Crowdsourcing for Research on Automatic Speech Recognition-enabled CALL

Catia Cucchiarini[1], Helmer Strik[1, 2, 3]
CLST[1], CLS[2], Donders[3]
Radboud University, Nijmegen, The Netherlands
{C.Cucchiarini, W.Strik}@let.ru.nl

Abstract

Despite long-standing interest and recent innovative developments in ASR-based pronunciation instruction and CALL, there is still scepticism about the added value of ASR technology. In this paper we first review recent trends in pronunciation research and important requirements for pronunciation instruction. We go on to consider the difficulties involved in developing ASR-based systems for pronunciation instruction and the possible causes for the paucity of effectiveness studies in ASR-based CALL. We suggest that crowdsourcing could offer solutions for analyzing the large amounts of L2 speech that can be collected through ASR-based CALL applications and that are necessary for effectiveness studies. We provide a brief overview of our own research on ASR-based CALL and of the lessons we learned. Finally, we discuss possible future avenues for research and development.

Keywords: Computer Assisted Language Learning, Automatic Speech Recognition, Pronunciation Instruction, Crowdsourcing

1. Introduction

Speaking skills have always been considered particularly challenging in language teaching, because of the time and individual attention they require for practice and feedback. This has been one of the reasons for the sustained interest in using Automatic Speech Recognition (ASR) technology in CALL applications. ASR technology has been around for more than 30 years and its potential for CALL has been emphasized from the beginning, but ASR-based CALL systems have not really found their way in language teaching contexts. This might have to do with a variety of factors. The relatively high costs involved in the development of new applications or in the acquisition of some commercial products might have been a hurdle to large-scale adoption, while for some products that are available for free privacy issues might have played a role. However, there is also another possible explanation for the general reluctance to embrace ASR technology in CALL. As a matter of fact, there are relatively few studies that have thoroughly investigated the effectiveness of ASR-based CALL in real-life environments, under realistic conditions with real users. This also applies to pronunciation instruction and training, which is the topic that has received most attention in ASR-based research and development, because of its potential for both language learning and speech therapy applications.

In the remainder of this paper we discuss the difficulties involved in developing ASR-based systems for pronunciation instruction, possible causes for the paucity of effectiveness studies and then consider possible solutions. In Section 2 we first discuss recent trends in pronunciation research and requirements for pronunciation instruction. We then consider important requirements for ASR-based CALL research in Section 3. Sections 4 and 5 provide a brief overview of our own research on ASR-based CALL and crowdsourcing, respectively. Discussion and conclusions are presented in Section 6 and 7.

2. Pronunciation Instruction

In pronunciation research there are different views on what the aim of pronunciation instruction should be. According to the “nativeness principle” (Levis, 2005: 370), pronunciation instruction should help L2 learners lose any traces of their L1 accent in order to achieve a nativelike accent.

The “intelligibility principle”, on the other hand, holds the view that pronunciation instruction should help L2 learners achieve intelligibility in the L2, which should be possible even if traces of an L1 accent remain. In line with this distinction, different constructs have been introduced in pronunciation research (Munro & Derwing, 1995a). Accent has been taken to refer to subjective judgments of the extent to which L2 speech is close to native speech and is usually expressed by scalar ratings. Intelligibility has been defined as the extent to which L2 speech can be correctly reproduced in terms of orthographic transcription (Munro & Derwing, 1995a). A third construct, comprehensibility, has been introduced to indicate the ease with which listeners understand L2 speech, again expressed through scalar ratings (Munro & Derwing, 1995a). Research has shown that communication can be successful even in the presence of a non-native accent (Munro & Derwing, 1995b). This combined with the knowledge that achieving a nativelike accent is beyond reach for most language learners, has led pronunciation researchers to advocate a focus on intelligibility in pronunciation instruction as opposed to nativeness (Levis, 2005; 2007; Munro & Derwing 2015).

