COMPUTER-BASED SECOND LANGUAGE PRODUCTION TRAINING BY USING SPECTROGRAPHIC REPRESENTATION AND HMM-BASED SPEECH RECOGNITION SCORES

Reiko Akahane-Yamada, Erik McDermott,
Takahiro Adachi, Hideki Kawahara and John S. Pruitt

ATR Human Information Processing Research Laboratories, Kyoto, Japan
2-2, Hikaridai, Seika-cho, Soraku-gun, Kyoto, 619-0288, Japan

ABSTRACT

How can we provide feedback to second language (L2) learners about the goodness of their productions in an automatic way? In this paper, we introduce our attempts to provide effective feedback when we train native speakers of Japanese to produce English /r/ and /l/. First, we adopted spectrographic representation overlayed with formant frequencies as feedback. Second, we investigated the correlation between human judgments of L2 production quality and acoustic scores produced by an HMM-based speech recognition system. We also adopted the HMM-based scores as feedback in the production training. Evaluation of the pre- and post-training productions by human judges showed that production abilities of the trainees improved in both training groups, suggesting that both spectrographic representation and HMM-based scores were useful and meaningful as feedback. These results are discussed in the context of optimizing L2 speech training.

1. INTRODUCTION

Human speech perception and production patterns become language-specific quite early in the course of language acquisition. Humans generally acquire the phonetic inventory of their first language (L1) without difficulty. However, it is not always easy for them to acquire the phonetic system of another language after once establishing the L1 phonetic system ([1], [2]). Accordingly, the learners of L2 have to overcome the difficulty either through daily exposure or through extensive training.

The goal of the present project is to provide an effective self-training method for adults to learn to produce speech segments in L2. There are several possible production training methods which differ mainly in their feedback. One, which is thought to be direct and helpful, is a method in which a visual representation of trainees' articulatory behaviors is presented as feedback. However, the real-time representation of articulatory behaviors is not easy with today's technologies and, in addition, special equipment is required.

Another method utilizes the visualized acoustic properties, such as those shown by a sound spectrogram, as feedback. This technique is already used for clinical purposes (e.g. SVIII by IBM com.), but has two substantial disadvantages when applied to self-training. First, trainees with no knowledge of speech acoustics have difficulty in reading and interpreting the visualized acoustic properties. Second, it is hard to correct articulation behavior from acoustic properties, since there is often no simple correspondence between gesture and acoustic structure.

The third method uses feedback based on the acoustic similarity between the trainees' production and a template. For instance, Kewley-Port and her colleagues have developed a production training system, so-called ISTRA, for hearing-impaired children, and this system has met with success ([3]). In her system, the acoustic similarity metric was estimated for the similarity between each new utterance and a stored template which represented the best recent utterances of the trainee.

In this paper, we developed two techniques, related to the latter two methods above, and conducted respective studies to evaluate their potential for L2 production training. In the first study, we adopted a spectrographic representation with the results from a formant-tracking analysis overlayed as feedback and trained Japanese speakers to produce English /r/ and /l/. In order to compensate for the disadvantages mentioned above, we started the training with prolonged /r/ and /l/, and gradually introduced shorter coarticulated utterances. The goal of this procedure was to help the trainee to notice the important cues in the visualized acoustic properties, and to learn the positions of articulators which produce the appropriate acoustic properties.

In the second study, we examined whether an HMM-based speech recognition system (the ATR HIP MECS system [4]) can assess the L2 learners' production quality in a manner consistent with human evaluation. A database of /r/, /l/ and /w/ productions by native AE speakers and Japanese speakers was used to train and test HMMs for this purpose. We compared the confidence scores from the MECS system and the human evaluations for this database, and also examined the recognizer's recognition accuracy for utterances from different ranges of pronunciation quality. Furthermore, we adopted this HMM score as feedback and trained Japanese speakers to produce English /r/ and /l/.

2. STUDY 1: SPECTROGRAPHIC FEEDBACK

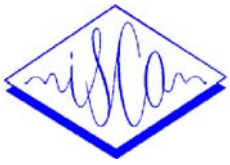
2.1. Method

Procedure

The experimental set-up employed a pretest–post-test design. Before and after the training period, pretests and post-tests were administered, where both production and perception ability of /r/ and /l/ were tested. The trainees were native speakers of Japanese (9 male and 1 female with an average age of 21, ranging from 18 to 24).

