Applying adaptive recognition of the learner's vowel space to English pronunciation training of native speakers of Japanese

William L. Martens[1] and Rui Wang[2]

[1]University of Sydney
[2]Deakin University

Abstract When native speakers of Japanese are taught English as a second language, there are difficulties with their training in pronunciation of American English vowels that can be ameliorated though adaptive recognition of the learner's vowel space. This paper reports on the development of an online Computer-Assisted Language Learning (CALL) environment that provides Japanese learners with customized target utterances of 12 single-syllable words that are synthesized according to an adaptive recognition of the learner's vowel space. These customized target utterances provide each learner with examples of each of 12 American English monophthongs in consonant-vowel-consonant (CVC) context in order to sound as if they had been uttered by the learners themselves. This adaptive process was incorporated into a successfully developed tool for Computer-Assisted Pronunciation Training (CAPT) which gave more appropriate pronunciation targets to each learner, rather than forcing the learners to attempt to match the formant frequencies of their own utterances to those of the target utterances as produced by a speaker exhibiting a different vowel space (i.e., a speaker with a different vocal tract length).

1 Introduction

When native speakers of Japanese are taught a second language (L2), adult learners typically have difficulty mastering certain phonemic contrasts between vowels in the target language (L2), especially if fewer vowel sounds are used in their native language (L1). For example, when learning English as a Second Language (ESL), native speakers of Japanese must learn to overcome difficulties in identifying each of the L2 vowels, as well as learning to produce those L2 vowels with confidence in their pronunciation. The results of a closely related study [1] that were published fifteen years ago showed that identification training of native speakers of Japanese yielded improved skills in pronouncing American English (AE) vowels that typically are difficult for native speakers of Japanese to distinguish. That study [1] demonstrated the effectiveness of a high-variability identification training procedure in improving native Japanese identification and production of five AE mid and low vowels exhibiting contrasts between vowel sounds that are exemplified in the following five AE words: "bad, bod(y), bud, bawd, bird."

In contrast to other popular approaches to Computer-Assisted Pronunciation Training (CAPT), distinct advantages are observed when using an approach based upon identification training with carefully selected pronunciation examples. An alternative popular approach is that based upon the display of visualized acoustic properties, such as those shown by a sound spectrogram [2]. Providing such visual feedback in L2 pronunciation training has been observed to exhibit two substantial disadvantages (as summarized in [2]):

"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 previously mentioned study [1], employing high-variability identification training for native Japanese ESL students, produced clear results supporting the current approach using training that provides L2 sound examples rather than visual feedback. Those results can be summarized briefly as follows: Before and after a six-week identification training period, performance in production of five AE mid and low vowels was assessed for 54 native Japanese participants, all ESL students at the University of Aizu. The rates of distinct pronunciations of those five vowels were measured through blind assessment by AE native speakers. The results of that study [1] revealed that identification training with feedback improved the students' production of the target AE vowels. In the current study, an alternative method of providing target utterances for the AE vowels was employed in an effort to circumvent a problem in pronunciation training that stems from variation in production between Japanese participants exhibiting differences in their vowel spaces (i.e., differences in the range of frequencies over which the first two vocal formants varied for those participants). This paper describes initial attempts at

providing Japanese ESL learners with customized target utterances of 12 short words that were synthesized according to an adaptive recognition of the learner's vowel space. These 12 words (listed below in Table 1) featured the five AE mid and low vowels of the previous study [1], along with words featuring seven additional vowels. The 12 AE vowel sounds were those for which formant frequency data are available for large groups of native speakers (e.g., results based upon the 90 AE speakers sampled in [3]).

IPA symbol	'CVC' word examples
i	heat
ɪ	hit
ɛ	bet
æ	hat
ɑ	hot
ʌ	hut
ɔː	hawk
ʊ	hook
u	hoot
ɝ	bird
oʊ	boat
eɪ	bait

Table 1. Symbols of the International Phonetic Alphabet (IPA) used to identify the vowel sounds that presented in the online sessions described in this paper (see https://www.internationalphoneticassociation.org). Note that the IPA symbols for the last two words indicate that they are diphthongs [eɪ, oʊ] rather than monophthongs, but these two are regarded as "smaller" diphthongs that involve less spectral movement than "true" diphthongs. Indeed, in the examples produced for the current study, the central vowel sound in these two 'CVC' words was pronounced with relatively constant formant frequencies, consistent with the h-V-d utterances reported by Hillenbrand, et al. [3].

