Language Learning & Technology
ISSN 1094-3501 CC BY-NC-ND

2023, Volume 27, Issue 1
pp. 1–19

ARTICLE

Automatic pronunciation assessment vs. automatic speech recognition: A study of conflicting conditions for L2-English

Enrique Cámara-Arenas, University of Valladolid

Cristian Tejedor-García, Radboud University Nijmegen

Cecilia Judith Tomas-Vázquez, University of Valladolid

David Escudero-Mancebo, University of Valladolid

Abstract

This study addresses the issue of automatic pronunciation assessment (APA) and its contribution to the teaching of second language (L2) pronunciation. Several attempts have been made at designing such systems, and some have proven operationally successful. However, the automatic assessment of the pronunciation of short words in segmental approaches has still remained a significant challenge. Free and off-the-shelf automatic speech recognition (ASR) systems have been used in integration with other tools with the hopes of facilitating improvement in the domain of computer-assisted pronunciation training (CAPT). The use of ASR in APA stands on the premise that a word that is recognized is intelligible and well-pronounced. Our goal was to explore and test the functionality of Google ASR as the core component within a possible automatic British English pronunciation assessment system. After testing the system against standard and non-standard (foreign) pronunciations provided by participating pronunciation experts as well as non-expert native and non-native speakers of English, we found that Google ASR does not and cannot simultaneously meet two necessary conditions (here defined as intrinsic and derived) for performing as an APA system. Our study concludes with a synthetic view on the requirements of a reliable APA system.

Keywords: *Automatic Pronunciation Assessment (APA), Automatic Speech Recognition (ASR), Automatic Assessment Tools, Second Language (L2) Pronunciation*

Language(s) Learned in This Study: *English*

APA Citation: Cámara-Arenas, E., Tejedor-García, C., Tomas-Vázquez, C. J., & Escudero-Mancebo, D. (2023). Automatic pronunciation assessment vs. automatic speech recognition: A study of conflicting conditions for L2-English. *Language Learning & Technology*, *27*(1), 1–19.
https://hdl.handle.net/10125/73512

Introduction

Automatic speech recognition (ASR) systems are proliferating, probably because they aim at a very desirable goal: allowing direct verbal communication with electronic devices (computers, smartphones, home automation systems, etc.) without interfering keyboards or tactile screens (Becker & Nguyen, 2017; McCrocklin, 2016; O'Brien, et al., 2018). For English as a second language (ESL) learners, ASR systems present unexpected potentialities resting on a commonsensical assumption that if they make themselves understood by an English ASR, their English is at least intelligible in terms of pronunciation. Therefore, recognition is translated into a positive assessment of performance in a second language, one would claim (Chen & Li, 2016). Available ASR technology (often free and off-the-shelf), if adapted properly, might help teachers with pronunciation assessment, especially as it would allow them to assess many students

simultaneously, unaffected by tiredness or subjectivity (Neri et al., 2010).

Automatic pronunciation assessment (APA) constitutes an emerging field of interest. There are several studies which have addressed the automatic assessment of the pronunciation of words (e.g., Luo, 2016; Tejedor-García, 2020) and sentences (e.g., Crossley & McNamara, 2013; Yarra et al., 2019). Teaching institutions are using automatic systems for the assessment of L2 spoken discourse (Seed & Xu, 2018). The important point is that while assessing contextualized words/sentences seems possible, the assessment of isolated words has remained a challenge (Cheng, 2018). However, improvement through APA-feedbacked training has been proved to be comparable to that resulting from classroom instruction (Luo, 2016; Tejedor-García 2020) as a direct correlation was reported between APA and human ratings of proficient speech (Crossley & McNamara, 2013). Human-machine correlation coefficient is in fact higher than inter-rater correlation (Cheng, 2018).