Testing

In the production test, recordings were made of the subject's productions of /r/-/l/ minimal pairs. English words contrasting /r/ and /l/ in five syllabic positions (word-initial singleton, word-initial consonant cluster, intervocalic, word-final singleton, and word-final consonant cluster) were used as the stimulus materials. A reproduction task was used in which the trainee read 34 English /r/-/l/ minimal word pairs from a randomly ordered list.



5th International Conference on Spoken Language Processing (ICSLP 98)
Sydney, Australia
November 30 - December 4, 1998

In the perception test, a two alternative forced choice (2AFC) task was used. In each trial, two members of a minimal pair were each displayed by a button shown on the CRT monitor. One of the members was then played over headphones. The subjects chose one of the words by pressing the appropriate key. English words contrasting /r/ and /l/ in five syllabic positions produced by three native speakers of AE were used as stimuli. There were approximately 300 stimuli in total.

Production Evaluation

The productions from pretest and post-test phases were later evaluated by native AE speakers who were trained in phonetics. In the first session, the intelligibility of the consonant was assessed. The sequence of the stimulus presentation was blocked by subject; the pre- and post-test versions from one subject were presented in a single block in random order. The evaluator identified the consonants by using a two alternative forced choice (2AFC) task contrasting members of the /r/-/l/ minimal pairs.

In the second evaluation session, evaluators rated the goodness of the productions with knowledge of the talkers' intended word using a 7 point scale (1: worst - 7: best). Again, the sequence of the stimulus presentation was blocked by subject, and the pre- and post-test versions from one subject were presented in a single block in random order.

Production Training

The production training procedure was based on the reproduction task. A spectrographic representation overlayed with tracks of the first three formants of the model's and the trainee's productions were provided as feedback. The training lasted about five hours on three separate days (i.e., approximately 100 minutes per day). On the first day of the experiment, there was a pre-training phase, which lasted about ten minutes, followed by training sessions. In the pre-training phase, the trainee was first taught the correct tongue positions for the phonemes /r/ and /l/ via verbal instruction by the experimenter. Following this, the trainee produced prolonged /r/ and /l/. Spectrographic representations with overlayed formant-tracking results of the model's and the trainee's productions were displayed simultaneously on the CRT monitor. The main acoustic cue on the spectrogram (i.e., low F3 frequency for /r/ and high F3 frequency for /l/) was brought to the trainee's attention.

In the training, sixty-eight /r/-/l/ minimal pairs produced by three AE male talkers were used as model productions which the trainee imitated. In order to manipulate the length of /r/ and /l/, model productions were analyzed and re-synthesized using the STRAIGHT algorithm developed by Kawahara ([5]). STRAIGHT allows for large changes in acoustic parameters, such as duration, with little or no change in the perceived naturalness of the speech. There were three versions for each word. In two of these versions, the portions of /r/ and /l/, which contrasted the words in the pair, were expanded to have durations three times as long and twice as long as the original duration; in the third version, the /r/ or /l/ portion was left unchanged (E3, E2 and E1 stimuli, respectively).

On the first day, after the pre-training phase, a training session using E3 stimuli was conducted. On the second and third day, training sessions using E2 and E1 were conducted, respectively. In each session, stimuli by three talkers were presented in three separate blocks.

On each trial, the target word was presented on the CRT monitor in English orthographic form, and then the model sound was played over headphones. The trainee produced the word by imitating the model sound, and immediately after this production, the spectrographic representations of speech by the model talker and the trainee were displayed together with the formant-tracking results (Fig.1).

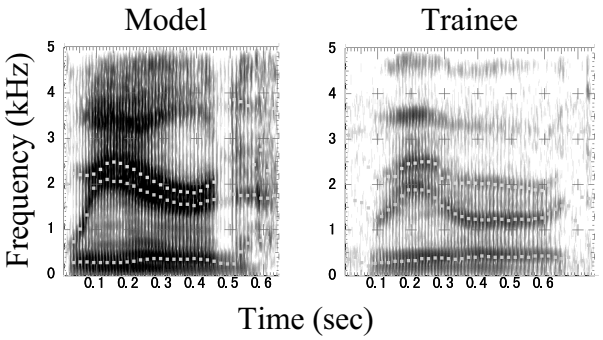


Figure 1: Example displays during production training for the word, "wired" by one of the model talkers.