It is clear from the range of variation in formant frequencies typically observed for male and female AE speakers (e.g., as directly compared in [3]), that there are no single target values for the formant frequencies of each vowel sound that should be regarded as the "correct" values. Of course, native AE listeners do learn to adapt to the vowel space of a given speaker so that individual AE vowel sounds can be readily identified in the context of other AE vowel sounds produced by the same speaker (as was clearly shown in [4]). This result can be understood from the upward shift in formant frequencies that characterizes the range of vowel sounds typically associated with a decrease in the physical size of the speaker (as shorter vocal tracks exhibit higher vocal formant frequencies).

To be perfectly clear, it should be emphasized here that no estimate of vocal-tract length (VTL) is required or attempted by the algorithm employed here to adaptively adjust pronunciation examples to the learner's vowel space. Although ample evidence exists (e.g., [5]) that parameters describing a given speaker's vowel space are highly correlated with that speaker's VTL, a more direct approach to vowel space normalization operates only upon parameters of the audio signals (i.e., those derived from captured speech, and those manipulated in speech sound synthesis and/or modification). Indeed, the observed pattern of formant frequencies (identified by speech signal analysis) can be used to predict VTL within about a centimeter [5], with errors in predicted length of less than a few percent (within the normal adult VTL range, with lengths extending from approximately 13 cm to 20 cm).

General details of the audio signal processing involved in such approaches are given in the Methods section of this paper. At the outset, it is more important here to present the concept underlying the adaptive approach to computer-assisted training in L2 pronunciation. The reader should note that the goal of this paper is not to promote a specific CAPT application; indeed, the goal is rather to promote general awareness of the need for personalized training, and to encourage a more thoughtful response to this need. Indeed, it should be asked whether there is indeed a fundamental need for such systematic pronunciation training. Why should native Japanese ESL learners work so hard on proper AE pronunciation? Perhaps it would be better to design a Computer-Assisted Language Learning (CALL) environment that would attempt to make Japanese ESL learners more comfortable with their own pronunciation of English. In this respect, it is thought that CAPT applications featuring more personalized training are more respectful of the individual's native cultural traits, treating individual differences between L2 speakers in a more responsible manner [6].

Rather than focussing training upon "native-like pronunciation" it has been argued that ESL learners' should focus upon communicative competence [7], such as skilled communication on particular tasks (i.e., within particular contexts). Instead of teaching AE pronunciation for its own sake, scenarios can be presented in which communication problems are addressed that potentially can stem from confusion that results from predictable pronunciation difficulties.

Such identified pronunciation difficulties are targeted through the novel CAPT approach taken here, fixing on the goal to serve clearer communication. This positive motivation is in strict contrast with more typical negative approaches that focus on pronunciation for its own sake. The current adaptive approach was motivated in part by a rejection of the punitive approach that could be taken in more strict pronunciation training, wherein L2 pronunciation is corrected relative to an external reference. The psychological ramifications of the punitive approach are objectionable as they lead to negative consequences such as undermined confidence, and a disdain for English that is spoken with a foreign accent [6].

2 Background

Since the publication in 1960 of Broadbent and Ladefoged's seminal study of contextual effects on vowel identification [8], it has been well established that human listeners adapt to the vowel space of the speaker to whom they are listening. Their results revealed the likelihood of a shift in the identification of a vowel sound embedded in a short consonant-vowel-consonant (CVC) utterance that results when the spoken context is manipulated. The classic example of this phenomenon is found when a CVC that is usually heard as the word "bet" is preceded by a sentence exhibiting generally higher formant frequencies, in which case that same CVC is more likely to be identified as the word "bit."

Because the vowel sound in the bit-CVC has a lower first formant frequency than the vowel sound in the bet-CVC, the latter is effectively shifted to occupy the position of the former in the vowel space of the speaker with the higher formant frequencies (as would be observed for a speaker with a somewhat shorter vocal tract). Despite the general awareness of this well-known phenomenon, pronunciation training for L2 learners has typically disregarded the individual's normal vowel space when guiding those learners to produce L2 vowel sounds that do not occur in their native language. For example, the CVC demonstrating a somewhat higher first formant frequency than that found in the word "bet" is found in the AE pronunciation of the word "bat." This discussion is focussed upon a potential problem that can be encountered when a native speaker of Japanese is guided to produce the word "bat" as spoken by a speaker with generally lower formant frequencies than those of that Japanese L2 learner.

