Error Detection for Teaching Communicative Competence

W. Lewis Johnson
Alelo Inc.
Los Angeles, CA USA
ljohnson@alelo.com

Abstract— **The primary goal of Alelo's language and culture products is to help learners develop communicative competence. This paper gives an overview of Alelo's instructional and technical approaches for developing communicative competence, and places pronunciation training within that broader context. Courses address a wide range of knowledge, skills, attitudes and other relevant factors pertaining to communicative competence, including pronunciation skills. Error detection and remediation play an important role; however they must be provided in a way that supports the broader goal of promoting communicative competence. Speech and language technology adapted for language learners provides a foundation for this work. Focused pronunciation activities, some of which require specialized speech models, support the learning process. Learner data is used to develop a profile of each learner's competencies and predict their future attrition and decay. This makes it possible to provide learners with individualized curricula focusing on their individual needs.**

Keywords-computer-aided language learning; speech technology; learner modeling; error detection and remediation

I. Introduction

The foremost objective of language learning is to communicate effectively in real-world settings and situations [2]. To achieve this, learners must acquire a variety of language-related knowledge, skills, abilities and other related factors (KSAOs). Pronouncing the sounds of the language is one of these skills, and one that is a common focus of computer-aided language learning. However it is not necessarily the most important skill. For example, the proficiency guidelines published by the American Council for the Teaching of Foreign Languages cite lack of accent as a criterion only at the highest levels of proficiency [1]. Below that level it is sufficient for learners to be understood by native speakers of the language. Various kinds of errors can impede understanding and lead to misinterpretation, including problems with grammatical structures and vocabulary, unusual phrasing, or failure to conform to the pragmatic norms of discourse in the language. Native speakers frequently misjudge misunderstandings of non-native speech as being due to poor pronunciation [7], which may exaggerate the perceived importance of pronunciation errors.

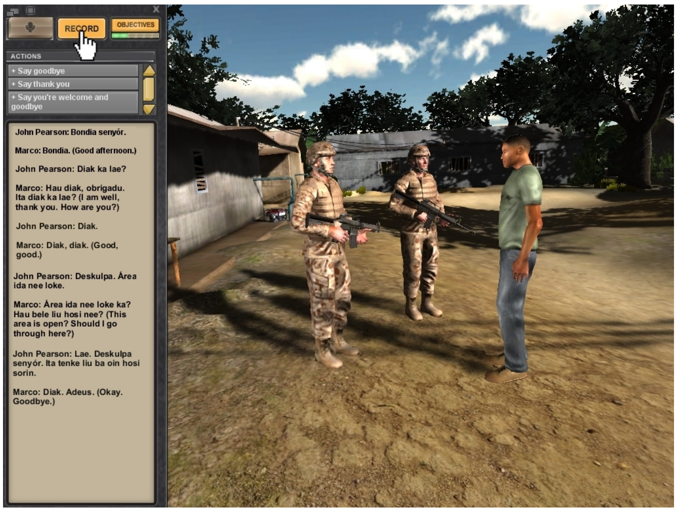
This paper gives an overview of the role of error detection and remediation within Alelo's instructional approach, which centers on the use of social simulations to help learners develop communicative skills. The approach is realized in a technology platform that is able to detect a range of language errors, both in the social simulations themselves and in other activities that help learners acquire the KSAOs that they will need to succeed in the social simulations. The paper discusses the role of learner language and learner errors in the underlying speech and language models. Then it focuses on methods for analyzing learner performance to provide feedback on pronunciation. Finally, the paper describes how learner performance data is used to assess communicative competencies in order to provide a learning experience customized to individual needs.

II. Instructional approach and examples

Alelo courses center on the use of social simulations for practicing and assessing communication skills. Social simulations engage learners in conversation with computer-generated characters. The computer characters interact with the learners in a manner appropriate to the social and cultural context of the conversation, and learners are encouraged to do the same.