2.2. RESULTS

Intelligibility scores (in terms of how often the AE listener's identification responses matched the talkers' intended words in the 2AFC task between members in minimal pairs) and goodness rating scores were calculated for the subject's pretest and post-test productions. Overall intelligibility and goodness across 10 trainees improved from 62.7% and 3.36 in the pretest to 85.1% and 4.18 in the post-test (Figure 2). Perception accuracy showed a small but significant improvement of 3.6% from pretest(57.0%) to post-test(60.6%).

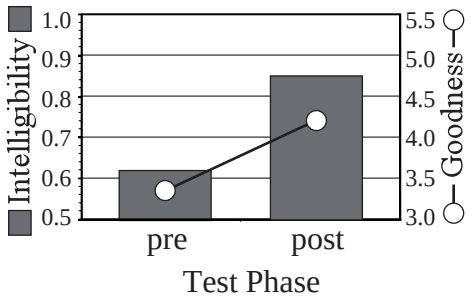


Figure 2: Intelligibility (bars) and goodness scores (circles) of trainee's pretest (pre) and post-test (post) production as judged by AE listeners.

3. STUDY 2: HMM-BASED PRONUNCIATION RATING

In a previous study[6], we investigated the effects of English language experience on native Japanese speakers' ability to produce English /r/, /l/, and /w/. In doing so, we employed native English speakers to evaluate our subjects' productions. In the current study, we compared these human evaluations of intelligibility to an evaluation generated by an HMM speech recognition system to determine whether such automated evaluation could provide reliable and meaningful scores to L2 learners. We further examined the effectiveness of production training using the HMM scores as feedback in order to determine the applicability of such automated evaluation.

3.1. Database

Speech materials

Three groups of subjects produced /r/, /l/, and /w/ in 10 vowel contexts by reading from a list after listening to a recorded model of the utterances. The subjects were: 144 speakers of Japanese who had never lived abroad, 144 speakers of Japanese who had lived in the USA from 1 to 15 years, and 38 native speakers of American English. All utterances used are single syllables of two phonemes, containing /r/, /l/ or /w/ in the initial position, and a vowel, /iy/, /eh/, /aa/, /ao/, /uw/, /ey/, /ay/, /oy/, /ow/, or /aw/ in the second position.

Human Evaluations

Approximately 100 American English speakers evaluated the productions by classifying them as /r/, /l/, /w/ or "other". The evaluators were blind to the intended phoneme and language group of the speaker. A per-utterance intelligibility between 0.0 and 1.0 was calculated by averaging the evaluators' classifications for the intended phoneme.

Speech Corpus for HMM Training and Testing

The guideline for the HMM design adopted here was that HMMs should be trained on high quality utterances (utterances with good intelligibility scores), from many different speakers, and tested on utterances of various intelligibility from a distinct set of speakers. The data set used to train the HMM-based speech recognition system consisted of 1140 utterances from 38 AE subjects and 3922 "perfect" utterances (utterances with intelligibility scores of 1.0) from 212 Japanese subjects, for a total of 5062 utterances and 250 speakers. The test set consisted of 2103 utterances from a distinct set of 73 Japanese subjects (the average intelligibility score of the test set was 0.80). This test set was subdivided into two smaller sets reflecting intelligibility intervals between 0.0 and 1.0.

3.2. Evaluation of HMM ratings

Hidden Markov Model design

Each utterance was transformed into a sequence of feature vectors, each consisting of 21 MFCC-based values, including MFCC deltas and delta energy, calculated for a 20 ms frame of speech at a 10 ms frame rate. To facilitate the focus on the pronunciation of /r/, /l/ and /w/, the vowels were transcribed as a single symbol. Maximum Likelihood Estimation (MLE) was used to estimate HMMs for the three phonemes of interest /r/, /l/ and /w/, as well as for the generalized vowel symbol, and an utterance initial/final silence symbol. The HMM for each category was assigned 3 states, each state consisting of a mixture with 8 gaussian components.

Recognition rate on testing data

The correct recognition rate for the resulting HMMs was evaluated on the testing set. The task was to identify each utterance correctly among the three possibilities, "sil(ence) r vow sil", "sil l vow sil" and "sil w vow sil". In addition to the recognition rate for the entire test set, the recognition rates for various testing subsets (each containing utterances from a given interval of intelligibility) were also evaluated. The results are summarized in Table 1.

Correlation with human ratings

A finer measure of the match between HMM and human judgments is to measure the correlation between human and machine evaluations for each utterance in the entire test set. For this purpose, HMM pronunciation scores were generated for each utterance. The score definition used here was similar to that proposed in [9], but posterior probabilities (for /r/, /l/ and /w/) were directly calculated at the segment level from the segment-level log-likelihoods, rather than at the frame level. These scores were then correlated with the human intelligibility ratings. The correlations for /r/, /l/ and /w/ were 0.78, 0.81 and 0.86, respectively, corresponding to an average correlation of 0.82.

