Automatic Speech Recognition: A Shifted Role in Early Speech Intervention?

Foad Hamidi, Melanie Baljko

Department of Electrical Engineering and Computer Science,
Lassonde School of Engineering, York University, Toronto, Ontario, Canada
{fhamidi,mb}@cse.yokru.ca

Abstract

Although *automatic speech recognition* (ASR) has been used in several systems that support speech training for children, this particular design domain poses on-going challenges: an input domain of non-standard speech and a user population for which meaningful, consistent, and well designed automatically-derived feedback is imperative. In this design analysis, we focus on and analyze the differences between the tasks of speech *recognition* and speech *assessment*, and identify the latter as a central issue for work in the speech-training domain. Our analysis is based on empirical results from fieldwork with Speech-Language Pathologists concerning the design requirements analysis for tangible toys intended for speech intervention with primary-school aged children. This analysis leads us to advocate for the use of only rudimentary ASR feedback.

Index Terms: speech intervention, automatic speech recognition

1. Introduction

In the context of *control* systems, *automatic speech recognition* (ASR) refers to a series of techniques combining signal processing, statistical modeling, and machine learning to interpret human speech typically by deciphering input acoustic signals into phones or other linguistic elements such as syllables, words or phrases [1]. Speech, as a mode of input, has been taken up in many ASR-based applications in the disability community, such as for speech-to-text communication technologies and for command interpretation systems for hands-free computer use [2]. However, there are key differences between these speech-based *control* systems and those system for *speech training*.

Speech training for children, as conducted in face-to-face sessions led by a *speech language pathologist* (SLP), involves eliciting speech that includes the problematic segment that has been targeted for intervention. The child is provided with corrective feedback (best practices from clinicians adopt a feedback approach at word-level or even coarser granularity). The SLP draws upon a repertoire of techniques for speech elicitation and feedback.

The potential of ASR to support computer-based tools to improve the efficacy of the traditional face-to-face clinician-client dyad and the potential to provide new modes of intervention, outside of face-to-face sessions with an SLP has been recognized previously [3]. Despite the recognized benefits, relatively few computer intervention systems that incorporate ASR have been developed and thoroughly evaluated. A recognized obstacle for the use of ASR in speech intervention systems has been that this technology oftentimes does not perform well for non-standard pronunciations and can lead to inconsistent feedback [4]. Other systems focus on the use of multimedia instructions (i.e., animation and audio) to aid parents and SLPs communicate feedback to children in the course of speech exercises, but do not use ASR (e.g., [5]). In our design analysis, we discuss these systems with a view to clarify and reposition the design objective for this particular design domain.

Many language learning and practice applications have been developed in recent years for smartphones and tablets [6]. Many of these applications are digital versions of flashcards and pictures to help SLPs in intervention (e.g., Phonics Studio). A few of these applications record speech and provide data gathering (e.g., Articulate It!). The potential benefits of these applications for speech training and intervention are clear, and the field looks forward to systematic usability and efficacy evaluation.

Our design analysis focuses on the theoretically oriented question of what is the feasibility of automatic corrective speech feedback for children? Having clear answers to this and other foundational questions are prerequisites for good applications. We provide a literature review of previous computer speech intervention systems that incorporate ASR, with view of identifying challenges and techniques to address them. A goal is to contribute toward the design of new-generation speech intervention system and to yield novel insights. To this end, we have conducted fieldwork with five clinic-based SLPs who work with pediatric populations, with a particular focus on the designs of tangible toys intended for use as part of and in support of speech intervention protocols.

2. Analysis

2.1. Challenges in Repurposing

Prior speech intervention systems have either incorporated extant ASR engines or have developed specific ASR engines for their projects (such as [7] and [8]). Reuse, in general, is often a good strategy, since it has the advantage of repurposing a large amount of work and effort gone into the original design of the ASR. However, this reuse has introduced a number of issues. The first issue concerns the nature of the ASR output. Speech intervention systems require the *analysis* of input speech that is relative to a given target. The required output needs to provide useful information about the differences between the elicited and the targeted speech unit, which is necessary in order to provide corrective feedback. Traditional conceptions of ASR systems provision for the *identification* of words within speech, where the content of that input speech is not known *a priori*. The *recognition result* from the ASR module is provided in the form of a lexicographical interpretation for some particular input acoustic signal. Thus, one can recognize a misalignment between what ASR module provides and the design

SLPAT 2013, 4th Workshop on Speech and Language Processing for Assistive Technologies, pages 55–61,
Grenoble, France, 21–22 August, 2013. ©2013 Association for Computational Linguistics

requirements. A key challenge in this design domain is the alignment and extraction of information that will be useful for corrective feedback, whereas a main challenge for ASR (more generally) is identification in the face of deviation from the training pronunciation.