This approach is incorporated into a variety of courses, in use around the world for language and culture education and training. Figure 1 shows an example of a dialog implementing the social-simulation approach, taken from a course in Tetum, the language spoken in East Timor. This course is intended for use by Australian military personnel preparing for overseas operations. In this dialog, the learner plays the role of a soldier named John Pearson (on the left) standing guard outside a restricted area. A computer-controlled character on the right, a Timorese man named Marco, wishes to enter the area. A conversation ensues. The learner engages with Marco by clicking on the Record button (top left) and speaking in Tetum. The system interprets the meaning of each learner utterance and the character responds accordingly.

In this example, the dialog began with basic greetings and rapport-building exchanges, as is customary in Timorese culture. The learner then attempted to inform Marco that the area is closed and off-limits. However the learner has made a mistake in using the word “loke” (open) instead of “taka” (closed), leading to confusion on Marco’s part. The learner is forced to correct himself and explain that the area is in fact closed.



John Pearson: Bondia senyór.

Marco: Bondia. (Good afternoon.)

John Pearson: Diak ka lae?

Marco: Hau diak, obrigadu. Ita diak ka lae? (I am well, thank you. How are you?)

John Pearson: Diak.

Marco: Diak, diak. (Good, good.)

John Pearson: Deskulpa. Área ida nee loke.

Marco: Área ida nee loke ka? Hau bele liu hosi nee? (This area is open? Should I go through here?)

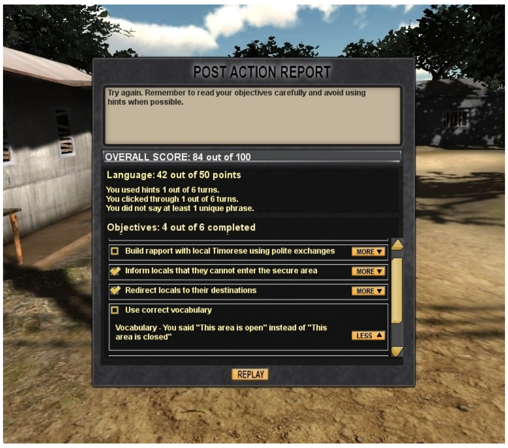
John Pearson: Lae. Deskulpa senyór. Ita tenke liu ba oin hosi sorin.

Marco: Diak. Adeus. (Okay. Goodbye.)

Figure 1. Example social simulation

The approach is motivated by research in how people learn in context [3]. The context affects how communicative skills are learned and how they are recalled and applied in real-world settings. The social-simulation approach therefore builds on the theory and methods of task-based language instruction [5]. As with other simulation-based learning approaches, the social-simulation approach is an experiential education method [4]. It gives learners opportunities to learn by doing and then seeing the results of what they did. Here, the learner had to explain to a man that he cannot pass, and saw the result of his mistake (the man asking for confirmation that he is free to enter). We refer to this type of error feedback as *organic feedback*, meaning that it is intrinsic to the behavior of the characters in the simulation. Learners find this to be a highly salient and meaningful way of receiving feedback on their performance.

At the conclusion of the simulation, the system generates a summary of the learner’s performance (Figure 2). This helps learners see the areas where they need to improve, and helps instructors to track learner progress so that they provide additional feedback and guidance. As this example illustrates, learners receive feedback on multiple aspects of their language performance. First, they get feedback on how well they performed the task, i.e., whether they accomplished the objective. They get feedback on the quality of their language performance, including how much they relied on hints, and whether they produced a variety of utterances instead of repeating the same memorized phrases over and over again. They also receive feedback on specific language errors. In this case, the focus is on improper use of vocabulary, a common problem for learners who are at the early stages of vocabulary mastery [11].



POST ACTION REPORT

Try again. Remember to read your objectives carefully and avoid using hints when possible.

OVERALL SCORE: 84 out of 100

Language: 42 out of 50 points

You used hints 1 out of 6 turns.
You clicked through 1 out of 6 turns.
You did not say at least 1 unique phrase.