Intelligibility	# Correct/Total	Accuracy
0.0-0.1	73/203	36.0
0.1-0.2	24/48	50.0
0.2-0.3	14/25	56.0
0.3-0.4	33/54	61.1
0.4-0.5	56/73	76.7
0.5-0.6	31/42	73.8
0.6-0.7	36/45	80.0
0.7-0.8	157/184	85.3
0.8-0.9	98/110	89.1
0.9-1.0	1256/1319	95.2
0.0-0.9	522/784	66.6
0.0-1.0	1778/2103	84.6

Table 1: HMM recognition rates for /r/, /l/ and /w/

3.3. Production Training using HMM scores

Procedure

In order to determine whether the training using HMM scores as feedback is effective or not in the production training, two native speakers of Japanese, one male (MS01; age 26) and one female (MS02; age 27), were trained to produce /r/ and /l/ in a short production training session. The experimental design was similar to the one in the STUDY 1: it employed a pretest–post-test design, and in each pretest and post-test, production recordings and perception tests were administered.

Training

After the pretest, the same pre-training instructions as in STUDY 1 were given. First, the tongue positions for /r/ and /l/ were verbally explained by the experimenter. Second, the training of prolonged /r/ and /l/ was administered by using spectrographic representations with overlayed formant-tracking results.

The production training procedure was based on the reproduction task where the HMM score was provided as feedback. Since our HMMs were trained by using only syllable initial consonants, we used 14 minimal pairs contrasting /r/ and /l/ in the initial position produced by two AE speakers, one male and one female, as model productions of which the trainees imitated.

On each trial, the target word was presented on the CRT monitor in English orthographic form, and then the model sound was played over headphones. The trainee produced the word by imitating the model sound, and immediately after this production, the % probability that the target consonant was recognized as the intended consonant by our HMMs was displayed. The trainees were allowed to retry producing the same word until they were satisfied. Trainees completed 4 sessions, each consisting of 28 trials, in about two hours.

The productions from pretest and post-test were later evaluated by two native AE evaluators who were trained in phonetics. The evaluation used the same procedures used in STUDY 1.

Result

The pretest and post-test production scores and the accuracy in the perception test are shown in Table2. The production ability improved from pretest to post-test (intelligibility: 93.2% to 97.2%, goodness: 5.31 to 5.46). Accuracy in the identification test also improved from 83.4% to 87.3%. It is clear that the trainees who participated in this experiment showed relatively high scores even for the pretest, compared to the average score for Japanese speakers. Note that the trainees in STUDY 1 showed

trainee	pretest	post-test
MS01		
intelligibility	97.9	99.0
goodness	6.09	6.11
perception	97.3%	100%
MS02		
intelligibility	88.5	95.4
goodness	4.52	4.81
perception	69.6%	74.6%

Table 2: Improvement in production and perception from pretest to post-test.

62.7% and 3.36. The production ability of MS02 (who showed lower performance at pretest) improved significantly from pretest to post-test (Fig.3).

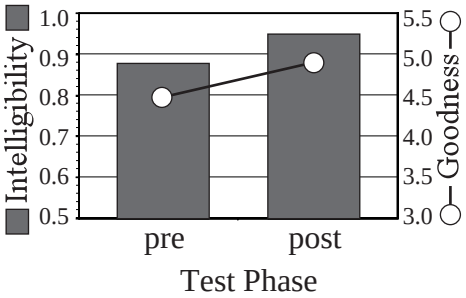


Figure 3: Intelligibility (bars) and goodness scores (circles) of MS02's pretest (pre) and post-test (post) production as judged by AE listeners.

4. DISCUSSION

Production training using visualized acoustic properties provided by spectrographic representations was found to be highly effective for improving the production ability of L2 sounds. In previous studies of perception training of /r/ and /l/, using the same materials produced by five talkers including the three talkers used in the present study, it was demonstrated that perception training also improved production ability ([7], [8]). However, after over 12,000 trials (about 20 hours), the pretest to post-test gains in intelligibility averaged only 7% increase in intelligibility. In contrast, the trainees in the present study improved 22% by the same tests after only 612 trials lasting about five hours. This suggests that the production training using spectrographic representations as feedback is far more advantageous for production learning than using perception training alone. On the other hand, the perception ability improved only 3.6% in accuracy, while it improved about 20% by perception training. Taken together, we can conclude that training in either perception or production domain modifies the trained domain, and there is a transfer to the other untrained domain, both from perception to production and production to perception. However, the amount of the change by this transfer is not as big as the change by direct training.

