INTERSPEECH 2021
30 August – 3 September, 2021, Brno, Czechia

Effects of voice type and task on L2 learners' awareness of pronunciation errors

Alif Silpachai[1], Ivana Rehman[1], Taylor Anne Barriuso[1], John Levis[1], Evgeny Chukharev-Hudilainen[1], Guanlong Zhao[2], Ricardo Gutierrez-Osuna[2]

[1]Iowa State University, USA
[2]Texas A&M, USA

{alif, ilucic, barriuso, jlevis, evgeny}@iastate.edu, {gzhao, rgutier}@tamu.edu

Abstract

Research suggests learners may improve their second language (L2) pronunciation by imitating voices with similar acoustic profiles. However, previously reported improvements have been in suprasegmentals (prosodic features such as intonation). It remains unclear if voice similarity applies to L2 segmentals (consonants and vowels). To address this issue, this study investigates how voice similarity facilitates awareness of pronunciation errors, a necessary step in pronunciation improvement. In two experiments, advanced L2 learners identified their pronunciation errors by comparing their production to the production of a resynthesized model voice using learners' voices as the base (Golden Speaker voice), or to an unfamiliar resynthesized voice with the same gender as the learner (Silver Speaker voice). In Experiment 1, L2 learners identified all syllables with vowel and consonant errors when comparing their production to the model voice. Their choices were compared to identifications by expert judges. In Experiment 2, learners were told how many errors the expert judges had identified before identifying the same number of errors. Results did not support facilitative effects of Golden Speaker voices in either experiment, but Experiment 2 resulted in higher identification percentages. Discussion of the challenges in self-identification of errors in relation to voice similarity are offered.

Index Terms: error identification, perception, pronunciation, golden speakers

1. Introduction

It is widely assumed that language learners are able to listen to a correct production of an utterance and identify differences between their own production and that of the model, but it is not clear how well L2 learners actually notice differences in their own production and the production of other speakers. Previous studies have suggested that L2 learners may improve their L2 pronunciation from imitating a voice that is similar to their own (henceforth a Golden Speaker voice). Probst et al. [1] reported that learners who imitated voices similar to their own (i.e., a voice produced by the same gender and having similar pitch and speech rate) improved their pronunciation more than learners who imitated voices that were dissimilar in these characteristics. It has also been shown that listening to one's voice produced in a resynthesized, native-like accent can improve L2 pronunciation. Japanese L1 speakers improved their English prosody more after practicing with their own voice that was resynthesized to match native prosody, compared to speakers who practiced with native voices [2]. More recently, Japanese learners of Italian improved their Italian intonation in three pragmatic contexts (request, order, and granting) after imitating their own utterances modified to match native prosody [3]. The effectiveness of resynthesized voices in these studies might encourage "behavioral shaping" in which learners compare their pronunciation to a model that sounds more similar to their own voice ([4, 5, 6, 7]). It remains unclear, however, whether such behavioral shaping plays a role in pronunciation of L2 segments. Improvements in L2 pronunciation observed in previous studies employing voice imitation have often been on suprasegmentals (e.g., [2, 3]).

It is possible that behavioral shaping leads to improvements in L2 pronunciation in general which encompasses the ability to detect one's pronunciation errors (domain-general hypothesis). Alternatively, the behavioral shaping may apply only to the suprasegmental domain given that segments differ from suprasegmentals in many respects (domain-specific hypothesis). If the domain-general hypothesis is true, L2 learners should have an enhanced ability to notice their L2 segmental errors, presumably a necessary step in improving their L2 pronunciation. However, if the domain-specific hypothesis is true, such ability should be not observed.

1.1. This study

To test these hypotheses, this study extended the literature to the segmental domain and investigated how well L2 learners noticed segmental pronunciation differences between their own production and that of a resynthesized model voice with correct pronunciation. Two types of synthesized model voices were used: a voice that is based on the L2 learner's own voice (Golden Speaker; GS), and a voice that was gender-matched but based on a different voice from the learner's (Silver Speaker; SS). The following research question was investigated in two experiments where less and more information about the number of errors in each sentence was provided to the learners:

To what extent do learners using the GS voice notice their errors compared to those using a SS voice?