Objectives: 4 out of 6 completed

□ Build rapport with local Timorese using polite exchanges
☑ Inform locals that they cannot enter the secure area
☑ Redirect locals to their destinations
□ Use correct vocabulary

Vocabulary - You said "This area is open" instead of "This area is closed"

Figure 2. Example scenario feedback

Viewed in this context, pronunciation skill and pronunciation training play a supportive and distinctly secondary role in the simulation and in feedback. Learners do not get feedback on pronunciation from the characters in the simulation. It is unusual in real life for native speakers to critique learner pronunciation in the course of a conversation, and such feedback would tend to break the sense of immersion in the simulation and turn it into a pronunciation exercise. The structure of assessments such as Figure 2 is informed by the needs and preferences of language instructors. In the case of the Tetum course, Australian military instructors want to know firstly whether the learner is able to accomplish the task, and secondly their command of phrases, vocabulary, and grammatical forms in performing it. Pronunciation accuracy is relevant, but has lower priority than these other factors. This prioritization is consistent with common standards for world language instruction (e.g., [2]).

This is not meant to imply that pronunciation accuracy is insignificant. On the contrary, the social-simulation approach requires careful attention to the language spoken by learners, including common patterns of pronunciation errors, and techniques for preventing them and remediating them. Otherwise dialogs such as the one shown might easily break down because the computer is unable to understand the learner's speech with sufficient robustness and reliability, and therefore cannot engage effectively in conversation with the learner.

III. Multimodal Communication with Learners

The foundation of the Alelo technical approach is a computational architecture for multimodal communication, designed for use in learning applications. This architecture is employed in all activities that involve social simulation, including interactions with animated characters such as the one with Marco in Figure 1. The following is a brief overview of this model; further details may be found in [8].

Processing in the Alelo architecture is a continuous cycle of learner behavior interpretation, intent planning, and behavior generation. Behavior interpretation involves processing input from the learner and inferring the communicative intent, i.e., the meaning that the learner intended to convey. In the intent-planning phase, the system decides what action to take in response, typically a communicative action. Then, in the behavior-generation phase, the system determines how to perform the action. This architecture has much in common with other conversational agent architectures, such as the SAIBA architecture [15]. What makes it unique is that it is designed for use in teaching intercultural communication skills.

We take a broad, comprehensive view of intercultural communication, including both verbal and nonverbal skills. The architecture takes input from the learner through both a verbal and a nonverbal channel, and then interprets the combination in the context of the culture to arrive at a behavior interpretation. The medium used for each channel depends upon the capabilities of the computing platform, as well as the learning objectives of the particular activity.

Verbal input is commonly, although not exclusively, obtained through speech processing. We have designed the architecture so that it can still function if the sound input channel or the speech recognition module has been disabled on the learner's computer. In such cases, learners may input their choices from menus instead. We are also looking to support text input, to help learners develop written language skills, as well as improve their mastery of grammatical forms in the language.

The nonverbal channel is used to capture gestures and body movements that have a communicative role in the target culture. This includes hand-gesture greetings (e.g., the palm-on-heart gesture common in the Islamic world), bowing, shaking hands, etc. In current courses, learners select these from menus; however, input could also be performed through a motion-capture interface such as Microsoft's Kinect system.

The output of the behavior-interpretation phase is a communicative act that describes the intended communicative function of the learner's input, together with features of the input that are useful for analysis of learner performance, e.g., a transcription of the spoken utterance and its duration. Communicative functions play a central role in the system, since the primary learning objective is to develop communicative competence. The dialog system processes communicative acts in real time, allowing non-player characters to respond appropriately to each dialog move the learner makes in the conversation. It also records and logs them for analysis and learner modeling. The learner modeling system uses the evidence from the learner's behavior to assess the learner's mastery of each of the communicative competencies in the curriculum [9].

IV. Support for Learner Language

A key feature of Alelo's technical approach is that it is designed to process and understand learner language, i.e., language forms produced by learners [6]. All components of the architecture are designed with the characteristics of learner language in mind, particularly the language of novice-to-intermediate learners, who have been the most common users of Alelo products to date.