Extant, general ASR modules (e.g., *Dragon Naturally Speaking* [9]) are mostly developed for speakers with clear speech. These modules are derived from human speech samples and are trained on clear "standard" speech. When the input speech differs from the modeled speech, due to reasons such as when the input speech is produced by a speaker with an accent or speech impairment, the performance of the ASR module degrades [2, 10]. The performance further degrades when speech is affected by environmental noise, distortion and sound quality change [11].

An *error,* in this context, can be understood as either a *recognition error,* where input is "correct" but the system fails to recognize it, or a *speech error,* where input speech significantly deviates from a standard model. Despite rapid improvements in ASR technology, some researchers believe that because sound and specifically speech is a noisy input channel, errors are an inevitable part of ASR technologies [1]. In the presence of non-standard speech, ASR modules produce low confidence scores for predicted candidates, reflecting the high possibility of recognition errors. In response, several research initiatives have focused on ASR specifically for dysarthric speech (e.g., [12, 13]) and/or the speech of children (e.g., [14]).

2.2. Prototype Systems

Kewley-Port et al.'s early system was developed using recorded templates of the child's best production, which were then used as standards against which to measure the acceptability of new utterances [15]. The researchers conjectured that recognition error rates as high as 20%, a rate within the capabilities of a small vocabulary speaker-dependent system, would be acceptable for articulation training. A more detailed assessment of the degree of success of the system was not provided. Adoption of this approach has been limited, however: training is required for each individual, and target words and phrases that consist of segments not producible by the child are not possible (thereby obviating application for speech intervention).

Speech intervention mediated by the *Speech Training, Assessment, and Remediation* (STAR) system, a system designed to distinguish between the segments /r/ and /w/, was achieved through a role-playing game with the premise that "aliens" need to understand the child's speech [16]. Evaluation was conducted in which likelihood ratios, as calculated automatically by the ASR module, were compared with perceptual quality ratings, as provided by human judges. The results showed high correlation between the two measures for substitution errors. In other words, the system worked well when /r/ and /w/ were misarticulated. However, the ASR module produced many false positives (i.e., the results correlated poorly for correctly articulated examples).

2.3. Box of Tricks

Vicsi et al. developed a speech intervention system, *Box of Tricks,* for children with hearing impairment [8]. *Box of Tricks* uses ASR to detect and to provide feedback about speech mistakes and was originally devised to support Hungarian, and has subsequently been expanded to also support English, Swedish and Slovenian. *Box of Tricks* is designed to train for vowels and also fricatives.

The goal of *Box of Tricks* is to teach children to modify their speech on the basis of visualizations of their speech signals. Picture-like images of energy, change in time, fundamental frequency, voiced or unvoiced detection, intonation, spectrum, spectrogram (cochleogram) and spectrogram differences were used for the visualization.

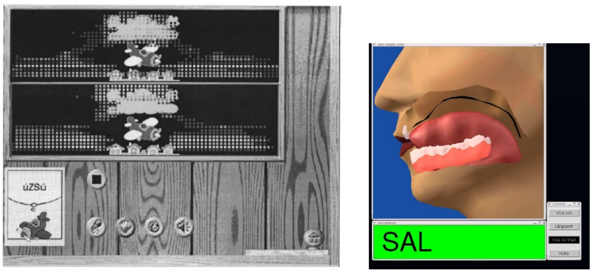

Figure 1: *Feedback from two systems designed for children: Box of Tricks (left) [8] and ARTUR (right) [7]*

For the visualizations, a filter was developed and applied that produces a representation based on inner ear processing rather than FFT spectra. The researchers hypothesized that the visualization generated by this filter would be a more intuitive representation of speech for their users than other types of visualizations. The representation of elicited pronunciation was shown in alignment with a representation of a target pronunciation. Parts of the representation were highlighted to signal more important features of the speech: to draw the children's attention to these parts that were highlighted by amusing background pictures. These visualizations, especially combined with gaming elements, would be more stimulating than numerical scores.