2. Experiment 1

2.1. Participants

A total of 37 participants completed Experiment 1. One student started but did not complete the study. All were Chinese international students recruited from STEM majors at Iowa State University. Their native languages included Mandarin, Cantonese, and other regional Chinese dialects. They were randomly assigned to training groups: 18 (11 female) were exposed to the GS voice for training, and 19 (11 female) were exposed to the SS voice for training. Demographic data are summarized in Table 2. Participants were compensated $25 for two sessions.

http://dx.doi.org/10.21437/Interspeech.2021-701

Table 2: Self-reported participant demographic data.

	Gender	Age	Time in the US	L2 Proficiency (e.g. TOEFL Scores)
GS	M = 7 F = 11	19 - 30 m = 24.0	2 mos.-5 yrs. m = 2.6 yrs.	iBT 72-107 m = 90.2 IELTS 7.5 DNR (n = 3)
SS	M = 8 F = 11	20 - 30 m = 25.4	2 mos.-18 yrs. m = 4.0 yrs.	iBT 70-108 m = 90.2 IELTS 6.5 DNR (n = 7)

2.2 Stimuli generation

To generate GS and SS utterances, we asked participants to read fifty sentences aloud. The sentences were from CMU ARCTIC [5] and were selected because they contained pronunciation features predicted to be difficult for speakers of Chinese languages.

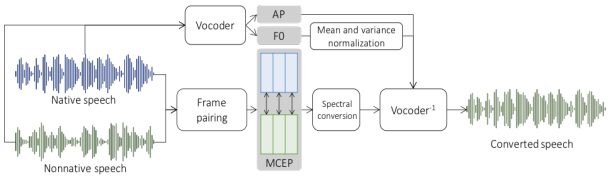

Figure 1. *Overall process of accent conversion.*

To generate the synthesized GS voice for each learner, we followed the system (Figure 1) proposed by [6]. During a training phase, we paired utterances from a reference native speaker recording with those from the nonnative learner. First, we used the STRAIGHT vocoder [7] to decompose each speech frame into three components: fundamental frequency (f_0), aperiodicity (AP), and spectral envelope, represented by a 25-dimensional vector of Mel Cepstra (MCEP). Then, we paired native and nonnative MCEP vectors based on their phonetic similarity, measured using a phonetic posteriorgram [8]. Finally, we built a Gaussian Mixture Model (GMM) that converted the native MCEP vectors into those for the L2 learner. During the test phase, we provided an utterance for the reference speaker, decomposed it into f_0, AP, and MCEP, used the GMM to convert the reference MCEP vector into the learner's MCEP vector, and recombined it with the reference f_0 (normalized to match the learner's f_0 range) and AP. The result was a GS utterance that had the linguistic content (i.e., segmental) and prosody (pitch and energy contour, and speaking rate) of the reference native utterance, but the voice quality of the learner. In Experiment 2, we further adjusted the speaking rate on the GS utterances to match the original speaking rate of each L2 learner [1].

SS voices were generated in a similar fashion, except that we mapped the reference native speaker utterances to those of another native speaker with similar gender to the learner. Thus, learners from GS and SS groups received stimuli with similar acoustic quality.

To ensure that the GS voices were heard as being significantly more similar to the L2 speaker's voice than the SS voices, we conducted a listening test using Amazon Mechanical Turk. English-speaking listeners (n = 18) rated Golden and SS voices as compared to the native Mandarin speakers' unsynthesized voices using a seven-point scale, where 1 is "definitely the same speaker", and 7 is "definitely a different speaker". The GS voices had a mean rating of 3.77 (95% confidence interval: [3.20; 4.34]) whereas the SS was rated 5.46 (95% confidence interval: [4.84; 6.08]). Thus, the GS voices were rated as significantly more similar to the original voices than the SS voices were ($p < .05$).

2.3 Procedure

Each participant was randomly assigned to either the GS model (n = 18) or the SS model (n = 19). The native voices used with the participants' voices to synthesize GS and SS model voices were from speakers of American English. Upon arrival, the participants provided informed consent and completed a demographic questionnaire. Participants completed two sessions: an initial recording phase and a mispronunciation detection phase. In the initial recording, each subject read the 50 sentences preselected from ARCTIC to provide an accurate model for speech synthesis. All recordings were made in a quiet room using a Samson C03U microphone. Each sentence was displayed on a computer screen and was controlled by the experimenter using an HTML program. Productions were recorded using Audacity.