Especially problematic is the case in which various Japanese L2 learners of AE pronunciation are provided with a single utterance of the word "bat" as a pronunciation example out of context, after which they then are instructed to produce that same sound. In such cases, it is inevitable that some (physically smaller) L2 learners will be attempting to match an utterance provided by a (physically larger) L1 speaker exhibiting lower formant frequencies, relative to the higher formant frequency values more appropriate to the lengths of their vocal tracts. That is, the L2 learner with a shorter vocal tract generally exhibits higher formant frequencies than those of a longer-VTL AE speaker providing the target utterance.

To make the point that such AE pronunciation training can serve to clarify communication for the L2 learner, a concrete example is offered here. As the difference in AE pronunciation of the words "pad" and "pod" make a meaningful distinction that can be embedded into a single conversational example, it is straightforward to demonstrate an L2 learner's need to clearly and distinctly produce those two vowel sounds. A conversation that was featured in online training sessions focussed upon two consumer products for which Japanese language pronunciation differs in an interesting way from their AE pronunciation: As shown in Figure 1, the words "iPad" and the iPod" are easily confused since the typical pronunciation of "iPad" by native speakers of Japanese is very similar to the typical pronunciation of "iPod" by AE native speakers (i.e., both are produced using the /ɑ/ vowel sound typical of the AE pronunciation of the word "pot"). The AE pronunciation of the "iPad" product name is produced using the /æ/ vowel sound that is not used in the Japanese language, as it features the vowel sound typical of the AE pronunciation of the word "bat"). In contrast, the vowel sound appearing in the typical pronunciation of "iPod" by native speakers of Japanese is more similar to the /o/ sound typically produced by AE native speakers in pronouncing the word "boat." This is the crux of the issue here, since that pronunciation could be said to refer to a non-existent product, which as a possibility appears as the encircled "???" in Figure 1.

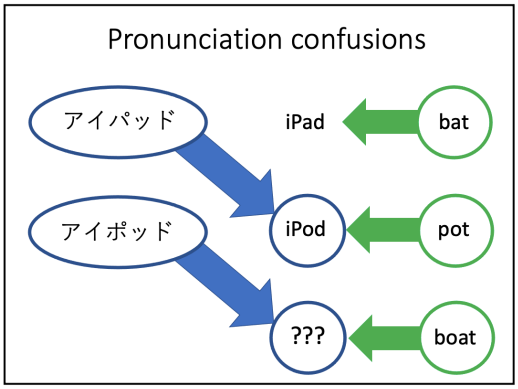

Fig. 1. Diagram shown to participants online while explaining the mismatch between Japanese loanword pronunciation (written here in *katakana*) and that of AE native speakers for the "iPad" and iPod" consumer products, illustrating the potential confusions that can result when the Japanese L2 learner does not switch from the typical pronunciation of Japanese loanwords to that typical of AE native speakers (contrasting "new" vs. "similar" phones, as described in [9]).

Why is this example relevant to the proposed adaptive recognition of the learner's vowel space to English pronunciation training of native speakers of Japanese? First, the reader is reminded that it has been well established that human listeners adapt to the vowel space of the speaker to whom they are listening [4]. Furthermore, results of research on identification training for non-native vowels [10][11] have shown that training should include full sets of vowels rather than focus only upon vowels presenting difficult phonetic contrasts (such as those exhibited by the 5 vowels presented in [1]). In the more recent study [10], the influence of training set sizes was shown for both native Japanese ESL learners and Korean ESL learners. The concept that is applied in the currently proposal CAPT approach is to make sure that the difficult phonetic contrasts are present as a subset in full sets of vowels (providing context larger than just the few difficult vowels). It is hoped that this approach will be appreciated as a complement to other CAPT systems, such as those employing automated speech recognition to provide feed-back to the Japanese ESL student [12].

3 Methods

This work described in this paper took an approach to the development of a novel online Computer-Assisted Language Learning (CALL) environment following the paradigm termed 'Research Through Design' by Zimmerman, et al [13]. Accordingly, the qualitative research methods employed called for no simple experimental studies; rather, the methods used here entailed an iterative design process that is commonly used during the initial stages of development of a CALL environment. To put this approach in context then, it is pointed out here that such 'Research Through Design' (RTD) has the goal of producing an artefact rather than producing scientific support for particular conclusions that might be drawn from the results of studies that directly test experimental hypotheses. The CALL environment developed through the RTD process is then regarded as the desired artefact, the validation of which can establish the designed system itself as the valued research output of the development study. The four steps of the RTD process can be described as follows:

Grounding — an investigation to gain multiple perspectives on the envisioned system and its associated problems.