In incorporating APA modules, many computer-assisted pronunciation training (CAPT) experts have turned to ASR technology which has reached remarkable levels of reliability (Hasan et al., 2017). It has been reported that such CAPT systems indeed mediate pronunciation improvement (Kukulska-Hulme, 2012; Rahimi, 2015). These systems operate upon the premise that a word recognized by the system is a word that was properly pronounced, while defective pronunciations will not be properly recognized (Thomson, 2012). From the user's point of view, the response of ASR is binary: recognition vs. nonrecognition/misrecognition. Any corrective feedback requires the intervention of another system or a human agent (Cucchiarini & Strik, 2018). Therefore, if ASR recognizes 'seat' where 'sit' was targeted, a second system might inform the user that they must produce an opener and more centralized vowel. For the purpose of assessment, however, a binary response minimally interfaced seems to be rudimentary but operational.

Although ASR-based APA within CAPT seems efficient (Liakin et al., 2015; Luo, 2016; Tejedor-García, 2020), the implicit premise that justifies its use has only been partially tested. Researchers in the field have mostly been concerned with intelligibility (Jenkins, 2002; Levis, 2005) and therefore with the extent to which human and automatic recognition overlap (Coniam, 1999; Derwing et al., 2000; McCrocklin & Edalatishams, 2020). Humans tend to understand accented speech better than ASR systems, although the latter are getting better at it. ASR systems have only been rarely confronted with the challenge of assisting with accent reduction in learners who have attained intelligibility and still feel an urge for improvement (Hermans & Sloep, 2018). In a preliminary approximation to the issue, we contend that the effectiveness of ASR technology for accent-reduction APA depends upon its satisfying two elemental conditions: an intrinsic condition and a derived condition.

It is to be expected from an English ASR system that whenever presented with the native rendering of an English word (with non-to-minimally transferred pronunciation), the system recognizes it (i.e., it returns the orthographic renderings of the input). This condition is intrinsic to all ASR systems: they must recognize all intelligible expressions (Adda-Decker & Lamel, 2000; Liakin et al., 2015). But in order to be effective as an accent-reduction APA tool, an ASR system should fail to recognize strongly transferred realizations. This condition is clearly not intrinsic to the system as ASR systems are designed to recognize and they tend to always recognize something (Liakin et al., 2015). An ASR-based accent-reduction APA system must then satisfy a derived condition, meaning it must filter out/qualify transferred pronunciations. This function can be extraneously imposed on an ASR system by a mediating system that re-interprets nonrecognitions or misrecognitions as negative assessments.

These conditions are in conflict with the core purpose of ASR . The intrinsic condition purports that the ASR is able to recognize expressions rendered in a diversity of voices, timbres, accents, contexts, and so forth. However, the derived condition demands that the ASR system is clever enough to refrain from this reductionistic tendency whenever confronted with transferred pronunciations. At this time, conventional ASR systems are unlikely to be able to effectively integrate both demands when functioning within APA systems. We intend to move beyond this initial deductive stance and determine the reliability of Google ASR as the core system within accent-reduction APA protocols for English.

Our study begins by contextualizing the issue at hand, including a general overview of ASR technology, a characterization of Google ASR, and a reflection on the demands imposed on ASR by different pedagogical approaches. We will then describe our Research Goals, and our Research Methods in terms of Procedures, Metrics, Participants, and Materials. The remaining sections will follow the traditional structure of Results, Discussion, and Conclusions.