Ten of the 50 sentences were selected for the mispronunciation detection phase based on pronunciation errors in the L2-ARCTIC corpus [9]. These 10 sentences were 12-16 syllables long and had an average of 4-6 pronunciation errors by the native Chinese speakers in L2-ARCTIC.

To determine each participant's actual pronunciation errors on the 10 sentences, four phonetically-trained expert judges convened and listened to all 370 experimental sentences (37 participants x 10 sentences), identifying the syllables containing at least one mispronunciation. At least three of four judges had to agree for syllables to be marked as mispronounced. Errors were identified as part of syllables rather than phones because of the lack of 1-1 correspondence between phones and letters.

In the mispronunciation detection phase, participants listened to each of their 10 sentences and compared them to the same sentences in either the GS or SS condition. They identified errors by clicking on syllables where they noticed pronunciation differences between their recording and the model voice. For example, the sentence "*For the twentieth time that evening, the two men shook hands*" (used as an example sentence in the study, as in Figure 2) was divided into syllables that could be clicked to identify where they heard differences between their speech and the Golden or SS voice.

To become familiar with the task, they listened to an example sentence in which they clicked the syllables that had differences between the example recording and the model voice. The example recording was not of their own production, but rather was randomly selected from the L2-ARCTIC corpus. Once they completed the example, they listened to their 10 sentences as many times as needed and identified mispronunciations they heard. They were not told how many mispronunciations to identify in each sentence. After finishing a sentence and moving to the next, they were not allowed to go back and change their answers.

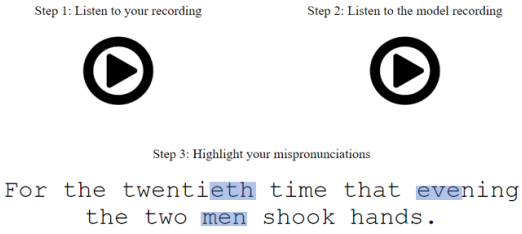

Figure 2. *User interface for the experiments. When participants moved the cursor over multisyllabic words, each syllable was highlighted separately.*

2.4 Analysis

The dependent variables included precision, calculated as the number of mispronunciations that both the participant and the judges identified (i.e. the ones where the participant concurred with the judges, or, in other words, the participant identified a true mispronunciation), divided by the total number of mispronunciations that the participant identified (both true and incorrectly perceived); and recall, calculated as the number of mispronunciations that both the participant and the judges identified, divided by the total number of mispronunciations that the judges identified (both those also identified by the participant, and those that the participant skipped).

For each dependent variable, two nested linear mixed-effects regression models were fitted to the data. Both models included random effects for sentence and participant. The first model was an intercept-only model, and the second model added Group as a fixed effect. Gains in goodness of fit of successive models were evaluated by the likelihood ratio test. Fixed-effect parameters of the full model were used to estimate means of the dependent variables for the control and treatment group. Wald estimates of the confidence intervals (CIs) for means were then derived from the model. The data were also examined by estimating a Bayes factor using Bayesian Information Criteria [10], comparing the fit of the data under the null and the alternative hypothesis.

2.5 Results

The question in Experiment 1 was whether learners in the GS group—the one with similar voice quality to their own—would more successfully identify deviations in their pronunciation than learners in the SS group who listened to a synthesized native-speaker model voice matched only for gender.

On average, precision was 0.51 (95% confidence interval: [0.41; 0.61]) in the GS group, and 0.56 (95% confidence interval: [0.46; 0.66]) in the SS group. The precision score reflected the number of true errors identified over the number of true and false errors identified by the participants. Adding Group as a fixed effect to the model did not significantly improve the fit of the model to the data: $\chi^2(1) = 0.82$, $p = 0.36$. An estimated Bayes factor (null/alternative) suggested the data provided "strong" evidence against an effect of Group on precision [11]. Participants with the GS voice did not show greater precision on error identification than those hearing the SS voice. On average, recall was 0.25 (95% confidence interval: [0.19; 0.30]) for both groups. Adding Group as a fixed effect to the model did not significantly improve the fit of the model to the data: $\chi^2(1) = 0.05$, $p = 0.83$.

2.6 Discussion

Participants in the GS group would were predicted to more successfully identify deviations in their pronunciation than participants in the SS group. There is no evidence that this was the case. Ultimately, the learners struggled to notice pronunciation differences between their own and the model voices, despite their relatively advanced L2 proficiency.