Learner language at this level tends to have relatively limited complexity, consisting of relatively short utterances. Learners at the novice level make frequent use of memorized phrases [1]. Learners tend to have a restricted vocabulary, which is specified in the course curriculum and therefore somewhat predictable. These factors, together with the task and dialog context, serve to constrain the complexity of the natural language the system needs to understand. Thus, for example, in the dialog context in Figure 1, Marco can expect the learner to engage in a relatively limited range of communicative functions, and express them in a limited number of ways.

At the same time, learner language can have a broad range of variability in terms of accent, pronunciation errors, and other errors in linguistic forms and usage. The behavior-interpretation system therefore must have sufficient tolerance for variability and sensitivity to error. Tolerance for variability needs to be sufficient to allow the system to successfully interpret the learner's speech in most cases and respond accordingly, particularly in a dialog context. Sensitivity to error is required in order to detect and classify learner errors, assess learner mastery of component linguistic skills, and provide constructive feedback.

The desired degree of tolerance and sensitivity depends upon the level of the course and the learning objectives of the particular learning activity. For beginners, the highest priority is building confidence and allowing them to experience success; therefore a high tolerance for pronunciation errors is important. As learners progress, the tolerance for errors should decrease, to encourage them to improve.

To achieve sufficient tolerance for variability in learner pronunciation, we train our speech recognition models using a mixture of native speech and learner speech. The incorporation of learner speech helps to ensure that the input system is relatively tolerant of variability in accent. The speech recognizer combines a language model built out of vocabulary and phrases from the course, and a "garbage model" that can match with low probability against any utterance. The garbage model ensures that each learner utterance is positively recognized with sufficient probability, thereby minimizing the occurrence of false recognitions.

The speech input system dynamically switches between language models as the learner progresses. As the learner advances to more complex material, the perplexity of the language model increases. This has the effect of progressively increasing the accuracy threshold for the learner's speech, since utterances need to be recognized with progressively higher probability to distinguish them from alternative phrases and from the garbage model.

Sensitivity to error is achieved by incorporating common learner errors into the language model. The choice of which errors to include depends on the objectives of the learning activity, the reliability with which errors can be detected, and what sort of feedback is appropriate in a given context. Since learning objectives cover a range of linguistic forms (vocabulary, phrases, and grammatical structures), functions (communicative functions and rhetorical structures), and practices (pragmatics and context-dependent determiners of usage), a variety of types of errors can occur, and these can potentially be captured in the language model. In practice we utilize such error models mainly in focused exercises involving specific communicative skills, and only to a limited extent in extended dialogs (as in Figure 1), since this would defeat the purpose of the latter. If we were to continually interrupt the dialog with feedback on grammar and pronunciation, for example, the activity would quickly cease to be an exercise in communication and become an exercise in grammar and pronunciation.

We have conducted evaluations of the performance of the spoken dialog system, and have reported the results elsewhere [13]. In [13] we evaluated the speech-understanding performance of our Subsaharan French course against human raters. The percentage of misunderstandings (where the system assigned an interpretation that was different from the expert raters' interpretation) was quite low, 3.5% of the conversational turns. When human raters judged utterances to be incomprehensible, the system also rejected them as "garbage" 95% of the time. However there were many utterances (33% of the total) that the raters found comprehensible, but the system rejected as garbage. In 63% of these cases, the learners had one or more pronunciation errors. After further analysis, we concluded that many of the learners were confused by French orthography, and so we corrected the problem by providing better spoken hints, not by changing the speech understanding algorithms.

At the same time, we want to ensure that learners and teachers have a positive subjective experience with the system. Do they feel that the speech input system has an appropriate degree of tolerance for error, so that the activity is neither too easy nor too difficult? In general the answer is yes, with the exception of tone errors in tonal languages such as Chinese. Since our speech models are built from segmental phonemes, our speech recognizer can't distinguish tones in continuous speech, and is therefore overly tolerant of such errors. This requires us to make special provision for teaching the pronunciation of tones in tonal languages, as will be described below.