Ideation — the generation of many possible different solutions to the problems.

Iteration — a cyclical process of refining the system concept with increasing fidelity.

Reflection — the critical evaluation of the created artefact not as a solution to particular problems, but as a means of determining whether the artefact satisfies the needs of the envisioned system.

The following section of this paper provides an overview of the results of that iterative design process that was employed to create a system that was truly satisfying to the user both in terms of the user experience and the ultimate outcome observed as improved English language pronunciation. This overview is based upon results observed for nine ESL students who were native speakers of Japanese, which began with the introduction of AE vowels to those ESL students in online sessions hosted by a native-AE-speaking instructor. Students were engaged in English language conversation that focussed upon pronunciation of AE vowel sounds. Through pictures depicting conversational scenarios, contrasts between the vowel sounds were discussed with reference to the set of 12 words listed in Table 1. Rather than introducing the symbols of the International Phonetic Alphabet (IPA) that can be used to identify these vowel sounds (as shown in Table 1), only the 12 exemplary words that are listed in Table 1 were actually included in the conversation with the learner (c.f., [14]). During the course of the session, audio samples of the learner's speech sound were submitted to analysis to find the range of formant frequencies typically produced, so that the target utterances for the learner's pronunciation practice could be synthesized according to a shift in the normalized formant frequencies of AE vowel sounds relative to the learner's vowel space. This preparatory step required the native Japanese ESL learner to pronounce repeatedly the five syllables that are spelled in *romaji* (roman characters) as "ha hi hu he ho," corresponding to the five Japanese syllabic symbols usually written using the following *katakana* characters:

"ハ ヒ フ ヘ ホ"

These five syllables were also recorded in a Japanese language sentence context in order to provide more definitive evidence for the area covered by the speaker's vowel space. These sentences were always of the form exemplified by the following sentence (written here in *romaji*): "Kore wa *ha*ba." Note that the *ha* syllable was varied between the five vowel sounds (listed above) between each of the produced sentences. Figure 2. shows the spectrogram resulting from a time-variant analysis of a recorded speech sample employing Linear Predictive Coding (LPC) to derive an all-pole (purely recursive) filter for each time frame [15]. The spectrogram shows the magnitude response of those filters for each time frame, using the colormap shown on the right of the plot to code the observed dB magnitude over time and frequency.

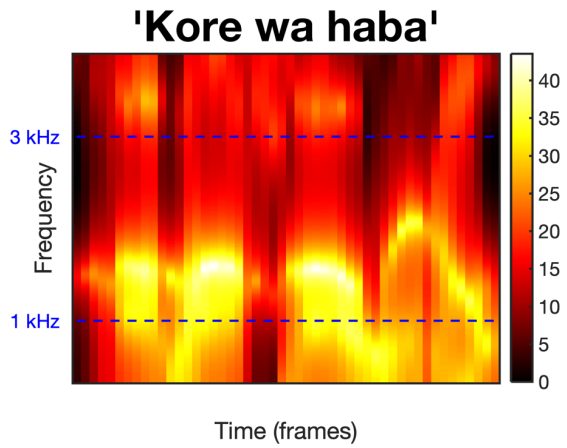

Fig. 2. Time-varying Linear Predictive Coding (LPC) analysis results for the utterance of the sentence appearing in the graph's title. For each frame (of 20-ms duration), a 30th-order all-pole filter was computed that allowed for the identification of the first few formant frequencies of the speaker's recorded speech sample. It is the magnitude response of those filters over the indicated range of frequency that was used to construct the spectrogram pictured here (using the 'hot' colormap shown on the right of the plot to code the observed dB magnitude).

It is beyond the scope of this paper to provide details of the audio signal processing that underlies the adaptive analysis of a learner's speech samples that enables further processing of target speech samples to match the learner's vowel space. Suffice it to say that the formant frequencies exhibited by the LPC-based filters can be made to match those of a targeted AE vowel as prescribed for the individual learner (i.e., as if properly produced in the context of other vowel sounds produced by each Japanese ESL student). An example of how two non-native vowel sounds can be taught in this context is shown diagrammatically in Figure 3.