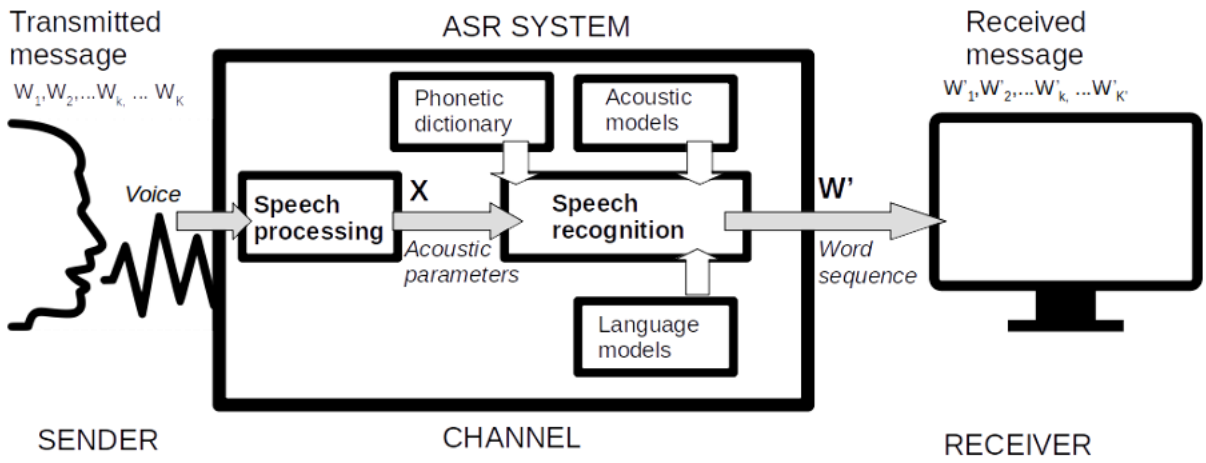
Automatic Speech Recognition

ASR Fundamentals

ASR can be described within the frame of verbal communication (Huang et al., 2001). As Figure 1 shows, senders produce messages consisting of word sequences: W = w1, w2, w3, …, wk. The system captures the acoustic signal through a microphone and generates a sequence X of acoustic parameters of the voice. A recognition module translates X into a sequence W' of words conveying the intended message (Huang et al., 2001).

Figure 1

ASR Processes



The derivation of W' from X relies on language and acoustic modelling. Acoustic models take bundles of acoustic features as input, and select between linguistic units made available by the language model. Deriving W' from X constitutes a matter of probability whereby the possibility of W' being equal to W is maximized. The language model contains a mathematical representation of the messages that could possibly be produced in a particular language together with a notion of their probability (frequency) based on written corpora. The probability of a sequence W, consisting of a collection of words (w1, w2, w3, etc.), considers both the frequency of each word within the corpora and the frequency of its collocation (n-gram) with the other probable words in the sequence (Bellegarda, 2004).

The acoustic model is based on the representation of X/W validated correspondences derived from a repertoire of actual recordings where W and X are known a priori. Phonological inventories assist here by indicating the expected phonological structure (as a chain of phonemes) of each word. Although acoustic models are mainly based on the notion of *phoneme as unit*, most models can handle a variety of specific acoustic features (e.g., the aspiration of plosives), bundles of features constituting the variant of a phoneme, phoneme clusters, whole words, or even whole phrases. Therefore, the acoustic models can dynamically cross and mix different combination levels in constituting their operative units (Deng et al., 2013).

The derivation of W' from X results from the probability that the acoustic features considered by the

acoustic model correspond to a particular message, Pr (X|W) and from the frequency of that particular message in the corpus, Pr (W). The interaction of both probabilities is expressed as a maximized product: W' = argmax W {Pr (X|W) *Pr (W)}; where W' corresponds to the word sequence generated by ASR. Both the probability of the message W and the probability of X to correspond with W' are contemplated in the decision rule (Huang et al., 2001).

Google ASR Specifications

Google ASR is a commercial off-the-shelf, general-purpose service available worldwide for more than 120 languages (Google Cloud, n.d.). It combines state-of-the-art technology with cloud-based computing. Gathering data from millions of individuals using the applications of the company, Google's machine-learning algorithms have obtained remarkable results (Këpuska & Bohouta, 2017; Meeker, 2017). McCrocklin and Edalatishams (2020) presented Google ASR as today's paradigm in the field, which makes it an excellent choice for our study.

The Google ASR system is user friendly. Users submit audio recordings through an online service and receive the system's output as a list of real-time text hypotheses ordered by their confidence rates (n-best list). Response time may depend on the internet connection (never affecting recognition at the server's end). Users can select the ASR language, enable a filter for profanities, and customize the number of hypotheses returned by the system. However, this black-boxed system does not allow for task customization, meaning external developers cannot provide their own custom language or acoustic models. With no incremental output, the service returns only the final list of hypotheses. Results for the same inputs vary over time since Google regularly updates its models, adapting to the audio data supplied by users. We observed accuracy improvements when re-submitting the same audio files after a period of six months.

