Towards the development of a comprehensive pedagogical framework for pronunciation training based on adapted automatic speech recognition systems

Saandia Ali[1]

Abstract. This paper reports on the early stages of a locally funded research and development project taking place at Rennes 2 university. It aims at developing a comprehensive pedagogical framework for pronunciation training for adult learners of English. This framework will combine a direct approach to pronunciation training (face-to-face teaching) with online instruction using and adapting existing Automatic Speech Recognition systems (ASR). The sample of learners chosen for the study are university students majoring in Arts, Literature or Communication at graduate and undergraduate level. These students might show an advanced mastery of grammar and syntax, but their spoken English remains heavily accented and may hinder effective communication. A considerable body of research has already investigated the efficacy of ASR systems for pronunciation training. This paper takes stock of how Computer Assisted Pronunciation Training (CAPT) software has been used and developed so far and looks at further potential improvements to address bad pronunciation habits among French learners of English.

Keywords: computer assisted pronunciation training, automatic speech recognition, CALL system design, ESL.

1. Introduction

Pronunciation is an area of teaching which is often neglected, probably because teachers lack time and often resources to enable them to tackle phonetic and phonological competences. In most French universities, classes are overcrowded (up

1. Rennes 2 University, Rennes, France and Jean Jaures University, Toulouse, France; saandia.vanessa.ali@gmail.com

How to cite this article: Ali, S. (2016). Towards the development of a comprehensive pedagogical framework for pronunciation training based on adapted automatic speech recognition systems. In S. Papadima-Sophocleous, L. Bradley & S. Thouësny (Eds), *CALL communities and culture – short papers from EUROCALL 2016* (pp. 7-13). Research-publishing.net. https://doi.org/10.14705/rpnet.2016.eurocall2016.530

to 40 students per group) and the emphasis is placed on fluency and communication skills rather than phonetic accuracy. In addition to this observation, most teachers do not feel confident with teaching pronunciation as they often haven't received any training themselves.

Under these circumstances, students experience performance anxiety, and they only have a limited amount of time for teacher-student interaction and individualized feedback. As mentioned by Eskenazi (1999), "[l]anguage learning appears [to be] most efficient when the teacher constantly monitors progress to guide […] remediation or advancement" (p. 450).

CAPT programs (Abuseileek, 2007) could help realise these goals by offering individual practice and feedback in a safe environment. Recent ASR based CAPT programs include Subarashii (Entropic HTK recognizer), VILTS (SRI recognizer), FLUENCY (Carnegie Mellon University SPHINX recognizer), Naturally Speaking (Dragon Systems), and FluSpeak (IBM ViaVoice recognizer).

We intend to build on these existing programs and on previous research to develop a set of tools to address bad pronunciation habits among French learners of English. In an attempt to do so, the rest of this paper will elaborate on the following questions:

- How have ASR systems been used to teach pronunciation?
- What improvements are still needed to develop an ideal pronunciation training framework for French learners of English?

2. Using ASR systems for pronunciation training: an overview of existing tools and previous research

2.1. Smartphone commercial apps

The simple act of googling 'pronunciation training apps' shows the considerable number of tools and software available to help people acquire good pronunciation. Two main types of pronunciation training apps can be found: those that target a wide variety of users ranging from students to other users, including tourists or occasional users, and those that were developed by teachers or researchers specialising in the domain of language learning and teaching. The first type of apps

(see for example pronunciation checker[2], English pronunciation checker[3] or Vowel Viz[4] for iPhones) are used to check or verify the accuracy of one's pronunciation in a number of contexts. Pronunciation checker is a multilingual app based on databases of up to 1000 words for each targeted language and enables the user to listen to the production of a word, practise saying it via a recording device and then obtain an evaluation of the resulting production as a 'score'. There are two proficiency levels: easy and hard. This type of app usually lacks depth and doesn't include any linguistic or didactic information as an input or as a diagnosis, which is often limited to a numerical score.

The second kind of app is based on more in-depth linguistic and sometimes pedagogical content (see for example Sounds pronunciation apps[5] by Macmillan, English pronunciation[6] by Kepham or Speech Ace[7]). Most of these apps focus on pronunciation training at segment level and include pre-training tasks and content which revolve around interactive phonemic charts and illustrated descriptions of the articulatory features of the sounds of English. Recording facilities are also included along with diagnoses of learners' productions that can be compared with targeted productions in the chosen model (often US or UK English).