Box of Tricks did not provide overt instruction to the children about how to correct their speech, however. Although the users were provided with feedback that indicated, in some fashion, the differences between their input speech and the desired, target speech, they were not provided with clear instruction for how this difference might be decreased. The researchers conjectured that this approach provides meaningful feedback to children and allows them to use the system by themselves. It was not clearly demonstrated that, in the absence of such corrective feedback, the children were able to incorporate the information into their motor learning, but neither was the conjecture disproved.

2.4. ARTUR

Bälter et al. developed a prototype of a computer system for speech intervention for children with hearing impairments to be used in the absence of SLPs [7]. The system aims to identify problematic pronunciations and provide corrective feedback. A computer-animated head with exposed internal parts of the face and mouth, referred to as the *Articulation TUtoR* (ARTUR), was constructed. ARTUR was utilized to provide feedback based on the input (albeit not synchronously with the elicitation). The researchers hypothesized that, for children with hearing problems, the visualization of the movement of vocal tract is more useful than acoustic signal visualization. A knowledge base of mappings was constructed: for each possible error, an appropriate corrective response was developed (some corrective responses were reused). The feedback, in the form of spoken commands and corresponding animation, was drawn from this knowledge base. The researchers conjectured that showing the hidden parts of vocal tract would be key to effective speech intervention. In the final implementation of the system, audio input is to be supplemented with video footage of the user for more accurate categorization of pronunciation error.

The system was tested with two groups of children in a Wizard-of-Oz study. The children in the first group were six years old and the ones in the second group were between nine and eleven years old. In addition to children, an adult with English as second language also used the program and provided feedback.

The empirical qualitative data demonstrated that the children, especially the older group, liked the idea of playing with a computer and being given explicit feedback. However, while they (and especially the older group) liked the program in general, they found the visual feedback confusing and unhelpful. This was found of both the image representation of speech organs and the accompanying animation. The children suggested that adding more game-like features, such as goals and rewards, to make it more engaging. Also, they found the user interface of the program, as well as the anatomy of the vocal tract (e.g. the hard palate), unclear. When compared to interaction with the SLP, older children described the interaction as more relaxed. In more recent work, ARTUR's interface was assessed for use in second language pronunciation training for adults and children [17].

In a study of the pedagogy of feedback conducted to inform an application of the system for second language training, Engwall and Bälter demonstrated that, even given the availability of accurate information for feedback, many interaction decisions such as when and how to deliver feedback need to be built into the design of a given application [17, 18]. The study was done in the context of second language learning but the results are still relevant to pronunciation training.

2.5. Speech Viewer II

A commercial (but no longer in distribution) speech therapy system, *Speech Viewer II,* was developed to help adults with speech impairments improve their speech [19]. This system visualized speech signals and waveforms. Figure 2 shows speech visualization produced by this system.

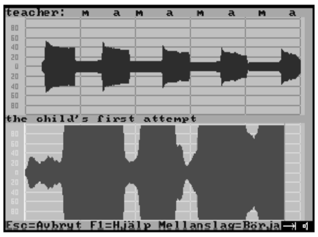
teacher: M a M a M a M a
80
60
40
20
0
20
40
60
80
the child's first attempt
Esc=Avbryt F1=Hjälp Mellanslag=Börja

Figure 2: *Speech Viewer II uses wave diagrams as feedback [18].*

Two studies have shown that this system does not work well for use by children with hearing impairments. The first study showed that the program did not have any advantages over traditional speech therapy for vowel training for children with profound hearing impairments [20]. The second study tested a vowel accuracy feedback with children with hearing impairments and showed that the system produced modest gains but exhibited inaccuracies and inconsistencies in feedback [21].

When the use of *Speech Viewer* was restricted to the improvement of prosodic features of speech for children with hearing impairments, better results were produced. Öster conducted a study with two deaf children who were trained using the program for ten minutes twice weekly over an eight week period [22]. For each child, a different skill was targeted: a fifteen-year-old boy who had difficulty with producing durational contrast between phonologically long and short vowels, and a thirteen-year-old girl who had difficulties producing voicing contrasts between voiced and voiceless velar stops. Both children were reported to have improvements in the areas targeted. Öster also conducted a study with a five-year-old deaf boy who had difficulty controlling the loudness and pitch of his speech [19]. While detailed information about the amount of training, methodology and the results of the intervention is not provided, the researcher reported that use of the program, and specifically its graphical interface, allowed the SLP to communicate better with the child, resulting in improved loudness and pitch.