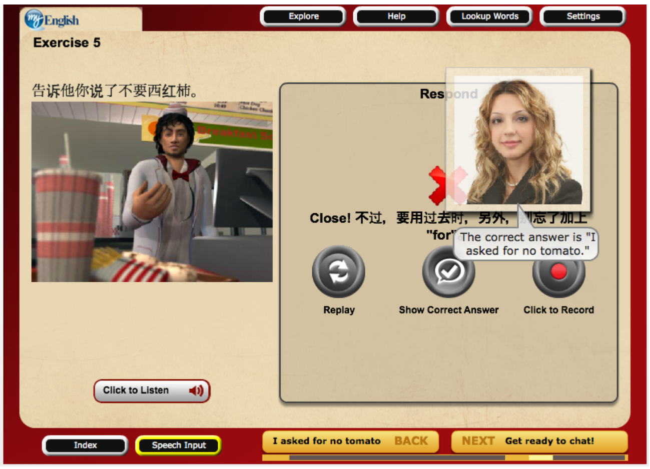
V. PART-TASK LEARNING

As noted above, successfully negotiating complex dialogs requires mastery of a combination of knowledge, skills, abilities, and other characteristics (KSAOs) pertaining to linguistic forms, functions, and practices. To help learners acquire these component KSAOs, Alelo courses provide a variety of structured part-task learning activities focused on a limited set of KSAOs. In these part-task exercises, learners receive much more detailed feedback on their performance of individual KSAOs, including pronunciation skills, than they do from extended dialogs.

Figure 3 shows a common type of part-task learning activity called a mini-dialog, which enables learners to practice individual communicative functions. This example is taken from the goEnglish course, an on-line course in colloquial American English developed for Voice of America. goEnglish is available in multiple languages and has over 100,000 registered users worldwide. Learning modules deal with a variety of situations in everyday life, work, and school in the United States.

This example is taken from a module on ordering food in a fast-food restaurant. There are a number of communicative practices involved in this activity that may be unfamiliar to people from other cultures. For example, when one orders a hamburger, one may choose from a range of toppings and condiments. If the cook gets the order wrong, the customer will need to negotiate with the restaurant staff to get it corrected. Part-task learning activities in the module introduce some of the individual communicative skills that can be helpful in such situations.

In this example, the learner has ordered a hamburger without tomato, and the hamburger arrives with a tomato slice on it. The learner's task is to inform the counter worker of the error. The learner decides what to say and clicks on the record button to say it. The software evaluates the choice and gives a response. There is no single right way to complete the exercise. Any well-formed utterance that conveys the intended meaning and is appropriate to the situation is rated as acceptable.



告诉他你说了不要西红柿。

Close! 不过，要用过去时，另外，别忘了加上 "for"

The correct answer is "I asked for no tomato."

Figure 3. A mini-dialog exercise

In this example, the learner's input was "I asked no tomato." This utterance illustrates a common learner mistake, i.e., omitting a function word, in this case "for." The system detects the error and provides an explanation and feedback, presented in part by the Virtual Coach (top right), who pops up and comments on the learner's response. Feedback typically includes a cognitive component (evaluation of the learner's response) and an affective one (encouragement and mitigation of embarrassment). This approach builds on research showing that pedagogical agents that interact with learners in a socially appropriate way can promote positive learner attitudes and yield improved student learning outcomes [10].

To detect and respond to such errors, the speech processing system employs grammar-based language models that match a variety of appropriate and inappropriate responses. Errors in grammar, morphology, semantics, and pragmatics are captured in this fashion. Pronunciation errors can be captured here as well, although in practice we tend to cover pronunciation in other activities that focus specifically on those skills. And since, as described above, a garbage model is active to capture utterances that do not match any of the expected responses, learners also get feedback if the system cannot precisely pinpoint the error.

VI. PHONETICS AND PRONUNCIATION EXERCISES

To further support the learning process, we provide part-task learning activities that focus on particular linguistic forms, including the sounds of the language. Some activities give learners opportunities to practice listening to and discriminating different sounds. Others give them practice speaking the sounds. These activities help to reinforce the phonetic skills that learners are acquiring in the conversational exercises.