2.2. Experimental research aiming at CAPT software development

Numerous studies have tackled the question of ASR efficacy for CAPT (see e.g. Hinks, 2001). In this section, we present a brief overview of three representative studies (i.e. Elimat & Abuseileek, 2014; Escudero & Tejedor-Garc, 2015; Kim, 2006) that led to the development and testing of experimental ASR-based software for pronunciation training in English. They provide three different examples of how ASR systems and pronunciation teaching strategies can be tested and reveal the remaining challenges of current ASR technology.

Escudero and Tejedor-Garc (2015) introduce the architecture and interface of a serious game intended for pronunciation training and assessment of Spanish students of English as a second language. Android ASR and text to speech tools make it possible to discern three different pronunciation proficiency levels, ranging from basic to native. The authors use minimal pairs to promote learners' awareness

2. https://play.google.com/store/apps/details?id=com.app.pronunciation_checker
3. https://play.google.com/store/apps/details?id=com.eapp.pc
4. https://itunes.apple.com/us/app/vowelviz/id740035896?mt=8
5. https://play.google.com/store/apps/details?id=com.macmillan.app.soundsfree
6. https://play.google.com/store/apps/details?id=com.study.english.pronunciation
7. http://www.speechace.com/

of the potential misunderstandings and wrong meanings that can result from too approximate productions of phonemes.

Elimat and Abuseileek (2014) use the 'Tell me more performance' program to test the efficacy of ASR systems as well as various teaching techniques (i.e. individual work, pair work, group work) to train third grade learners of English. The best results were obtained with the group of students who worked individually with the ASR system.

The study of Kim (2006) resulted in the creation of Fluspeak, which is an ASR based pedagogical software used to teach US English pronunciation. It was tested with 36 university students through a hybrid teaching approach mixing Face to face teaching with individual work with the software. The study included a comparison between human scoring and automatic scoring with Fluspeak. Although Fluspeak gave good results with beginners focussing on phoneme production, it gave poor results overall for advanced learners trying to gain fluency.

On the whole and to our knowledge, most CAPT softwares show promising results and very positive impacts on the pronunciation of segmental sounds among various types of learners. Prosodic features and fluency generally speaking are areas of pronunciation training that still seem to require further research and development.

2.3. Towards enriching an ASR based pronunciation training system with linguistic and pedagogical content

Drawing conclusions from previous research and from an evaluation of commonly used CAPT software, this section provides an outline of the intended enrichment and development steps that need to be taken to develop a comprehensive pedagogical framework for pronunciation training. Three main steps were identified:

- selecting an open source ASR system to be adapted and further enriched to suit our purposes;
- enriching input data with prosodic information: selecting prosodically labelled corpora (L1 English, L1 French, L2 French);
- providing didactized content: targeted feedback, diagnosis, post-task courses and further practice.

Several open source speech recognition toolkits are available for research and development (see (Gaida et al., 2014; Povey et al., 2011). Gaida et al. (2014), for instance, compare the most commonly used open source softwares and show that the Kaldi toolkit is the most efficient and easier to adapt than the CMU sphinx toolkit for instance.

The common approach to recognize speech is to take a waveform, split it in utterances by silences and then try to recognize what is being said in each utterance. In order to do so, all possible combinations of words need to be tested and matched with the audio so as to select the best matching combination. Three models are used to complete the matching process: the acoustic model (acoustic properties for each phoneme of the target language), the phonetic model or phonetic dictionary (with the mapping from word to phone) and a language model (defining which word can follow another and restrict possible combinations).

Starting from the Kaldi toolkit, prosodic information can be added at the level of the acoustic model, which is usually based on large corpora annotated at segment level. We propose to use our own corpus developed in previous studies (see Ali, 2010; Ali & Hirst, 2009) to train Kaldi with prosodically annotated data in English. The chosen intonation model for this corpus is defined in Hirst and DiCristo (1998) and based on automatic modeling of rhythm and intonation via the Momel-Intsint algorithm (see Hirst & Espesser, 1993). Learner corpora (Diderot Longdale corpus and CIL corpus) will also be used to train the ASR system to recognize the productions of French learners of English at various proficiency levels (beginner, intermediate, advanced).

Once the recognition process has successfully taken place, pedagogical content will be added. Three kinds of tasks will be introduced:

- reading tasks based on isolated words (to assess phoneme production in monosyllabic words and word stress in polysyllabic words);
- reading tasks based on full utterances (to assess rhythm and intonation);
- conversation and guided interaction with virtual agents (to develop interaction skills, fluency and discourse level prosodic features).

Explicit feedback and diagnosis will be provided for each type of task using recording facilities along with Praat and Momel-Intsint representations to visualize productions and compare them to the target models.