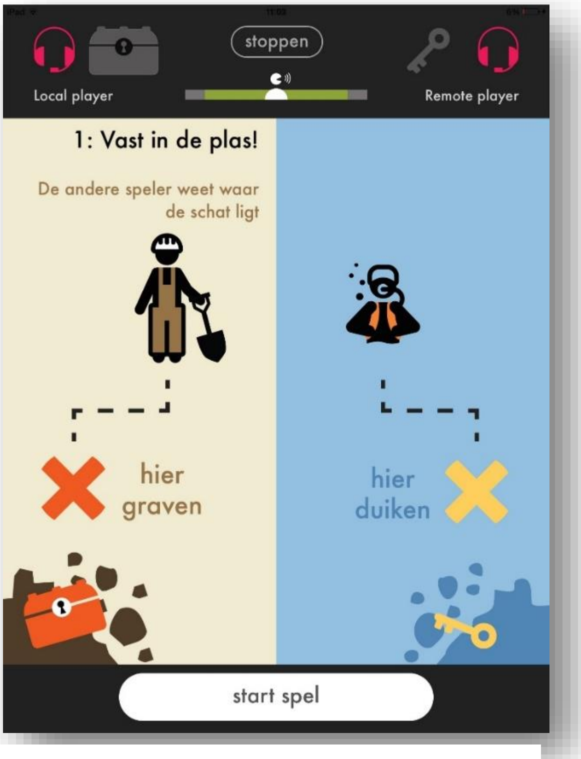
3. Requirements for ASR-based pronunciation research

In line with these distinctions, pronunciation researchers are interested in research that investigates to what extent ASR-based pronunciation instruction contributes to improving constructs such as accent, intelligibility or comprehensibility of L2 learners. However, convincing evidence is lacking (Thomson & Derwing, 2015). Most of the research on ASR-based pronunciation training has been conducted offline on annotated speech corpora (Cucchiarini & Strik, 2017). In general, such studies evaluate the accuracy of specific algorithms (Stanley, Hacioglu, & Pellom, 2011; Qian, Meng, Soong, 2012; Lee, Zhang, & Glass, 2013) in identifying pronunciation errors or in grading L2 speech. To investigate the effectiveness of ASR-based CALL complete systems are needed, in which these algorithms are incorporated to provide speaking practice and feedback on the utterances produced by L2 learners under realistic conditions. In addition, a certain amount of learning content is needed so that learners can practice for a sufficient amount of time. It is the kind of

longitudinal research that is needed to increase our understanding of the contribution of ASR-based CALL to pronunciation teaching and language learning in general. Unfortunately, there are not so many complete systems that employ ASR and that could be used in open, online effectiveness research in real life conditions. This has to do with a series of difficulties (Cucchiarini & Strik, 2017).
First of all, the limited availability of large corpora that can be used to develop, test and optimize the specific speech technology that is required for learning applications. Another difficulty is related to the nature of the expertise required, which is highly varied and interdisciplinary as it covers engineering, system design, pedagogy and language learning. This can also pose problems in finding the necessary funds for this type of cross-disciplinary research.

4. Our own research on ASR-based CALL

In our own research over the last twenty years we have pursued the goal of developing complete ASR-based CALL systems. This research has been conducted in close cooperation with speech technologists, language learning researchers and teachers. The aim was to develop systems that could be used to conduct more comprehensive research contributing insights to both speech technology and language learning research (Cucchiarini et al. 2009, 2011, 2014; Strik, 2012; Strik et al. 2012; Van Doremalen et al., 2010, 2013; 2016). An important aspect in this research was also how to boost user motivation either by providing appealing, useful feedback (Bodnar et al., 2016, 2017; Cucchiarini et al., 2009; Penning de Vries et al., 2015, 2016, 2019) or by introducing gaming elements, see e.g. Figure 1 (Ganzeboom et al. 2016).



1: Vast in de plas!