2.6. OPTACIA

Öster et al. have conducted initial experiments with the OPTACIA system, which is similar to Speech Viewer, and produces visual maps for training Swedish sibilant fricatives, fricatives with higher-frequency and acoustic energy than non-sibilant fricatives, to hearing-impaired users [23]. The system is designed to supplement speech intervention. The user is provided with a visual representation of his or her speech that is shown in relation to a visualization of a target pronunciation. The researchers hypothesize that having this feedback will help increase the frequency of correct pronunciations. In this system, the produced diagrams will be described by the SLP and used as a tool during therapy to visualize specific components of speech.

The speech of three severely hearing-impaired children when pronouncing the fricatives was recorded and mapped against the created maps and it was found that the visualizations corresponded well with the speech produced.

While this system shows it is possible to create visualizations that correspond with non-standard input speech, it did not discuss the usefulness of this approach for children. The input data was restricted to sibilant-vowel combinations rather than words, and the visualizations were shown in terms of time and frequency, an unintuitive approach for children. The project was in its initial phase and no user studies were conducted.

2.7. visiBabble and VocSyl

The visiBabble system, manifested either as a tangible toy or as a software application, processes infant vocalizations in real-time and produces brightly colored animations, intended to provide positive reinforcement of the production of syllabic utterances, intended as an early speech intervention and support for later language and cognitive development [24, 25].

In a similar vein, the VocSyl system also used speech and vocalization analysis and visualization to engage children's speech [26, 27, 28] using a software application. VocSyl uses a suite of audio visualizations to represent different audio features of speech (pitch, loudness, duration and syllables) in abstract visual representations that are presented to children in real-time.

visiBabble and VocSyl are intended to encourage children with speech delays. VocSyl was originally designed for motivating children with *Autism Spectrum Disorder* (ASD) to encourage speech vocalizations [28]. An initial study of VocSyl with 5 children with ASD showed that audio and visual stimulation increase the rate and duration of speech like vocalizations. Hailpern et al. found that each of the children responded to at least one form of feedback and that only some participants responded to visual stimuli whereas others responded to auditory stimuli or a combination of visual and auditory stimuli. They also found that it is likely that visualizations should be customized to some extent for each person [28].

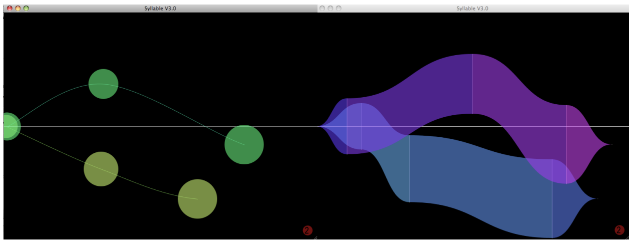

Figure 3: *VocSyl visualizations to illustrate multisyllable words [27]*

A more recent application of the VocSyl system supports the production of multi-syllabic speech production in children with autism speech apraxia and speech delays. One of the goals is to provide children with a persistent visual representation of their speech that would facilitate reflection and a new experience of language skills. The goal is to use visualizations to illustrate differences in utterances and help with the ability to combine syllables both as word combinations and in single multisyllabic words [26, 27].

Figure 3 shows the interface of *VocSyl.* Syllables are represented by discreet elements (left screenshot) or regions in continuous visualizations (rights screenshot) and emphasis, pitch change and pacing are represented by the diameter of the graphical element and position on the y-axis and x-axis, respectively. The researchers involved two children with ASD, two children with SPD and four children without disabilities in the design of the system.

While the system does not currently provide corrective feedback, it focuses on engagement and motivation and, also, provides the visualizations as a communication aid to help SLPs demonstrate specific aspects (i.e., syllable location and volume) of the vocalizations. It is apparent that if corrective feedback were given in the absence of SLPs or parents to facilitate their interpretation, the children would not have been as motivated to continue using their speech.