3. Conclusion

Related studies such as Elimat and Abuseileek (2014) have shown that the ideal ASR software for CAPT should include at least five phases: ASR, automatic scoring on the basis of the comparison between a student's utterance and a native's utterance, error detection and error diagnosis. Starting from these essential characteristics and an evaluation of existing software, further improvements and preliminary steps were proposed in this paper in an attempt to develop a pronunciation training framework for French learners of English. The first steps mainly consist in enriching an existing open source ASR system with prosodic information to tackle the limitations of ASR tools when used to provide feedback at sentence and discourse level. This could be achieved by training ASR systems with both native and non-native speakers' prosodically labelled corpora. Further steps include the provision for enriched pedagogical content once the recognition process has successfully taken place.

References

Abuseileek, A. (2007). Computer-based pronunciation instruction as an effective means for teaching stress. *The jalt call Journal, 3*(1-2), 3-14.

Ali, S. (2010). Etude de la relation entre l'annotation des formes et des fonctions en anglais britannique contemporain. Linguistique. *Université de Provence-Aix-Marseille I. HAL Id: tel-00460431*. https://tel.archives-ouvertes.fr/tel-00460431

Ali S., & Hirst, D. (2009). Developing an automatic functional annotation system for British English Intonation. *Proceedings of Interspeech 2009, Brighton, 2207-2210*.

Elimat, A. K., & Abuseileek, A. F. (2014). Automatic speech recognition technology as an effective means for teaching pronunciation. *The JALT CALL Journal, 10*(1), 21-47.

Escudero, D., & Tejedor-Garc, C. (2015). Implementation and test of a serious game based on minimal pairs for pronunciation training pronunciation training. *Proceedings of SLaTE 2015* (pp.125-130).

Eskenazi, M. (1999). Using a computer in foreign language pronunciation training: what Advantages? *Calico Journal, 16*(3), 447-469.

Gaida, C., Lange, P. L., Petrick, R., Proba, P., Malatawy, A., & Suendermann-Oeft, D. (2014). Comparing open-source speech recognition toolkits. *Technical Report, 12*.

Hinks, R. (2001). Using speech recognition to evaluate skills in spoken English. *Working Papers, 49* (pp. 58-61). Lund University, Department of Linguistics.

Hirst, D. J., DiCristo, A. (1998). *Intonation systems: a survey of twenty languages*. Cambridge: Cambridge University Press.

Hirst, D. J., & Espesser, R. (1993). Automatic modelling of fundamental frequency using a quadratic spline function. *Travaux de l'Institut de Phonétique d'Aix 15* (pp. 75-85).

Kim, I. S. (2006). Automatic speech recognition: reliability and pedagogical implications for teaching pronunciation. *Educational Technology and Society, 9*(1), 322-334.

Povey, D., Ghoshal, A., Boulianne, G., Burget, L., Glembek, O., Goe,l N., Hannemann, M., Motlicek, P., Qian, Y., Schwarz, P., Silovsky, J., Stemmer, G., & Vesely, K. (2011). The Kaldi speech recognition toolkit. *Proceedings of the ASRU, 4*.

Research-publishing.net

Published by Research-publishing.net, not-for-profit association
Dublin, Ireland; Voillans, France, info@research-publishing.net

CALL communities and culture – short papers from EUROCALL 2016
Edited by Salomi Papadima-Sophocleous, Linda Bradley, and Sylvie Thouësny

Typeset by Research-publishing.net
Cover design by © Easy Conferences, info@easyconferences.eu, www.easyconferences.eu
Cover layout by © Raphaël Savina (raphael@savina.net)
Photo "bridge" on cover by © Andriy Markov/Shutterstock
Photo "frog" on cover by © Fany Savina (fany.savina@gmail.com)
Fonts used are licensed under a SIL Open Font License

ISBN13: 978-1-908416-43-8 (Paperback - Print on demand, black and white)
Print on demand technology is a high-quality, innovative and ecological printing method; with which the book is never 'out of stock' or 'out of print'.

ISBN13: 978-1-908416-44-5 (Ebook, PDF, colour)
ISBN13: 978-1-908416-45-2 (Ebook, EPUB, colour)

Legal deposit, Ireland: The National Library of Ireland, The Library of Trinity College, The Library of the University of Limerick, The Library of Dublin City University, The Library of NUI Cork, The Library of NUI Maynooth, The Library of University College Dublin, The Library of NUI Galway.

Legal deposit, United Kingdom: The British Library.
British Library Cataloguing-in-Publication Data.
A cataloguing record for this book is available from the British Library.

Legal deposit, France: Bibliothèque Nationale de France - Dépôt légal: décembre 2016.