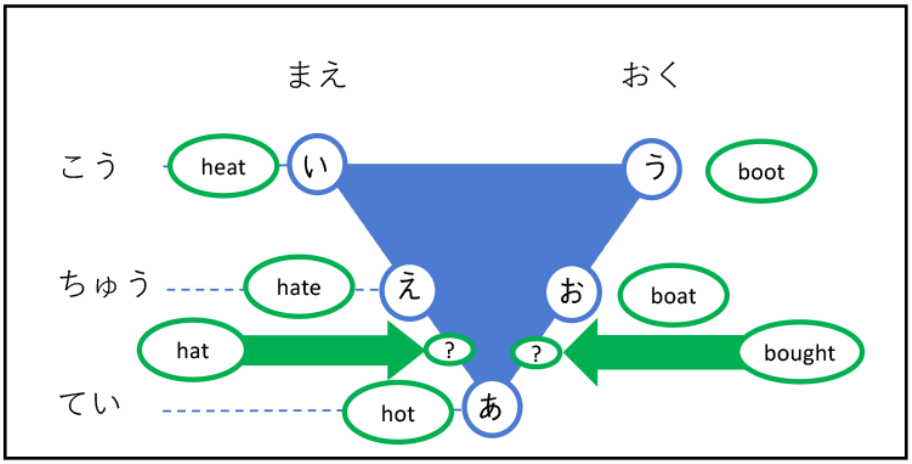

Fig. 3. Diagram showing Japanese vowel space provided to participants online to show five English-language words as examples of the five vowel sounds occupying Japanese vowel space, with an additional two English-language words that incorporate non-native exceptional cases that require the production of "new" (rather than "similar") phones [4]. The isolated vowel sounds of Japanese language are denoted by the corresponding *katakana* characters while the non-native (exceptional) vowel sounds are denoted by encircled "question mark" characters, as these phones have no proper *katakana* characters. The terms describing articulation of the mouth and tongue are added in the margins using their spelling in *hiragana*, which along the horizontal "backness" dimension correspond to the English-language terms for "front" and "back" vowels (with the corresponding terms in Japanese denoted as "まえ" and "おく").

Figure 3 here presents a diagram that was provided to Japanese ESL students online to show five English-language words as examples of the five vowel sounds occupying Japanese vowel space, with the addition of two English-language words that incorporate non-native vowel sounds as exceptional cases, which required the production of "new" (rather than "similar") phones [9]. By teaching that the two non-native vowel sounds are situated closely between vowel sounds that are "similar" to the native vowel sounds of Japanese, the ESL student can be directed to articulate these non-native sounds in a way that is "midway" between the adjacent native sounds. It was found, indeed, that this approach works best if the examples of the non-native vowel sounds are produced as if spoken by the ESL student (and so occupied expected locations within the student's own vowel space). Although no experimental test was executed to generate scientific data to document this finding, the final "reflection" step of the adopted RTD process included a critical evaluation of the CAPT system that is summarized in the following section of this paper.

4 Results and Discussion

As explained above (in this paper's Methods section), the four steps of the RTD process were followed to produce and validate an artefact, which was the designed system itself. This artefact was developed iteratively, not as a solution to particular problems, but as CALL environment that satisfies the general needs of the envisioned system. To begin with, and overall appraisal of the learning experience provided by the system will be presented.

During the online lessons with the introduced the CAPT system, the first nine participants reported that they greatly enjoyed the lessons. They reacted to the synthesized utterances with some amusement, since they were surprised to hear the unfamiliar (non-native) utterances in what sounded like their own voices. Several participants also commented that by listening to the utterances, as compared to listening only to the instructors' voice, they felt it was easier for them to "hit the target" for pronunciation improvement. Thus, there was an unsolicited validation of the development proposal, which was the following: Training in the pronunciation of AE vowels should be enhanced though adaptive recognition of the ESL learner's vowel space.

The reader might be interested to examine the different vowel spaces that were observed during the initial training sessions of the nine Japanese ESL students who participated in this development study. Two examples of the observed vowel spaces are presented in the upper panel of Figure 4, which plots formant frequencies for a a relatively large male speaker (diamond plotting symbols) and a relatively small female speaker (circular plotting symbols). Only utterances of five native Japanese syllables were analyzed for these two speakers here, those syllables that can be spelled in *romaji* as "ha hi hu he ho," and are alternatively written using the *katakana* characters "ハ ヒ フ ヘ ホ." The formant frequencies of these two native Japanese speakers can be compared with the mean formant frequencies characterizing male and female AE vowel spaces plotted in the lower panel of Figure 4. These formant frequencies are plotted separately for a group of 45 male and 48 female native AE speakers.