De andere speler weet waar de schat ligt

Figure 1: In “treasure hunters”, serious gaming is used to motivate patients to practice for ASR-based speech therapy (Ganzeboom et al. 2016).

The more recent systems have been equipped with logging capabilities (Bodnar et al., 2017; Penning de Vries et al., 2016), so that they can collect huge amounts of speech data produced by L2 learners practicing with the system, while at the same time recording all system-user interactions. These logged data can provide useful knowledge on learners’ progress, increasing our insights not only into the ultimate outcome of learning, but also into the processes that are conducive to learning.
One of the problems we have encountered in this research is, however, how to process and analyze these large sets of speech data that are produced by language learners or patients during practice or therapy and that need to be scored and analyzed to study the effectiveness of ASR-based applications. To be able to provide information on learning and effectiveness, these data need first of all to be transcribed and/or scored, to obtain the subjective judgments necessary to measure the constructs mentioned above (accent, intelligibility, comprehensibility). This is extremely time-consuming and expensive. In fact, the amount of data is such that manual annotations are actually not feasible. A possible alternative solution to obtain annotations and scoring of vast amounts of speech data at relatively low costs would then seem to be to employ crowdsourcing, as will be explained in the next section.

5. Crowdsourcing for ASR-based CALL

In ASR-based CALL pronunciation research crowdsourcing could play a more prominent role by providing transcriptions or intelligibility scores, which can in turn be used for effectiveness evaluation. In our own research, for example, we have used crowdsourcing to obtain evaluations of intelligibility of L2 learner speech (Burgos et al., 2015, Sanders et al., 2016) and pathological speech (Ganzeboom et al., 2016).
For the study described in Ganzeboom et al. (2016) an online listening experiment was carried out. Participants were invited by email or via Facebook. They filled in a questionnaire to gather some meta-information about native language, gender, age, etc. In total 36 listeners participated, 8 male and 28 female (age range 19-73), who rated 50 utterances on intelligibility in three ways:

- Likert: 1. very low, to 7. very high
- Visual Analogue Scale (VAS): 0. very low, to 100. very high
- Orthographic Transcription (Orthog. Transc.)
- The latter was used to calculate three extra scores:
- OTW = Orthog. Transc. scored at Word level
- OTP = Orthog. Transc. scored at Phoneme level
- OTG = Orthog. Transc. scored at Grapheme level

VAS and Likert are intelligibility scores on utterance level and were calculated as scores representing a percentage (%) of intelligibility. The VAS scores were already on a 0-100 scale, while the scores on the 1-7 Likert scale were transformed to percentage scores by first subtracting 1 and then multiplying by 16.67 (i.e. 1=0%, 2=16.67%, 3=33%, ..., 7=100%).
To obtain an intelligibility score at word level (OTW), we compared the raters’ orthographic transcriptions to the reference transcriptions, we counted the number of identical word matches and calculated a percentage correct score.

Intelligibility scores at the grapheme and phoneme level (OTG and OTP, resp.) were automatically obtained from the orthographic transcriptions through the Algorithm for Dynamic Alignment of Phonetic Transcriptions (ADAPT) (Elffers, et al. 2013) which computes the optimal alignment between two strings of phonetic symbols using a matrix that contains distances between the individual phonetic symbols. For the intelligibility scores on phoneme level (OTP), the orthographic transcriptions were converted to their phonemic equivalent using the canonical pronunciation variants from the lexicon of the Spoken Dutch Corpus (Oostdijk, 2000). Some results are presented in Table 1. For more details see Ganzeboom et al. (2016).

n = 50	M (SD)				
		VAS	OTW	OTP	OTG
Likert	63.1 (21.1)	.998	.733	-.763	-.773
VAS	63.2 (19.0)		.732	-.755	-.764
OTW	78.3 (16.1)			-.805	-.869
OTP	8.0 (6.5)				.954
OTG	8.9 (7.4)				

Table 1: Means (SDs) and correlations of the five intelligibility measures (n = 50 speech fragments).