The failure to find such effects may have been due to an unexpected difficulty in the task, that is, identifying differences between the model voice and their own voice. The difficulty, in turn, may have decreased the participants' motivation to perform the task. There are several possible explanations for the difficulty. First, the assumption that learners have the ability to identify errors is misplaced. The Perceptual Assimilation Model [12] suggests that hearing the difference between two L2 sounds may be challenging when L2 learners perceive these sounds as being acceptable productions of one phoneme as opposed to two separate phonemes which presumably are not yet part of the learners' phonologies. This may have been the case with the participants in this study, who may not have noticde errors beyond those that they expected because they lacked training in careful listening to speech.

To minimize task difficulty and foster extrinsic motivation, we conducted Experiment 2 with a more explicitly designed task and provided a monetary bonus for participants who performed well. This modification was undertaken to determine whether differences among the model voices would manifest under more favorable conditions.

3. Experiment 2

We conducted Experiment 2 to test the hypothesis that task difficulty and the lack of motivation obscured the relationship between voice type and error identification. Thus, instead of using an open-ended nature task like in Experiment 1, the task was scaffolded by telling each participant the number of errors they were to find for each sentence. To increase motivation, monetary bonuses were given to participants with high percentages of accurate identifications. Effects of motivation, both intrinsic and extrinsic, on L2 learning and performance have been explicitly stated in previous research (e.g., [13, 14]).

3.1 Participants and Stimuli

Most participants (n = 26) in Experiment 1 returned to the laboratory four months later for Experiment 2. They were assigned to the same conditions as in Experiment 1. Three participants who had few mispronunciations in Experiment 1 were not called back. Eight other participants did not respond to the recruitment email. All subjects were again paid for their participation. Two participants in the GS condition were excluded from the final analysis because of extremely quick response times (less than five minutes total) and scores that were much lower than other participants, indicating inattention to the stimuli and task. The stimuli were the same as those used in Experiment 1 except that the GS stimuli were also matched to the learners' speech rates, as suggested in [1].

3.2 Procedure

After giving consent, participants followed the same procedures as Experiment 1 with one exception: they were told that expert judges had detected a number of pronunciation errors in their sentences, and that they had to identify the same number of errors. To increase motivation, extra money was offered for

more successful identifications. Those identifying less than 60% of the errors identified by the expert judges received the base rate of $15, while identifications from 60-69% received an additional $5. Those who identified 70-79% received an additional $10, and those identifying more than 80% of errors received $15 more. The most anyone actually received was $5 additional (n = 6) as no one identified more than 66% of errors.

3.3 Results

As in Experiment 1, the subjects listening to the GS and SS voices were equally successful in identifying errors. Given that the number of selected syllables was fixed and equal to the number of syllables marked in the experts' annotations, precision was by definition equal to recall. On average, precision/recall was 0.51 (95% confidence interval: [0.42; 0.60]) in the GS group, and 0.49 (95% confidence interval: [0.40; 0.58]) in the SS group. Adding Group as a fixed effect to the model did not significantly improve the fit of the model: $\chi^2(1) = 0.14$, $p = 0.7$. An estimated Bayes factor (null/alternative) suggested that the data were 23:1 in favor of the null hypothesis. In other words, the data can be interpreted as decisive evidence against an effect of Group on precision.

Although results did not show an effect of voice, subjects in both groups doubled their identification of errors when they were told how many errors to identify. This indicates that scaffolding the task promoted greater identification accuracy by helping subjects understand the extent of the noticing that is expected. It is unlikely that the increased identification rates were due to greater familiarity with the sentences. The second time the participants identified errors was four months after the first, the amount of time spent on identification was approximately 10 minutes average (Experiment 1) and 20 minutes (Experiment 2), and they received no feedback on their identification accuracy in either experiment.

4. Discussion

We extended the literature on the use of a Golden Speaker voice in L2 pronunciation learning to the segmental domain. The results of both experiments indicated that the GS voice was not superior to the SS voice when it came to promoting identification of pronunciation errors. Thus, the GS > SS hypothesis was not supported in either experiment. These findings were unexpected given the previous evidence that voices that were prosodically similar to the learner's voice were superior at promoting pronunciation improvement [1, 2, 3, 4].