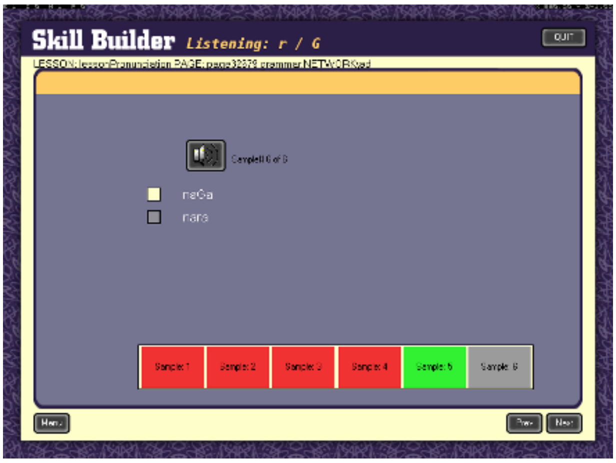


Figure 4. A sound discrimination exercise

Figure 4 shows an example listening exercise, taken from an Iraqi Arabic course. It gives learners the opportunity to practice distinguishing the apical rhotic /r/ from the voiced velar fricative /ɣ/, transliterated here as "G." English speakers often have difficulty distinguishing these sounds. Learners listen to a series of words containing one or the other of these sounds and indicate which sound they hear. This practice helps make them aware of the differences in sounds and better able to discriminate between them.

Pronunciation practice activities give learners practice speaking the sounds of language. These also focus on the sounds that learners tend to confuse and have difficulty discriminating. Figure 5 shows one such pronunciation practice exercise for Iraqi Arabic, again focusing on /r/ and /ɣ/. Learners are presented with minimal pair words that differ only in the target sound. They hear a native speaker pronounce each word, then they attempt to pronounce them. The system rates how close each learner utterance is to the two alternatives, and provides graphical feedback on a moving slider (top center). As the learner repeats the exercise, the displays at the bottom show their cumulative performance in producing these sounds. They continue until they are able to produce the sounds with sufficient reliability.

These exercises employ acoustic models that are constructed specifically for discriminating such sounds. We collect recordings of both native Arabic speakers and language learners speaking the target minimal pair words and build the models from the recordings. This capability is still experimental, while we collect additional training data to increase recognition reliability.

One of Alelo's development partners, VIFIN (Videnscenter for Integration), has developed an additional pronunciation practice activity using Alelo technology and has integrated it into a course they developed using Alelo's SocialSim^{TM} technology platform. This activity, called the Pronunciation Trainer, is shown in Figure 6. It is intended to help learners of Danish become familiar with its sounds. It presents the learner with a set of Danish words, each of which is an example of a particular phone in the Danish language. Learners repeatedly listen to and practice saying the words. A Danish speech recognizer developed collaboratively by VIFIN and Alelo attempts to recognize each word. A pedagogical agent named Harald (center right) provides feedback after each attempt. Each successful word recognition is treated as positive evidence that the learner has mastered the target phones, and each unsuccessful recognition is treated as negative evidence. The Pronunciation Trainer also serves as a reference tool. Learners can search not only words to practice pronunciation, but also single letters to find all words containing the letter with pronunciation variants.