For Likert, VAS and OTW, higher scores correspond to higher intelligibility (higher percentage correct); for OTP and OTG lower scores correspond to lower distance and thus higher intelligibility. All correlations were significant (p < .01).

Important for research data in general, and especially for data obtained by means of crowdsourcing, is their reliability. In our study the reliability of each of the five intelligibility measures was calculated using Intraclass Correlation Coefficients (ICC) based on groups of raters. The ICC values for all 36 raters together were very high, ranging from .95 (OTP, OTG) to .97 (Likert, VAS, OTW). As such a large number of raters may not always be achievable, we also calculated average ICCs based on randomly selected smaller subsets of the data (e.g. 9 subsets of 4 raters, or 6 of 6 raters). On average, for the utterance and word level scorings sufficient reliability is obtained with four raters (resulting in mean ICC values ranging from .79 to .84), while for subword scorings at least six raters are required (resulting in mean ICC values ranging from .79 to .80).

Fig. 2. Crowdsourcing experiment Palabras. At the end, participants can share their final score on Facebook.

In the L2 speech crowdsourcing experiment Palabras (see Figure 2), a web application was developed for obtaining transcriptions of Dutch words spoken by Spanish L2 learners that was accessible via Facebook. Participants would listen and write down what they heard. Different types of feedback were provided, like percentage correct, words still to transcribe and the majority transcription (Sanders et al. 2016).

Also in this case the quality of the data was checked by applying filters to remove transcribers who did not conform to our quality criteria (with other native languages than Dutch, who did not reach our threshold of intra and inter transcriber agreement, who entered more than once when the server was slow in response). In total useful data were obtained from 159 participants, which is definitely more than would have been the case with traditional experiments.

6. Discussion

So far crowdsourcing has been mainly used to produce language resources like learner speech corpora (Eskenazi et al., 2013), to obtain speech recordings with annotations (Loukina et al. 2015a, b), or to collect more complex and realistic speech data such as dialogues through conversational technologies (Sydorenko et al. 2018).

The experiences described in Section 5 would seem to be good reasons for extending the use of crowdsourcing to the larger sets of data that are obtained through the loggings in ASR-based CALL systems. These would constitute an enormous rich source of information for improving both the technology and the learning systems. In addition, these annotated data and speech files could be used to further train and adapt the algorithms employed in the system and thus to enhance the quality of the ASR technology.

This approach could be extended to ASR-based CALL that addresses other aspects of L2 speaking to obtain annotations of learner speaking performance, evaluations of L2 proficiency in grammar and vocabulary or of turn taking abilities, pragmatic competence, politeness strategies and formulaic language in spoken dialogue applications. An additional solution could be so-called implicit crowdsourcing, which could be applied by collecting additional speech data and subjective evaluations when users engage with ASR-based CALL systems. In other words, in this case the users of CALL systems would form the crowd. There are some important caveats to be taken into account, though. First of all, GDPR puts limitations to using spoken data in crowdsourcing as speech data are by definition sensitive data. Speech intrinsically contains information on identity and other personal features. Speech corpora often impose restrictions to making speech fragments audible to the public. In any case prior explicit consent has to be obtained for employing user data for research and development purposes. Finally, the reliability of the subjective data obtained through crowdsourcing has to be checked before these data are used for further research.

7. Conclusions

ASR-based CALL applications hold great potential for innovative research on language learning and future developments for language teaching. Effectiveness studies could help clarify their added value, but so far these studies have been few and far between, among other things because they require subjective judgments of large amounts of L2 speech. Crowdsourcing can be usefully applied for this purpose. For the two crowdsourcing initiatives described in section 5, the results were satisfactory as larger sets of data could be annotated and scored than would have been the case with traditional experiments. In turn these data provided useful insights into important aspects of intelligibility scoring measures with different degrees of granularity. To conclude, there seem to be good reasons for extending this approach to ASR-based CALL that addresses other aspects of L2 speaking to obtain much wanted subjective annotations and evaluations of learner speaking performance.