2.8. Field Interviews

Like Fell et al. [24, 25], we are also interested in the development of tangible interactive toys for the support of speech intervention [29]. To this end, we conducted open-ended interviews with five SLPs who work with children (our target user population is ages 4-7). We reached these SLPs by direct contact.

All the interviewed SLPs felt that a toy that focuses on speech elicitation would be useful. Three of the SLPs already use props such as dolls and physical toys, as well as, images and flash cards to engage children. These toys allow for the development of narrative and the engagement of the children's attention. They stressed that it is useful to have toys that when working with small children (ages 4-7) can be touched and grasped and are also durable.

Two of the SLPs who were interviewed used iPads to play games that involve speech. Surprisingly, they preferred games that encourage speech through stories and play but are not specifically developed for speech intervention and have simple interfaces, (e.g., My PlayHome). One SLP commented that she prefers to use non-computational material during intervention because too much technology can be distracting for the children.

It was noted that, sometimes, initial engagement of children is difficult and it takes a long time to establish a relationship with them to the point where they start using their speech more freely. It was also noted that capturing the child's natural speech (i.e., speech spoken in the absence of the SLP) would be helpful in assessing intervention needs. One SLP records samples of her client's speech during some of her sessions. She uses these samples for future comparison of intervention outcomes, analysis of speech in the absence of the client.

All SLPs indicated that having no or little feedback that is consistent and accurate is better than inconsistent or incorrect feedback, especially in the absence of the SLP who can mediate between the technology and the child. However, they mentioned that some measure of progress is necessary so that not all speech is rewarded equally. Additionally, the SLPs suggested that automatic tracking and record keeping of exercises are useful functions that a computational toy could provide.

Three of the interviewed SLPs discussed the context of multilingual communities. Working with children who are multilingual is quite common in Toronto, and in Canada more generally, due to the presence of many new immigrants. These SLPs noted that many immigrant children whose first language is not English face difficulties when moving to a new country where English is the main language and noted this condition as a contributor to speech delays. The issue is complex, as the home language is often not English, the parents and caregivers are not fluent and are not in a position to assist with speech exercises at home. Additionally, as the children grow up, they are faced with the challenge of switching between English and their home language. These challenges can place stress on interfamily relations and cause disconnect between children and their parents. These SLPs highlighted the particular need for a toy that is able to switch between languages and that supports communication between children and parents in languages other than English. School board policies oftentimes specifically encourage parents to speak and read with their children in the home language, as a support for language development.

3. Discussion

3.1. The challenges of feedback

As indicated previously and reinforced by our SLP interviews, the provision of meaningful and consistent feedback to children is a key priority. The feedback must not only be evaluative (i.e., indicating whether a given speech target was reached or not), but must also provide corrective analytic feedback (i.e., analyze and semantically interpret degrees of deviation between elicited speech and a given speech target). Analyses even from a decade ago identified that ASR technology is not able to provide corrective analytic feedback and can only provide evaluative feedback [30].

For an ASR module operating in a speech intervention context, if it is unable to produce a lexicographic candidate for a given speech input, then the system must decide whether this error is an instance of speech error due to poor speech or a recognition error due to poor system performance. In this situation, the system may exploit additional information in the form of information about the expected input.

Researchers have determined that systems, such as those reviewed in prior sections, which rely on abstract visualizations as feedback do not seem to work well for children. Neri et al. have identified a major problem with providing comparable waveforms, a popular form of feedback (e.g., *Speech Viewer II*), to the user [31]. Although showing target and input waveforms in alignment can be motivating for the user (i.e., to try to emulate the target waveform by modifying their pronunciation), it does not necessarily lead to behavior modification (i.e., correction of articulation). Moreover, Neri et al. argue that such alignments may be misleading, since it is possible for two articulations to both be "on target" and yet have waveforms that are very different from each other; they argue that even a trained phonetician cannot extract information needed to correct pronunciation from this feedback, let alone a user who does not have any training in interpreting this form of feedback [31].

3.2. Shifting from analysis to elicitation motivation

Although ASR is challenged by certain requirements of this design domain (namely the need for corrective analytic feedback), it supports admirably well another requirement: that the system be engaging, interactive and motivates repeated speech productions by the child. A key observation here is that incorporating ASR can make the computer system responsive to speech even if it does not provide detailed feedback. In the context of speech intervention, even rudimentary feedback can be of value, since it can motivate children to try multiple repetitions of words and phrases. This approach has been used in a remarkable study by Mitra et al., in the context of accent reduction, which found that even rudimentary feedback was helpful [32].