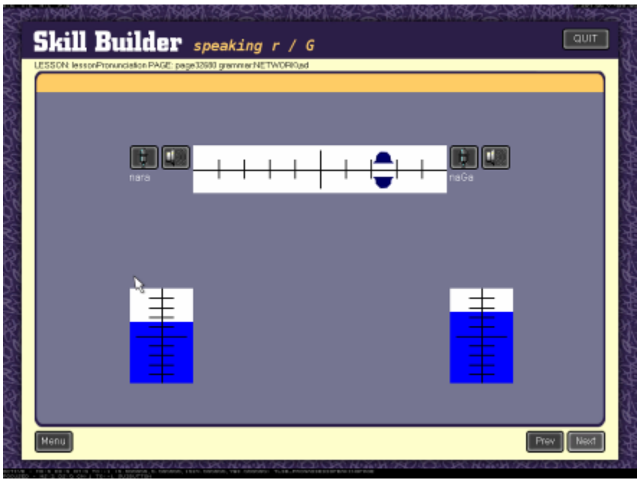
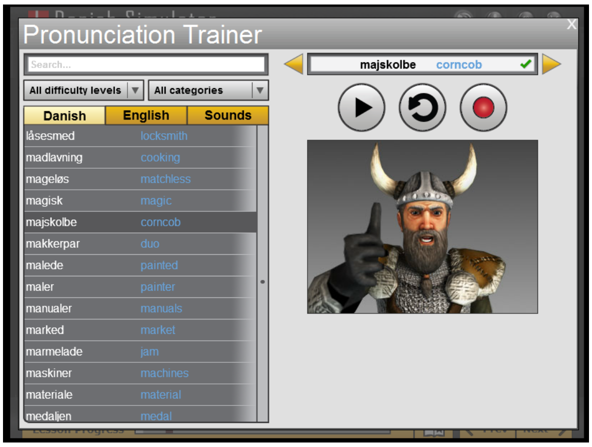

Skill Builder speaking r / G
QUIT
LESSON: lessonPronunciation PAGE: page32680 grammar:NETWORK.jsd
nara
naGa
Menu
Prev
Next

Figure 5. A phone practice exercise



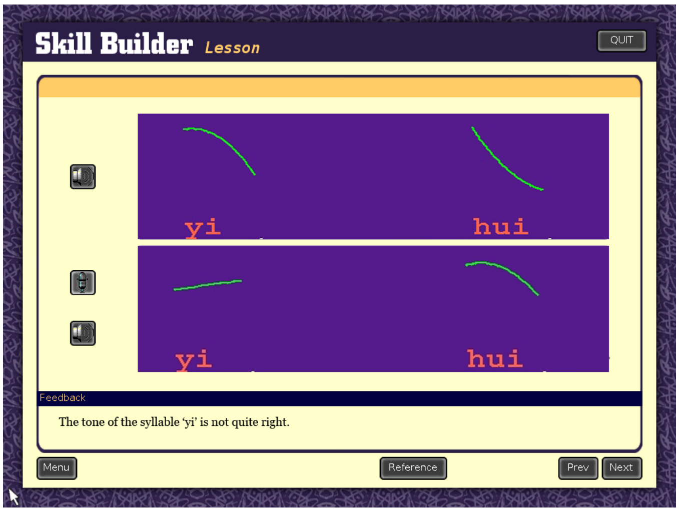
Danish	English	Sounds
låsesmed	locksmith	
madlavning	cooking	
mageløs	matchless	
magisk	magic	
majskolbe	corncob	
makkerpar	duo	
malede	painted	
maler	painter	
manualer	manuals	
marked	market	
marmelade	jam	
maskiner	machines	
materiale	material	
medaljen	medal	

Figure 6. VIFIN's Pronunciation Trainer

As mentioned above, tone production is a significant part of pronunciation in tonal languages such as Mandarin Chinese. Speech recognition algorithms typically process segmental phones and are insensitive to tones. So when we apply Alelo methods to tonal languages, we provide specialized pronunciation activities focusing on tone analysis and feedback.

Figure 7 shows the user interface for a prototype tone practice exercise called Tone Warrior. In this exercise, learners practice speaking two-syllable phrases, and are evaluated on their ability to produce accurate tones in these phrases. Two-syllable phrases expose learners to the complex interactions between the tones of adjoining syllables in languages such as Chinese, without introducing the added complexity of prosodic contours in continuous speech.

Tones are represented by pitch or fundamental frequency (f_0) [14], and analyzed using a super-resolution pitch detection (SRPD) algorithm [12]. The interface presents smoothed pitch contours that allow the learner to compare the shape of tones spoken by a native speaker with their own tones. The pitch detection algorithm can also distinguish qualitative tone shapes, such as the shapes of the four tones in Mandarin Chinese, and so can detect when the tone shape of a particular word is incorrect, as in this example.



The tone of the syllable 'yi' is not quite right.