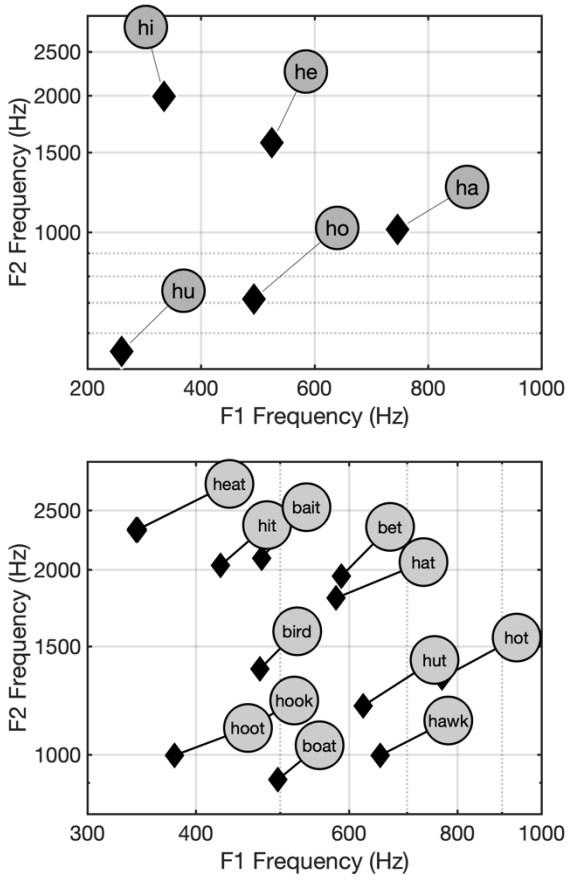

Figure 4. Upper panel: An example of the typical vowel spaces characterizing two native speakers of Japanese, an individual male (diamond symbols) and female (circular symbols). The plot shows individual formant frequencies derived from recorded h-V utterances of two of the nine participants in the current study. Lower panel: The mean AE vowel spaces as characterized separately for male and female native speakers by plotting for 12 monophthongs the mean formant frequencies observed for 48 female AE speakers (circular plotting symbols), and connecting these via line segments to the mean formant frequencies observed for 45 male AE speakers (diamond plotting symbols). The plotted mean formant frequencies were based upon analysis of the 12 h-V-d utterances reported in 1995 by Hillenbrand, et al. [3].

The graphic presented in the upper panel of Figure 4 gives examples of the vowel spaces exhibited by participants in the current case studies, which are complemented by mean observed AE formant frequencies in the lower panel in order to clarify for the reader what can be expected in general for male versus female speakers. These differences also suggest who great the shift in formant frequencies might need to be when such utterances are first spoken by a teacher with larger VTL, and are then synthesized for a Japanese ESL student with a smaller VTL. A detailed analysis of the related AE production results for all 9 Japanese ESL students is beyond the scope of this paper. While Figure 4 depicts the average formant frequencies for female speakers of American English in producing vowel sounds such as those that are observed for the 12 CVC-words inscribed in the circular symbols, the figure is also suggestive of the potential mismatch between the vowel spaces of a language teacher and learner. For example, were the teacher a male AE speaker, he likely would produce target utterances that would not be appropriate to the vowel space of a female speaker with a shorter vocal tract (exhibiting characteristically higher formant frequencies). Using the system described in this paper, Japanese ESL students were given the opportunity to attempt to produce the listed 12 CVC-words while being guided by target utterances that were synthesized with formant frequencies positioned in a manner appropriate to their own individual vowel space. Based upon initial experiences with this system, this application is under further development to allow for the assessment of L2 English-language learning experiences in non-native production and perception of AE vowels by native speakers of Mandarin and Korean language, similar to those reported in [16][17]. Comprehensive analysis of learners' produced vowel space characteristics and identification performance has begun, following the example set by the experimental studies reported in [1].