In this study, sixteen children between the ages of twelve and sixteen were chosen from an Indian English median school, where English was the primary medium of teaching but was spoken with a strong accent. They were grouped into four groups and given access to a computer for three hours a week. The children were provided with "Ellis", an English language learning program with no ASR support, four classic English-language films that they could choose to watch during their time at the computer and the previously mentioned *Dragon Naturally Speaking* program. The children were given the objective of making their speech understood by the *Dragon Naturally Speaking* program that either accepted or rejected input speech and did not provide corrective feedback. No further instructions were provided following an initial demonstration of the resources. Rather surprisingly, the approach was effective. To measure improvements in speech and whether they carry to real-life situations, four human judges were provided with video clips of children speaking at different evaluation points. A measure of the percentage of words correctly recognized was calculated. Significant improvements over a five-month period were observed. Furthermore, the word recognition rates by the ASR module were correlated with the human judges' assessments of pronunciation accuracy (e.g., an improvement of 117% was observed, as assessed by the human judges and of 79%, as assessed to the ASR module).

In another study of second language training, class observations and teacher interviews, revealed that in practice very little feedback is given to the students [17]. Reasons for limiting feedback were to maintain a positive atmosphere and communicative flow. A study of literature on the pedagogy of feedback shows that according to many theories (e.g., [33]), the *encouragement* of speech and communication is as important as its correction.

Scientific researchers in this domain may quickly conclude that the need for high-quality corrective analytic feedback clearly motivates the need for further work into automated speech analysis. And such work is ongoing. For instance, efforts in the area of *acoustic training*, which entails to the process of recording representative speech samples from a user to create or to augment an acoustic database [2]. In particular, Rudzicz has recently developed and validated a highly specialized ASR module for dysarthric speech [12]. Another project, the *Universal Access* (UA) dysarthria speech database has gathered a collection of speech samples from individuals with speech dysarthria that can be used to incorporate knowledge about dysarthric speech into an ASR application [13]. The other approach from the Assistive Technology domain to increase effectiveness of ASR for users with non-standard speech, *input restriction,* may also been seen as providing a useful avenue for speech intervention, since the approach relies on simplifying the recognition task by restricting the input to a limited number of isolated words, rather than continuous speech. This approach has been used widely, and improves accuracy rates (e.g., Rosengren et al. showed that adapting the vocabulary for each user improved accuracy rates from 28% to 62% [34]), and has been employed in some of the previously described systems (e.g., [19, 23]).

But in a parallel stream to the specific ASR research and development work underway, one may consider the broader design parameters of the application domain: a designer may see this situation not so much as an obstacle, but rather an occasion or opportunity to contemplate more generally the role of ASR in speech intervention systems, systems which are needed for the here and now, for deployment on a time-scale that is not hinged to the outcomes of medium- and long-term automated speech analysis research projects, and for contexts in which an SLP is already present and mediating the speech intervention session (who is trained and experienced with the design of corrective feedback).

Although it would seem unintuitive, we conjecture that rudimentary feedback provides more value than other more detailed types of corrective feedback. Rudimentary feedback preserves a main point of "value" of ASR, which is as the main driver for motivating, interactive technology-mediated experiences. These encounters motivate the elicitation of multiple and repeated speech productions over a sustained period of days or weeks. Engwall et al. [17, 18] correctly identified the need for nuanced and carefully designed strategies to deliver corrective feedback. We argue that the same care and attention is needed for the motivation strategy for eliciting productions. And though these two aspects are clearly intertwined, we are currently pursuing "low-tech" strategies in which there is a radical rethinking of the role of ASR in a speech intervention system. We recognize the limitation of ASR to analyze non-standard speech and instead use it to facilitate and motivate the use of speech as an input mode. The task of providing detailed feedback can be left to the SLP (the "human agent") and the use of ASR, and the computational media more generally, can be recruited for user engagement, motivation and the elicitation of speech productions.

A point that needs mention is that the findings discussed here is based on an assumption about the lexical unit being short and the language having a relative low ratio of morphemes to words (e.g., as in English); we expect the results to generalize to other moderately analytic languages, but may not generalize more broadly, for example to synthetic languages (e.g., Greek).