8. Bibliographical References

Bodnar, S., Cucchiarini, C., Penning de Vries, B., Strik, H., & van Hout, R. (2017). Learner affect in computerised L2 oral grammar practice with corrective feedback. Computer Assisted Language Learning, 30, 223-246.

Bodnar, S.E., Cucchiarini, C., Strik, H. & Hout, R.W.N.M. van (2016). Evaluating the motivational impact of CALL systems: current practices and future directions. Computer Assisted Language Learning, 29, (1), 186-212.

Burgos, P.; Sanders, E.; Cucchiarini, C.; Hout, R. van; Strik, H. (2015) Auris populi: crowdsourced native transcriptions of Dutch vowels spoken by adult Spanish learners. In: Proc. of Interspeech 2015, 2819-2823.

Cooke, M., Barker, J., & Lecumberri, M. L. G. (2013). Crowdsourcing in speech perception. In: M. Eskenazi, G-A. Levow, H. Meng, G. Parent & D. Suendermann (Eds.), Crowdsourcing for speech processing: Applications to data collection, transcription and assessment (pp. 137-172). Somerset, GB: Wiley.

Cucchiarini, C., Bodnar, S., Penning de Vries, B., van Hout, R., & Strik, H. (2014). ASR-based CALL systems and learner speech data: new resources and opportunities for research and development in second language learning. Proceedings of LREC, Reykiavik.

Cucchiarini, C., Heuvel, H. van den, Sanders, E.P. & Strik, H. (2011). Error selection for ASR-based English pronunciation training in 'My Pronunciation Coach'. Proceedings of Interspeech, 1165-1168, Florence, Italy.

Cucchiarini, C, Neri, A., Strik, H. (2009). Oral proficiency training in Dutch L2: The contribution of ASR-based corrective feedback. Speech Communication, 51 (10), 853-863.

Cucchiarini, C., Strik, H. (2017). Automatic speech recognition for L2 pronunciation assessment and training. In O. Kang, R. Thomson & M. Murphy (Eds.) The Routledge handbook of English pronunciation.

Derwing, T. M., & Munro, M. J. (2015). Pronunciation fundamentals: Evidence-based perspectives for L2 teaching. Amsterdam: John Benjamins.

Elffers, B., van Bael, C., and Strik, H. (2013). ADAPT: Algorithm for Dynamic Alignment of Phonetic Transcriptions. Internal report, CLST, Radboud University Nijmegen, The Netherlands.

Eskenazi, M., G. Levow, H. Meng, G. Parent & D. Suendermann (eds.) (2013). Crowdsourcing for speech processing: Applications to data collection, transcription assessment. New York: Wiley.

Ganzeboom, M.S.; Bakker, M.; Cucchiarini, C.; Strik, H. (2016) Intelligibility of Disordered Speech: Global and Detailed Scores. In: Proceedings of Interspeech 2016, pp. 2503-2507; San Francisco, CA, USA.

Hu, W., Qian, Y., Soong, F.K., Wang, Y. (2015). Improved mispronunciation detection with deep neural network trained acoustic models and transfer learning based logistic regression classifiers. Speech Communication, 67, 154-166.

Lee, Y. Zhang, & J. Glass, (2013). Mispronunciation detection via Dynamic Time Warping on Deep Belief Network-based posteriorgrams. Proceedings ICASSP 2013, Vancouver, BC, 8227–8231.

Levis, J.M. (2005). Changing contexts and shifting paradigms in pronunciation teaching. TESOL Quarterly, 39(3), 369-377.

Levis, J. (2007). Computer technology in teaching and researching. Annual Review of Applied Linguistics, 27, 184–202.