The results for HMM-based pronunciation quality assessment suggests that automatic speech recognition can provide helpful feedback to a language learner. A correlation of 0.82 between human and HMM ratings was found, indicating that the HMM scores are quite consistent with human scores. In addition to the per-utterance pronunciation scores generated by the recognizer, additional feedback can be obtained by considering how often utterances are recognized correctly. We found that a test set made up entirely of utterances with perfect human intelligibility scores was recognized 95% correctly. As shown in Table 1, the recognition rate falls off rapidly with decreasing intelligibility levels to near chance levels for the lowest intelligibility interval. Clearly, in an average, statistical sense, the recognition rate is related to human judgments of production intelligibility.

However, on an utterance-by-utterance basis, the HMM-based score does not always reflect an appropriate judgment: a good production will sometimes receive a poor HMM score, and vice versa. One may wonder whether feedback with such errors may diminish the effectiveness of the training. We examined this hypothesis by performing actual production training using HMM scores as feedback. Even though we had a limited number of subjects, and their pretest performance was so high that they had only little room for improvement, the result demonstrated that the training was effective. Part of the reason for the success of the training may lie in the fact that we allowed trainees to retry until they were satisfied in each trial. By this method of re-trials, i.e., producing the same word with feedback of the HMM score up to several times, subjects could ignore the occasional inadequate responses by the HMM.

Although further examinations are necessary to understand which aspects of the training procedure played the most substantial role, a production training method which uses both visualized speech representation and the speech recognition-based evaluation score may potentially be a powerful method for self-training of L2 speech segments. Furthermore, we believe that these production training methods used together with the perceptual training could possibly help L2 learners to develop new phonetic categories more robustly than conventional methods.

5. ACKNOWLEDGMENTS

We are grateful to Jesica Downs-Pruitt at University Washington, Prof Winifred Strange at University of South Florida and Prof Donna Erickson at Kanazawa University for their help and comments. We also thank Rieko Kubo and Tomoko Takada at ATR Human Information Processing Research Laboratories for running the experiments.

6. REFERENCES

1. C.T. Best, "A direct realist view of cross-language speech research", In *Speech Perception and Linguistic Experience*, Strange, W. Eds., (York Press, Timonium MD, 1995), pp.171-204.
2. J.E. Flege, "Second-language Speech Learning: Theory, Findings, and Problems", In *Speech Perception and Linguistic Experience*, Strange, W. Eds., (York Press, Timonium MD, 1995), pp.233-272.
3. D. Kewley-Port and C.S. Watson, "Computer assisted speech training: Practical considerations", In *Applied Speech Technology*, Syrdal, A., Bennett, R. & Greenspan, S. Eds., (Boca Raton: CRC Press, 1995), pp. 565-582.
4. A. Biem, E. McDermott and E. Woudenberg, "The ATR HIP Minimum Error Classifier System (MECS)", Technical Report, ATR Human Information Processing Research Laboratories, (1998).
5. H. Kawahara, "Speech representation and transformation using adaptive interpolation of weighted spectrum: Vocoder revisited", ICASSP'97, SPCH5P, p.1303-1306, (1997).
6. R.A. Yamada, W. Strange, J.S. Magnuson, J.S. Pruitt and W.D. Clarke III, "The intelligibility of Japanese speakers' production of American English /r/, /l/, and /w/, as evaluated by native speakers of American English", Proceedings of the ICSLP94, p.2023-2026, (1994).
7. R. Akahane-Yamada, Y. Tohkura, A.R. Bradlow and D.B. Pisoni, "Does training in speech perception modify speech production?", Proceedings of the ICSLP96, p.606-609 (1996).
8. A.R. Bradlow, D.B. Pisoni, R. Akahane-Yamada, and Y. Tohkura, "Training Japanese listeners to identify English /r/ and /l/: IV. Some effects of perceptual learning on speech production", *JASA*, vol. 101, 2299-2310 (1997).
9. Y. Kim, H. Franco, L. Neumeyer, "Automatic Pronunciation Scoring of Specific Phone Segments for Language Instruction", Proceedings of Eurospeech, vol. 2, 645-648 (1997).