4. Conclusion

In this paper, we have reviewed a number of systems that employ ASR for speech or pre-speech intervention. We discuss how ASR technology as of present often provides unreliable and approximate feedback in the presence of ambiguous or erroneous speech, which results in unintuitive and inconsistent feedback that can be confusing and ineffective to users.

Extant intervention systems that use ASR face the main challenge of designing effective feedback. There remains a misalignment between the original design goal of ASR modules (i.e., recognition of speech) and their repurposed role in computer speech intervention systems (i.e., analysis and assessment of speech). Research demonstrates that abstract representations such as waveforms and closeness scores are unintuitive for children and have not been helpful in correcting speech. Our fieldwork shows that SLPs themselves highly value ASR and computational media more generally for its effect in motivating users and eliciting repeated speech productions.

While *input restriction,* a method used previously in systems developed for users with dysarthric speech and strong accents can be employed to improve the performance of ASR modules, based on reported interview results with SLPs and the literature review, a more radical shift in the role of the ASR module is suggested. This method involves using ASR to engage rather than evaluate speech, given the goal of facilitating sustained practice through the elicitation of multiple repetitions of target words and phrases. As demonstrated by Mitra et al., it can be effective to subordinate the accuracy of ASR to its use as a facilitator and "encourager" of interaction [32].

5. Acknowledgements

We would like to thank the SLPs who took time and provided us with useful information through the interviews. This research is supported by the Natural Sciences and Engineering Research Council of Canada Discovery Grant, "Embodied Interaction Design for Articulation Elicitation and Feedback".

6. References

[1] Danis, C. and Karat, J., "Technology-driven design of speech information systems", Proc. of DIS'95, ACM, 17-24, 1995.

[2] Hawley, M. S. "Speech recognition as an input to electronic assistive technology", British Journal of Occupational Therapy, 65(1): 15-20, 2002.

[3] Zhao, Y., "Speech technology and its potential for special education", Journal of Special Education Technology, 22(3): 35-41, 2007.

[4] Hsiao, M. L., Li, P. T., Lin, P. Y., Tang, S. T., Lee, T. C., and Young, S. T. "A computer based software for hearing impaired children's speech training and learning between teacher and parents in Taiwan", In Engineering in Medicine and Biology Society, Proc. of the 23rd Annual International Conference of the IEEE (Vol. 2, pp. 1457-1459). IEEE. 2001.

[5] Menzel, W., Herron, D., Bonaventura, P., and Morton, R. "Automatic detection and correction of nonnative English pronunciation", In Proceedings of Workshop Intergrating Speech Technology in the (Language) Learning and Assistive Interface, InStil , 49–56, 2000.

[6] Teachers with Apps. "31 Speech And Language Apps For iPad", Retrieved from: http://www.teachthought.com/literacy-2/31-speech-and-language-apps-for-ipad/. 2013.

[7] Bälter, O., Engwall, O., Öster, A., and Kjellström, H., "Wizard-of-Oz test of ARTUR: A computer-based speech training system with articulation correction", In Proceedings of ASSETS'05, ACM, 36-43, 2005.

[8] Vicsi, K., Roach, P., Öster, A. M., Kacic, Z., Barczikay, and Tantoa, A., Csatári F., Bakcsi Zs. and Sfakianaki A., "A multimedia, multilingual teaching and training system for children with speech disorders", International Journal of Speech technology, 3, 289-300, 2001.

[9] Dragon Systems, "Dragon NaturallySpeaking SDK, C++ and SAPI Guide and Reference", Dragon Systems, 1999.

[10] Teixeira, C., Trancoso, I. and Serralheiro, A., "Recognition of non-native accents", In Proceedings of EUROSPEECH-1997, 2375-2378, 1997.

[11] Huang, X., Acero, A., and Hon, H. W., "Spoken Language Processing: A Guide to Theory, Algorithm, and System Development", Prentice Hall, 2001.

[12] Rudzicz, F. "Using articulatory likelihoods in the recognition of dysarthric speech. Speech Communication", 54:430–444, 2012.

[13] Kim, H., Hasegawa-Johnson, M., Perlman, A., Gunderson, J., Huang, T. S., Watkin, K., and Frame, S., "Dysarthric speech database for universal access research", In Interspeech, 1741-1744, 2008.