Another result of the iterative RTD process was the selection of particular solution for synthesis of pronunciation targets that was preferred for the application described in this paper. It should also be noted that the proposed adaptive approach to synthesis of non-native vowel sounds was quite successful for all nine participants in this initial investigation, with appropriate formant frequencies set for individual ESL learners applied according to their own vowel spaces. But it was not just the formant frequencies that were individualized, as the individual ESL learners heard targets that clearly sounded as though the ESL learners themselves had uttered them. This was accomplished via LPC-based analysis and synthesis using the learner's own source signals for synthesis of the non-native utterances, as illustrated in Figure 5, which gives an overview of this process which will be readily understood by those readers skilled in related signal processing techniques (as taught in [15].)

The technical process employed here is that which is generically termed cross-synthesis, since the learner's source (excitation) signal resulting from LPC analysis of the learner's utterance of one vowel sound is used as the input to an LPC-based source-filter synthesis of a different utterance. As such a separating of the information in an utterance into source and filter enables identity resynthesis, cross-synthesis enables the creation of novel utterances that clearly resemble those that are produced by the learner. Naturally, these provided good examples of utterances that the learner is able to produce, as was expected given the initial assumptions made during the "Grounding" stage of the RTD process in which multiple perspectives on the envisioned CALL system were entertained. For the shift in vowel space between speakers of different VTL (e.g., from a native AE example to the vowel space of a Japanese ESL student with smaller VTL), the shift in formant frequencies can be realized quite simply.

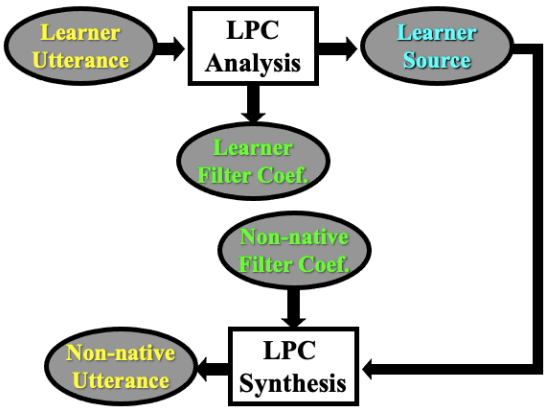

Figure 5. Flow chart illustrating the signal processing employed for LPC-based source-filter cross-synthesis of non-native utterances. LPC analysis produces two outputs for each of the learner's utterances, here designated as the "Learner Source" and the "Learner Filter Coef." LPC synthesis applies filter coefficients for a different vocal tract configuration to the Learner Source so that the learner's own glottal source spectral and temporal characteristics can be heard in the output "Non-native Utterance."

In the final stages of development, the merit of the developed CALL environment was assessed in terms of the benefits of the employed CAPT system. There may be other training mechanisms that if used in parallel will naturally complement the proposed system relying on adaptive recognition of the learner's vowel space. native-Japanese L2 learners of English language pronunciation. With regard to the problem of how best to provide pronunciation feedback, the proposed CAPT system solves a fundamental dilemma. It is understood that pronunciation feedback given to L2 learners should always be based upon an explicit mispronunciation model that also should be assessed for reliability and validity. When the model is implemented through an automatic process that utilizes audio signal processing on speech samples of the L2 learners, such as that of Hirabayashi & Nakagawa (2010), reliability can be established in the most straightforward manner. On the other hand, the validity of such automatic feedback processing is not always well established. For example, the validity of pronunciation scores generated using the method proposed in 2010 by Hirabayashi & Nakagawa [18] was assessed by comparing the pronunciation scores of L2 leaners with the pronunciation scores of a native speaker of the language to be learned. Such an approach need not respect the differences in vowel space between L2 learner and native speaker, which naturally would be considered by an impartial human judge of pronunciation quality.

The mispronunciation model of Ronen, et. al. [19] attempted to identify the expected set of mispronunciations for a given pairing of native and second languages. They recognized the subjective nature of the mispronunciation problem, and so they resorted to the validation of their results by correlating their automatically machine-generated scores with the scores produced by human judges (i.e., checking for the match between machine judgments and human judgments). In recently launched studies, the currently proposed CAPT system based upon adaptive recognition of the learner's vowel space is being tested in the manner exemplified in [1]. While it has been established that vowel space characteristics influence vowel identification accuracy (see, for example, [20]), it remains to be seen whether production of AE vowels by native speakers of Japanese can be improved by identification training based upon non-native vowel sounds presented in the context of learner's own vowel space (as enabled with the current CAPT system).