Comparing a GS voice to one's own voice did not enhance the ability to identify mispronounced segments. However, it remains unclear whether the approach was ineffective because it was in the segmental domain as opposed to the suprasegmental domain. To address this issue, future research should include segmentals and suprasegmentals. The ineffectiveness of the approach might have been influenced by the design of the study. Unlike previous GS studies that reported improvements in production of L2 suprasegmentals, this study did not include voice imitation. It is possible that exposure to a GS voice is more effective for improving L2 pronunciation when learners imitate the voice compared to when they only listen to the voice.

Surprisingly, the task we used was unexpectedly difficult. Even in Experiment 2, participants were told how many errors to identify and were encouraged to be extra careful by offering bonus payments, but only identified one-half of their true errors. Thoughtful scaffolding promoted greater identification accuracy, but the success rate suggests that comparing their own production to a model voice remained challenging. Pronunciation practitioners assume that high-proficiency language learners can compare their production to a model voice and draw from that comparison to improve their production [15, 16], but our results do not support this.

This may be because learners, having noticed one or two errors, do not notice anything else. The sentences all averaged between four and six errors in the L2-Arctic corpus. This could be too many to identify in an open-ended task. Second, some errors may be simply hard to notice. Research on L2 perception has shown that L2 production errors that are similar to, yet different from, L1 phonemic categories are especially challenging to notice [17]. The Perceptual Assimilation Model [12] demonstrates that such differences is particularly challenging because L2 learners do not possess the perceptual categories needed to hear such differences. Applied to this study, it is likely that for some errors, participants did not identify errors because they did not hear them.

Third, identifying errors in sentences with up to six errors may have been too cognitively demanding for participants when provided only implicit feedback from a model voice. Identifying errors may have been difficult because the participants had to both identify errors and simultaneously ignore other aspects of the spoken signal [18]. For future studies, reducing competing speech signals may allow language learners to become better listeners to their own speech.

This study of pronunciation awareness suggests directions for future research. The first is to look at how L2 listeners can be helped to become more aware of their errors when comparing their speech to that of a model voice. This study and a number of others suggest that reducing cognitive load increases the possibility for a more successful task completion for fluency [19] and for comprehension [20]. A second direction involves a detailed analysis of the kinds of errors that L2 speakers were successful in identifying. The errors could be compared to a perception test to look at whether the L2 speakers had more trouble hearing errors that they also did not identify, indicating their difficulty was related to perception challenges with particular sounds. (A reviewer asked whether aperiodicity alone could sufficiently characterize idiosyncratic voice quality. We leave this issue open for future research with more participants.)

5. Conclusions

This study examined the effect of voice similarity on speakers' awareness of their pronunciation errors. In both experiments, participants performed similarly when comparing their production to similar and non-similar voices. When the identification included no information about what the participants were to find, their identification of true errors was worse than when they were given guidance about the number of errors to identify. These results suggest that better identification of pronunciation errors is dependent upon better conditions for noticing. Our results must leave open the question of whether GS voices are better for noticing. Our Mechanical Turk experiment showed that GS utterances were closer than SS to original L2 utterances, and voices that are even more similar may result in better noticing.

6. Acknowledgements

This study was funded by National Science Foundation grants 1623622 and 2016984.

7. References

[1] K. Probst, Y. Ke, and M. Eskenazi, "Enhancing foreign language tutors—in search of the golden speaker," *Speech Communication,* vol. 37, no. 3-4, pp. 161-173, 2002.

[2] K. Nagano and K. Ozawa, "English speech training using voice conversion," *Proceedings of the First International Conference on Spoken Language Processing*, vol. 90, pp. 1169-1172, 1990.

[3] E. Pellegrino and D. Vigliano, "Self-imitation in prosody training: a study on Japanese learners of Italian," *Proceedings os SLaTE* (pp. 53–57), 2015.

[4] D. Felps, H. Bortfeld, and R. Gutierrez-Osuna, "Foreign accent conversion in computer assisted pronunciation training," *Speech Communication*, vol. 51, no. 10, pp. 920–932, 2009.

[5] R. Gutierrez-Osuna and D. Felps, "Foreign accent conversion through voice morphing," Department of Computer Science and Engineering, Texas A&M University, Tech. Rep., 2010.