[14] Potamianos, A., Narayanan, S., and Lee, S. "Automatic speech recognition for children. In Eurospeech, 97, 2371-2374, 1997.

[15] Kewley-Port, D., Watson, C.S., Elbert, M., Maki, D., and Reed, D., "The Indiana speech training aid (ISTRA) II: Training curriculum and selected case studies", Clinical Linguistics and Phonetics, 5(1): 13-38, 1991.

[16] Bunnell, H.T. Yarrington, D.M. and Polikoff, J.B., "STAR: Articulation training for young children", In Proceedings of InterSpeech'00, 4, 85-88, 2000.

[17] Engwall, O. Analysis of and feedback on phonetic features in pronunciation training with a virtual teacher. Computer Assisted Language Learning, 25(1): 37–64, 2012.

[18] Engwall, O. and Bälter, O. Pronunciation feedback from real and virtual language teachers. Computer Assisted Language Learning, 20(3): 235–262, 2007.

[19] Öster, A-M. "Teaching speech skills to deaf children by computer-based speech training", In Proceedings of the 18th International Congress on Education of the Deaf, 1995.

[20] Ryalls, J., Michallet, B, and Le-Dorze, G., "A preliminary evaluation of the clinical effectiveness of vowel training for hearing-impaired children on IBM's Speech Viewer", Volta Review, 96, 19–30, 1994.

[21] Pratt, S, Heintzelman, A., and Deming, S., "The efficacy of using the IBM Speech Viewer vowel accuracy module to treat young children with hearing impairment", Journal of Speech and Hearing Research, 36, 1063–1074, 1993.

[22] Öster, A-M. "Applications and experiences of computer-based speech training", STL-QPSR, 1, 59-62, 1989.

[23] Öster, A-M. House D., Green P., "Testing a new method for training fricatives using visual maps in the Ortho-Logo-Pedia project (OLP)", Phonum 9- Fonetik, 89-92, 2003.

[24] Fell, H. J., MacAuslan, J., Gong, J., Cress, C. and Salvo, T. “visiBabble for pre-speech feedback”, CHI Extended Abstracts, 767-772, 2006.

[25] Fell, H. J., Cress, C., MacAuslan, J., Ferrier, L., J. “visiBabble for reinforcement of early vocalization”, ASSETS 2004: 161-168, 2004.

[26] Hailpern, J., Harris, A., La Botz, R., Birman, B., and Karahalios, K. “Designing visualizations to facilitate multisyllabic speech with children with autism and speech delays”, In Proceedings of the Designing Interactive Systems Conference, 126-135, 2012.

[27] Hailpern, J., Karahalios, K., DeThorne, L., and Halle, J. “Vocsyl: Visualizing syllable production for children with ASD and speech delays”, In Proceedings of the Assets 12th international ACM SIGACCESS conference on Computers and accessibility, 297-298, 2010.

[28] Hailpern, J., Karahalios, K., and Halle, J. “Creating a spoken impact: encouraging vocalization through audio visual feedback in children with ASD”, In Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, 453-462, 2009.

[29] Hamidi, F. “Using interactive objects for speech intervention”, SIGACCESS Access. Comput. 96, 28-31, 2010.

[30] Hincks, R., “Speech recognition for language teaching and evaluating: A study of existing commercial products”, In Proceedings of ICSLP’02, 733-736, 2002.

[31] Neri, A., Cucchiarini, C., and Strik, W., “Automatic speech recognition for second language learning: How and why it actually works”, In Proceedings of ICPhS’03, 257–1160, 2003.

[32] Mitra, S., Tooley, J., Inamdar, P. and Dixon, P., “Improving English pronunciation - an automated instructional approach”, *Information Technologies and International Development,* 1(1): 75-84, MIT Press, 2003.

[33] Morley J. “The pronunciation component in teaching English to speakers of other languages”, TESOL Quarterly, 25:481–520, 1991.

[34] Rosengren, E., Raghavendra, P., and Hunnicutt, S., “How does automatic speech recognition handle severely dysarthric speech?” in I. Placencia Porrero and P. de la Bellacas [Eds], In Proceedings of the 2nd TIDE Congress, IOS Press, 336–339, 1995.