5 Conclusions

An online CALL environment was developed that provided Japanese learners with customized target utterances of 12 single-syllable words synthesized according to an adaptive recognition of the learner's vowel space. An automatic generation of those target utterances was accomplished through characterization of each learner's vowel space so that the formant frequencies used in synthesis of the utterances could be selected to conform appropriately to the formant frequencies that were calculated for the five vowel sounds produced by each learner in pronouncing a set of Japanese language terms. It was found that normalizing the vowel space used in synthesizing AE words in CVC form provided Japanese learners with a more comfortable and ultimately more effective learning experience for their training in pronunciation of American English.

References

1. S. Lambacher, W. L., Martens, K., Kakehi, C., Marasinghe, G. Molholt, The effects of identification training on the identification and production of American English vowels by native speakers of Japanese, Applied Psycholinguistics, **26**(2), 227-247 (2005)
2. R. Akahane, R. Yamada, T. Adachi, & H. Kawahara, Second language production training using spectrographic representations as feedback. Journal of the Acoustical Society of Japan (E), **18**(6), 341-343 (1997)
3. J. Hillenbrand, L. A., Getty, M. J., Clark, K. Wheeler, Acoustic characteristics of American English vowels. The Journal of the Acoustical Society of America, **97**(5), 3099-3111 (1995)
4. D. T. Ives, D. R. Smith, R. D. Patterson, Discrimination of speaker size from syllable phrases. The Journal of the Acoustical Society of America, **118**(6), 3816-3822 (2005)
5. A. C. Lammert, S. S. Narayanan, On short-time estimation of vocal tract length from formant frequencies. PloS one, **10**(7), e0132193 (2015)

6. A. Saito, Y. Heo, J. Perkins, Building Confidence in L2 Speaking through the Expanding Circle Communication: Practicing English as an International Language (EIL). Journal of Language Sciences, **27**(2), 199-228 (2020)
7. Y. Heo, A. Saito, Developing EFL Learners' Communicative Competence through Technology-mediated TBLT (2021)
8. D. E., Broadbent, P. Ladefoged, Vowel judgements and adaptation level. Proceedings of the Royal Society of London. Series B. Biological Sciences, **151**(944), 384-399 (1960)
9. J. E. Flege, The production of "new" and "similar" phones in a foreign language: Evidence for the effect of equivalence classification. Journal of Phonetics, **15**(1), 47-65 (1987)
10. K. Nishi, D. Kewley-Port, Training Japanese listeners to perceive American English vowels: Influence of training sets. Journal of Speech, Language, and Hearing Research, **50**(6), 1496–1509 (2007)
11. K. Nishi, D. Kewley-Port, Nonnative speech perception training using vowel subsets: Effects of vowels in sets and order of training. Journal of Speech, Language, and Hearing Research, **51**(6), 1480–1493 (2008)
12. K. Igarashi, I. Wilson, Improving Japanese English pronunciation with speech recognition and feed-back system. In SHS Web of Conferences, **77**, 02003 (2020)
13. J. Zimmerman, E. Stolterman, J. Forlizzi, An analysis and critique of Research through Design: Towards a formalization of a research approach. In proceedings of the 8th ACM conference on designing interactive systems, pp. 310-319 (2010)
14. K. Nakatsuka, A. Nogita, I. Wilson, Web application to convert English into helpful characters for pronunciation learners. In SHS Web of Conferences, **77**, 02005 (2020)
15. F. Keiler, D. Arfib, U. Zölzer, Efficient linear prediction for digital audio effects. In: Proceedings of the COST G-6 Conference on Digital Audio Effects (DAFX-00), Verona, Italy (2000)
16. J. E. Flege, O. S., Bohn, S. Jang, Effects of experience on non-native speakers' production and perception of English vowels. Journal of Phonetics, **25**(4), 437-470 (1997)
17. X. Wang, The acquisition of English vowels by Mandarin ESL learners: A study of production and perception (Doctoral dissertation, Theses (Dept. of Linguistics)/Simon Fraser University) (1997)
18. K. Hirabayashi, S. Nakagawa, Automatic evaluation of English pronunciation by Japanese speakers using various acoustic features and pattern recognition techniques. In Eleventh Annual Conference of the International Speech Communication Association, (2010)
19. O. Ronen, L., Neumeyer, H. Franco, Automatic detection of mispronunciation for language instruction. In: Fifth European Conference on Speech Communication and Technology (1997)
20. A. T. Neel, Vowel space characteristics and vowel identification accuracy. Journal of Speech, Language, and Hearing Research, **51**, 574–585 (2008)