Like most current ASR systems, Google ASR can be characterized in terms of four basic variables:

- *Speaker dependency*: Speaker dependent systems specialize in recognizing specific speakers. Speaker independent systems are designed for the recognition of speech produced by any speaker in the target language. The acoustic models of the latter are usually trained on extensive speech samples. The systems included in popular operating systems (e.g., Cortana) are speaker independent.
- *Message complexity*: Some ASR systems are designed for recognizing a limited number of inputs (e.g., digits); some are designed for the recognition of word series (e.g., personal names). Systems for continuous recognition are designed for the identification of specific predefined sentences (e.g., commands). Finally, there are the more complex systems for the recognition of spontaneous speech, which can handle false starts, hesitations, grammatical inconsistencies, and so on. Google ASR can work in any of the modes described above.
- *Lexical capacity*: Some systems work with small lexicons (e.g., numbers from 0 to 9). For the recognition of commands, inventories usually contain a few hundred items. For more demanding tasks, like dialog systems, thousands of lexical items are used, such is the case for Google ASR
- *Audio quality*: ASR systems are designed for specific bandwidth and sound capturing devices, from telephone lines to high quality microphones. The amount of speech data processed daily by Google services warrants its robustness and compensates for limitations of audio quality.

ASR efficiency can be measured in terms of their word error rate (WER; Huang et al., 2001). This magnitude has a quantitative and a qualitative dimension. Quantitatively, WER considers discrepancies between the number of words produced by the speaker and the number of words recognized by the system. This dimension contemplates two error types: word-insertion that occurs when the recognized sequence includes more items than the spoken sequence, and word-deletion that takes place when the recognized sequence includes less items. Qualitatively, WER considers issues or word substitution, that

is, cases where the system recognizes a word other the one spoken. The efficiency of some ASR systems is comparable to that of human transcription in terms of their WER. Based on this rate, Google developers ascertain that their system has achieved a word accuracy rate of 95% for English language, placing it at the threshold of human accuracy (Meeker, 2017).

Pedagogical Uses of ASR Systems

In using ASR for the assessment of pronunciation, we placed it at the end of a chain of decisions. Different teaching approaches can target accent-reduction in relation to suprasegmental (prosody, intonation, etc.) or segmental competence (Munro & Derwing, 2015; Thomson & Derwing, 2015). While suprasegmental competence is typically tested against the production of long words, phrases, sentences, or larger portions of speech, segmental assessment targets sounds (i.e., vowels), syllables, pseudo-words, or words in isolation.

As mentioned above, we explored the use of ASR for APA in the context of segment-oriented pronunciation teaching based on minimal pairs assuming an accent-reduction perspective. This form of training and assessment imposes demands which are passed over to ASR. However, there is a kinship between this form of training and the function of ASR. The minimal-pair technique is used to raise phonological awareness by exposing students to word pairs that contain one contrasting segment (e.g., shoe/chew, did/deed, other/udder; Celce-Murcia & Goodwin, 2014), but the challenge in producing any of the components of a minimal pair is that the other component is not recognized instead.

In our study, ASR-based APA systems were those that use words as their unit of recognition and assessment. This condition is shared by all general-purpose ASR systems. The assumption here was that the quality of the pronunciation of an intended target (i.e., the word the student tries to pronounce) depends on whether the system recognizes it over the other possibilities or not. Alternative APA systems rely on the phonemes. These speech-recognition modules perform a phonetic segmentation of the input. The process involves identifying phoneme boundaries and performing phoneme parametrization based on spectral data (Li et al., 2017). The quality of pronunciation is then calculated as a function of the distance between the acoustic parameters of the produced phonemes and those of the expected phonemes. When relying on commercial general-purpose ASR systems such as Google ASR, Siri or Alexa, the developer of APA tools can only access word-unit outputs. This necessarily limits their use for APA purposes as we will show in the present study.