Figure 7. Tone Warrior pronunciation activity

VII. Conclusions

This paper has provided an overview of automated error detection and feedback in the context of Alelo's language and culture courses, and has placed pronunciation error detection and feedback in that context. Alelo's approach emphasizes communicative competence, consistent with commonly recognized proficiency standards. Alelo's speech technology is designed to support robust human-computer conversational interactions, in the context of social simulations. Pronunciation error detection plays a role in this context, alongside detection of other types of language errors, insofar as it supports the broader goal of promoting communicative competence.

Proficiency standards indicate that as learners advance, the accuracy of their pronunciation should also improve. Alelo's spoken dialog technology is consistent with this model, since it requires learners to produce more accurate language in advanced dialogs than in beginning-level dialogs.

As Alelo continues to develop its learning methods and supports more advanced levels of language proficiency, pronunciation error detection can play a more significant role. We therefore see an expanding role for pronunciation practice activities that complement conversational practice activities. There is also a role for pronunciation assessment within conversational practice activities, as a component of an overall summary of the learner's competencies. However we will continue to view pronunciation as just one skill among the many linguistic KSAOs that learners must master, in support of the overarching goal of promoting communicative competence.

Acknowledgment

The author wishes to thank the members of the Alelo team who helped with the preparation of materials for this article, including Jidong Tao, Rebecca Row, and Mickey Rosenberg. Projects described here were developed under sponsorship from the Australian Army Simulation Wing, Voice of America, USMC PM TRASYS, and other agencies.

References

[1] American Council on the Teaching of Foreign Languages, "ACTFL proficiency guidelines: Speaking, writing, listening, and reading," Alexandria, VA USA: ACTFL, 2012.

[2] American Council on the Teaching of Foreign Languages, "Standards for foreign language learning: Preparing for the 21st century," Alexandria, VA USA: ACTFL, 2012.

[3] J.D. Bransford, A.L. Brown, and R.R. Cocking, "How people learn: Brain, mind, experience, and school," Washington, DC: The National Academies Press, 2000.

[4] J. Dewey, "Experience and education," New York: Collier Books, 1938.

[5] R. Ellis, "Task-based language learning and teaching," Oxford: Oxford University Press, 2003.

[6] R. Ellis and G. Barkhuizen, "Analyzing learner language," Oxford: Oxford University Press, 2005.

[7] S.M. Gass and L. Selinker, "Second Language Acquisition," New York: Routledge, 2008.

[8] W.L. Johnson, L. Friedland, A.M. Watson, and E.A. Surface, "The art and science of developing intercultural competence," in P.J. Durlach and A.M. Lesgold (Eds.), "Adaptive technologies for training and education," pp. 261-285, New York: Cambridge University Press, 2012.

[9] W.L. Johnson and A. Sagae, "Personalized refresher training based on a model of competency acquisition and decay," in Proceedings of the 2nd International Conference on Applied Digital Human Modeling, in press.

[10] W. L. Johnson, and N. Wang, "Politeness in interactive educational software," in C. Hayes & C. Miller (Eds.), Human-Computer Etiquette, London: Taylor & Francis, 2010.

[11] B. Laufer-Dvorkin, "Similar lexical forms in interlanguage," Tübingen: Narr, 1991.

[12] Y. Medan, E. Yair, and D. Chazan, "Super resolution pitch determination of speech singals," *IEEE Trans. ASSP*, vol 39, pp. 40-48, 1991.

[13] A. Sagae, W.L. Johnson, and S. Bodnar, "Validation of a dialog system for language learners," in Proceedings of the SIGDIAL 2010 Conference, pp. 241-244. Tokyo: Association for Computational Linguistics, 2010.

[14] Ye Tian, Jian-Lai Zhou, Min Chu, and Eric Chang, "Tone recognition with fractionized models and outlined features," in Proceedings of *ICASSP* 2004, Quebec, Canada, pp. 105-108, 2004.

[15] H. Vilhjalmsson and S. Marsella, "Social Performance Framework," in Proceedings of the AAAI Workshop on Modular Construction of Human-Like Intelligence, Menlo Park: AAAI, 2005.