prosody and generalization to segmental accuracy and novel sentences. The subjects appeared to allocate their attentional resources hierarchically. As the training program was explicitly designed to deal with prosody, this element took precedence and their attention initially, but as subjects became more confident with this aspect of their language production, they were able to notice other areas of the language also (e.g. specific consonants and vowels, liaison, etc.). In addition, the results of Experiment 2 suggest that prosodic cues facilitate the recall of lexical information following exemplar-based learning models in which all attended perceptual details of events are stored as traces in memory.

Taken together, the above findings also emphasize the pedagogical applications of speech technology. Its value can be seen not only in the quantitative results but also in the raising of the subjects’ awareness of the various components of language as shown by their questionnaire responses following training.

The use of computers in the development of spoken language skills for second-language learners has also been discussed by other researchers often with English as the target language [6,7], although the pitch display software program used in the present study is not language specific. The use of speech technology can be integrated into traditional language curricula still maintaining the current pedagogical emphasis on communicative language teaching.

5. REFERENCES