In most uses of the minimal-pair technique, target words are selected according to certain criteria, the first of which is that the words exist in the language (Tejedor-García, 2020). This demand, shared by Google ASR, imposes specific limitations in dealing with specific phonemes. For example, we can find many words in English that contrast in vowels /i:, ɪ/ (beat/bit, wheat/wit, seen/sin, etc.), but pairs contrasting in /ʃ/-/ʒ/ or /u:/-/ʊ/ are scarce. Minimal pairs typically include frequent and concrete words (Avery & Ehrlich, 1995), but in order to include rarer phonemes, minimal-pair-based training often needs to resort to infrequent and exotic words. We departed from the pedagogical constraint of only using frequent and concrete words, prioritizing phonological awareness and competence as independent and legitimate goals (Cámara-Arenas, 2013). This maximized the range of available contrasting pairs and presented an extra challenge to Google ASR (Mirzaei et al., 2016), which is based on repertoires of real pronunciations produced by real users in real communicative events..

Research Goals

As our concern was testing the usability of Google ASR for automatically assessing the pronunciation of ESL learners (standard British accent), we disregarded the variety of interfacing systems that might be used for its integration within an APA structure. The specific demands made on ASR in the current study involved the recognition of frequent and infrequent (but always existing) short English words. Our research questions were as follows:

RQ1. Does Google ASR satisfy the intrinsic condition (recognition of native pronunciations) and guarantee operationality within an accent-reduction APA system?

RQ2. Does Google ASR satisfy the derived condition (misrecognition/nonrecognition of transferred pronunciations) and guarantee operationality within an accent-reduction APA system?

RQ3. Are the levels of (dis)satisfaction found in RQs 1 and 2 interrelated?

RQ4. Can Google ASR be reliably used as the core component of an accent-reduction APA system for English as a second language?

Methodology

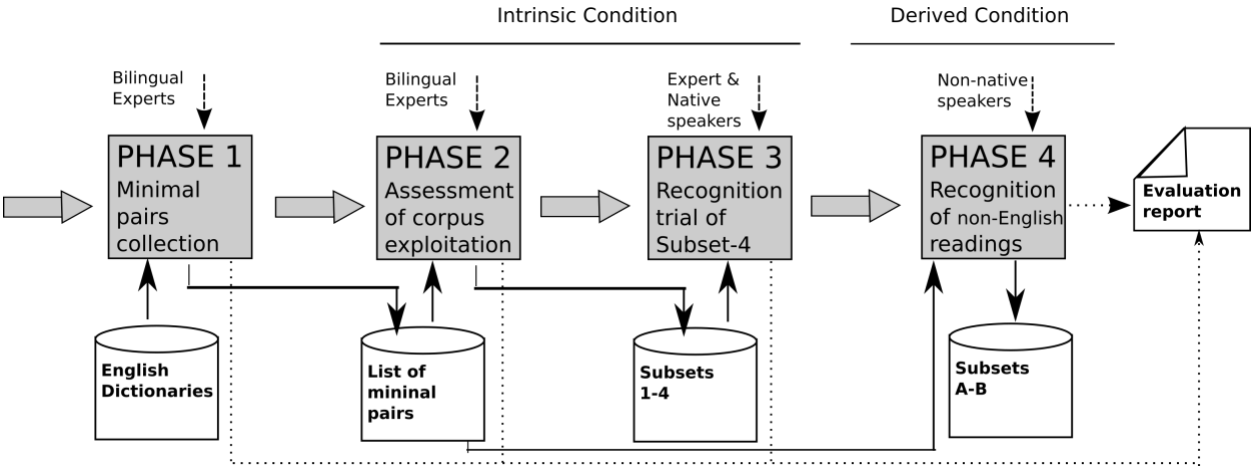
Our study involved pronunciation experts, native speakers (British English) and a miscellaneous group of ESL students with varied levels. The study was articulated around four inter-dependent phases. The second and third phases addressed the intrinsic condition, while the last stage explored the derived condition.

Experimental Phases

Figure 2 describes the four phases of our study. Phase 1 included the compilation of a corpus of minimal pairs. Phase 2 involved the recognition trial of the corpus compiled in phase 1 by two English pronunciation experts. This led to the characterization of the whole corpus that returned a collection of homogeneous subsets 1 to 4, the last of which (subset 4) incorporated the most misrecognized and/or unrecognized words. Phase 3 consisted of a trial of subset 4 by native speakers and experts. Phase 4 explored the recognition of the whole corpus in transferred readings which were carried out by the expert and non-expert participants of the study.