[6] S. Aryal, D. Felps, and R. Gutierrez-Osuna, "Foreign accent conversion through voice morphing," in *Interspeech*, 2013, pp. 3077-3081.

[7] C. Watson and D. Kewley-Port, "Advances in computer-based speech training: Aids for the profoundly hearing impaired," *Volta Review*., vol. 91, pp. 29–45, 1989.

[8] CMU_Arctic speech synthesis databases. Retrieved from http://www.festvox.org/cmu_arctic/

[9] S. Ding, C. Liberatore, S. Sonsaat, I. Lučić, A. Silpachai, G. Zhao, E. Chukharev-Hudilainen, J. Levis, and R. Gutierrez-Osuna, "Golden speaker builder--An interactive tool for pronunciation training. *Speech Communication*, vol. *115*, pp. 51–66, 2019.

[10] H. Kawahara, A. D. Cheveigné, H. Banno, T. Takahashi, and T. Irino, "Nearly defect-free F0 trajectory extraction for expressive speech modifications based on STRAIGHT," *Proceedings of Interspeech* pp. 537-540, 2005 Retrieved from isca-speech.org/archive/interspeech_2005/i05_0537.html

[11] T. J. Hazen, W. Shen, and C. White, "Query-by-example spoken term detection using phonetic posteriorgram templates," *Proceedings of IEEE Workshop on Automatic Speech Recognition & Understanding*, pp. 421-426, 2009.

[12] G. Zhao, S. Sonsaat, A. O. Silpachai, I. Lučić, E. Chukharev-Hudilaynen, J. Levis, and R. Gutierrez-Osuna, "L2-ARCTIC: A non-native English speech corpus," *Proceedings of Interspeech* pp. 2783-2787, 2018. Retrieved from https://lib.dr.iastate.edu/engl_pubs/226/

[13] E. J. Wagenmakers, "A practical solution to the pervasive problems of p values," *Psychonomic Bulletin & Review, vol. 14, no.* 5, pp. 779-804, 2007.

[14] H. Jeffreys, *The Theory of Probability*, Oxford: Oxford University Press, 1998.

[15] C. T. Best and M. D. Tyler (2007). "Nonnative and second-language speech perception: Commonalities and complementarities," In O-S. Bohn & M. Munro (Eds.), *Language Experience in Second Language Speech Learning: In Honor of James Emil Flege,* pp. 13-34, 2007, Amsterdam*:* John Benjamins.

[16] K. Saito, J. M. Dewaele, M. Abe, M., and Y. In'nami, "Motivation, emotion, learning experience, and second language comprehensibility development in classroom settings: A cross-sectional and longitudinal study," *Language Learning*, vol. *68, no.* 3, pp. 709-743, 2018.

[17] H. Sumiyoshi and C. Svetanant, "Motivation and attitude towards shadowing: learners' perspectives in Japanese as a foreign language," *Asian-Pacific Journal of Second and Foreign Language Education*, vol. 2, no. 1, pp. 1-21, 2017.

[18] A. Dlaska and C. Krekeler, "Self-assessment of pronunciation," *System*, vol. 36, n. 4, pp. 506-516, 2008.

[19] A. Dlaska and C. Krekeler, "The short-term effects of individual corrective feedback on L2 pronunciation," *System*, vol. 41, no. 1, pp. 25-37, 2013.

[20] J. E. Flege, "Second language speech learning: Theory, findings, and problems," In W. Strange (Ed.), *Speech Perception and Linguistic Experience: Issues in Cross-Language Research,* pp. 233-277. York: York Press.

[21] E. W. Healy, M. Delfarah, E. M., Johnson, and D. Wang, "A deep learning algorithm to increase intelligibility for hearing-impaired listeners in the presence of a competing talker and reverberation," *The Journal of the Acoustical Society of America, vol. 145*, no. 3, pp. 1378-1388, 2019.

[22] L. Strachan, S. Kennedy, and P. Trofimovich, "Second language speakers' awareness of their own comprehensibility: Examining task repetition and self-assessment," *Journal of Second Language Pronunciation*, vol. 5, no. 3, pp. 347-373, 2019.

[23] P. Robinson, P., S. C. C. Ting, and J. J. Urwin, "Investigating second language task complexity," *RELC Journal*, vol. 26, no. 2, pp. 62-79, 1995.