Loukina, A., Lopez, M., Evanini, K., Suendermann-Oeft, D., Zechner, K. (2015). Expert and crowdsourced annotation of pronunciation errors for automatic scoring systems, Proceedings INTERSPEECH-2015, 2809-2813.

Munro, M. J., & Derwing, T. M. (1995a). Foreign accent, comprehensibility, and intelligibility in the speech of second language learners. Language Learning, 45, 73-97.

Munro, M. J., & Derwing, T. M. (1995b). Processing time, accent, and comprehensibility in the perception of native and foreign accented speech. Language and Speech, 38, 289–306.

Munro, M. J., Derwing, T. M., & Thomson, R. I. (2015). Setting segmental priorities for English learners: Evidence from a longitudinal study. Int. Review of Applied Linguistics in Language Teaching, 53(1), 39-60.

Oostdijk, N. (2000). The Spoken Dutch Corpus: Overview and first evaluation. Proceedings of LREC 2000, 886–894, Athens, Greece.

Penning de Vries, B.W.F., Cucchiarini, C., Bodnar, S.E., Strik, H. & Hout, R.W.N.M. van (2015). Spoken grammar practice and feedback in an ASR-based CALL system. Computer Assisted Language Learning, 28 (6), 550-576.

Penning de Vries, B., Cucchiarini, C., Bodnar, S., Strik, H., & van Hout, R. (2016). Effect of corrective feedback for learning verb second, Int. Review of Applied Linguistics in Language Teaching (IRAL), 54(4), 347-386.

Penning de Vries, B., Cucchiarini, C., Strik, H., van Hout, R. (2019). Spoken grammar practice in CALL: The effect of corrective feedback and education level in adult L2 learning, Language Teaching Research.

Qian, X., Meng, H., Soong, F. (2012). The Use of DBN-HMMs for Mispronunciation Detection and Diagnosis in L2 English to Support Computer-Aided Pronunciation Training. In: Proc. of Interspeech, 775-778, Portland.

Sanders, E.P.; Burgos, P.; Cucchiarini, C.; Hout, R.W.N.M. van (2016) Palabras. Crowdsourcing transcriptions of L2 speech. In: Proceedings of the Int. Conf. on Language Resources and Evaluation (LREC) 2016, pp. 3186-3191.

Stanley, T., Hacioglu, K & Pellom, B. (2011). Statistical machine translation framework for modeling phonological errors in computer assisted pronunciation training system. SLaTE 2011, Venice, Italy.

Strik, H. (2012). ASR-based systems for language learning and therapy. International Symposium on Automatic

Detection of Errors in Pronunciation Training (IS-Adept). KTH: Stockholm, Sweden, June 6-8.

Strik, H. Colpaert, J., Van Doremalen, J. & Cucchiarini, C. (2012). The DISCO ASR-based CALL system: practicing L2 oral skills and beyond. Proceedings LREC, Istanbul.

Sydorenko, T., Smits, T., Evanini, K. & Ramanarayanan, V. (2018). Simulated speaking environments for language learning: insights from three cases. Computer Assisted Language Learning.

Thomson, R. I., & Derwing, T. M. (2015) The effectiveness of L2 pronunciation instruction: A narrative review. Applied Linguistics, 36(3), 326–344.

Van Doremalen, J., Boves, L., Colpaert, J., Cucchiarini, C., Strik, H. (2016). Evaluating ASR-based language learning systems: A case study. Computer Assisted Language Learning, 29(4), 833-851.

Van Doremalen, J., Cucchiarini, C., & Strik (2010). Optimizing automatic speech recognition for low-proficient non-native speakers. EURASIP Journal on Audio, Speech, and Music Processing.

Van Doremalen, J., Cucchiarini, C., & Strik (2013). Automatic pronunciation error detection in non-native speech: the case of vowel errors in Dutch. Journal of the Acoustical Society of America, 134, 1336-1347.