Figure 2

Phases of the Experimental Procedure



Several constrictions were imposed on phase 1. In order to keep the extension of the corpus within manageable limits, experts searched for words that contained stress /æ, ʌ, ɒ, e, ɪ, iː/. The corpus included mostly monosyllabic words which were susceptible to be used in at least one minimal-pair contrast involving another vowel. All the words came from the *Oxford English-Spanish Dictionary*, *Longman's Pronunciation Dictionary*, and *Merriam-Webster's 11th Collegiate Dictionary*. Experts were asked to target the following contrasts:

/æ/ vs /e/ (e.g., pan/pen)

/æ/ vs /ʌ/ (e.g., fan/fun)

/æ/ vs /ɒ/ (e.g., sack/sock)

/ʌ/ vs /ɒ/ (e.g., done/dawn)

/e/ vs /ɪ/ (e.g., bet/bit)

/i:/ vs /ɪ/ (e.g., seat/sit)

In phase 2, the two experts recorded/uploaded a reading of the entire set to an internet environment (see Materials) which collected the audio files to be presented to Google ASR. The system returns a size-customizable n-best list of probable targets, ordered by their confidence rates. We set the number of positions contained within the n-best list to five. In using Google ASR for APA, prospective developers will decide upon a *level of tolerance* of the system based on the n-best list and determine whether a word might be considered well-pronounced only if it is registered in the first position of the list or when registered in any of the customized positions. Intermediate degrees would also be possible.

The results of phase 2 were expressed in the form of four subsets. Each subset contained words that met the same tolerance criteria. For instance, subset 1 contained the words read by both experts that were recognized by the ASR in position 1; subset 2 included those recognized in positions 1 or 2; and subset 3, those recognized in positions 1 through 3. One of the outcomes of phase 2 was the isolation of subset 4, which contained the most unrecognized or misrecognized words, to be tried further in the subsequent phase. One usability issue was the number (sufficient, limited, insufficient) of the available items in every subset.

In phase 3, the same procedure was followed by the participating experts and native speakers. The 22 words of subset 4 were read aloud three times (66 reading events per participant) by each of the 13 participants (totaling 858 events) directly to Google ASR (through a smartphone application), making sure that ambiance conditions were optimal. After each reading, the position of the target word within the n-best list generated by the ASR was registered. Phase 3 sought to confirm the homogeneity of subset 4, while validating the experts' readings.

In phase 4, a miscellaneous group of native Spanish speakers read the 288 words with a Spanish-transferred pronunciation. The instructions provided for these participants in phase 4 included: "read it with a marked Spanish accent," "read the list as if they were Spanish words," and "Make sure that the vowels and the consonants that you use are all Spanish". These instructions were to be subjectively interpreted by the different speakers, but they guaranteed that the pronunciations presented to ASR in phase 4 were deviating and non-standard. Similar to phase 2, phase 4 results were characterized through the isolation of subsets A through D, where the tolerance level of each subset increases from A, containing the words that returned in position 1, to D, containing all the words returned in any position.

Metrics

The ASR system was customized to return an n-best list with five words. The most probable candidate appeared in position 1 of this list. From positions 1 to 5, the probability generated by the ASR decreased. Table 1 illustrates the results of a reading where the words 'porch' and 'soot' (not in the corpus) were recognized in positions 1 and 3, respectively.

Table 1

Computation Sample of P in Google ASR Output to an Expert Reading of 'Porch' and 'Soot'

Input	**n-best list results**					
Intended Word	**Position 1**	**Position 2**	**Position 3**	**Position 4**	**Position 5**	**P**
'porch'	**porch**	porge	Porche	pooch	Poarch	1
'soot'	Suit	foot	**soot**	search	Suet	3

The following notions were designed in order to carry out the analysis and characterization of the results:

- Predictability/Unpredictability (P, –P): The notion of Predictability refers to the n-best position where the ASR registers a given input. Predictability values are the integers 1 to 6 (Pϵ {1,2,3,4,5,6}); value 1 stands for maximal P, and value 6 represents a non-predicted target. In the example above, P('porch') = 1 and P('soot') = 3. Predictability can be expressed in terms of levels:
    - Level 1, total predictability, P = 1
    - Level 2, high predictability, P = 2
    - Level 3, moderate predictability, P = 3
    - Level 4, moderate unpredictability, P = 4
    - Level 5, high unpredictability, P = 5
    - Level 6, total unpredictability, P = 6
- Predictability Index (PI): This is an index resulting from the mathematical treatment of position tags (Predictability). For each reading of the same word, we computed P-values as 1 to 6, as described above, and calculated the average to define the Predictability Index, which also ranged between 1 (maximal PI) and 6 (minimal PI). Table 2 exemplifies PI with three reading-trials of the word 'soot' by a single speaker. We also characterized PI levels in terms of total, high, moderate, and so forth.

Table 2

Computation Sample of PI in Google ASR Output to an Expert Reading of 'Soot'

Input: 'soot'	**n-best list results**					
Readings	Position 1	Position 2	Position 3	Position 4	Position 5	P
R1('soot')	suit	foot	**soot** (3)	Search	suet	3
R2('soot')	suet	suit	Stewart	Sue it	Sweat	6
R3('soot')	suit	**soot** (2)	shoot	Sute	shipt	2
Predictability Index ('soot') = (3+6+2)/3 = 3.6						

- Confusion Rate (CR): CR consisted of the quotient of the number of times a targeted word was not returned in any position within the n-best list, and the total number of times that word has been read. The values of CR ranged between 0 and 1 (i.e., Maximal CR). The confusion rate of the example above was calculated as CR('soot') = 1/3 = 0.3 (30%) (i.e., a moderately low CR).

- Strict Confusion Rate (SCR): SCR included the quotient of the number of times a targeted word was not recognized in position 1 within the n-best list, and the total number of times that word was

read. The values of SCR also ranged between 0 and 1 (i.e., Maximal SCR). The strict confusion rate of the example above is SCR ('soot') = 3/3 = 1 (100%).

- Mean and Mode: the mean and mode of the n-best position was calculated for the recognized words. In the example above, Mean ('soot') = (3+2)/2 = 2.5, and Mode ('soot') = 2.5.

These metrics were computed both by word and by speaker. Satisfactory performance of the ASR in terms of the intrinsic condition implied PI-values near 1, and CR/SCR-values near 0 regardless of (non)standard pronunciation. Satisfactory performance of the ASR in terms of the derived condition imply PI-values near 6, and CR/SCR-values near 1 for non-standard pronunciation. The Chi-square test was used to compare the performance of the ASR system when used with bilingual experts and native speakers.

Participants

This study involved three groups of participants. In phase 1, the corpus was built upon request by two Spanish-English bilingual speakers with expertise in English phonetics. In phase 2, these two speakers tested the entire corpus against Google ASR, generating the data that guided our identification and selection of subsets 1 to 4.

In phase 3, 11 native speakers of English and the two experts who had participated in phases 1 and 2 tested subset 4 under the same experimental conditions. The native speakers consisted of teachers and students from the University of Leicester. Most of them used Received Pronunciation (a standard variety of British pronunciation), and some incorporated occasional regional features (Cheshire, Manchester, Lincolnshire, London). Four participants were more than 30 years old, while the rest were between 18 and 25. Four participants were female and nine were male.

Phase 4 also involved the two experts, and nine Spanish ESL students with different levels of competence (from beginner to proficient). All of them were asked to produce Spanish-transferred pronunciations of English.

Materials

In phase 1, experts used the dictionaries mentioned above and elaborated their lists of contrasts on Excel 2010. For phases 2 and 4, a web application was developed for Google Chrome 70-80, Mozilla Firefox 60-70, and Safari 13 that allowed the participants to record/upload their readings of the entire corpus. The application's interface was straightforward. A word in bold was displayed together with a word counter. Below these two elements, visible "start" and "stop" buttons were placed for initiating/resuming recording. In the upper right part of the screen, simple recording instructions remained visible throughout the process. Participants were asked to record in a quiet environment, using compatible web browsers and their own microphones and speakers/headphones. Once each word was recorded, the speaker could listen to their recording and re-record it, or they could continue to the next word. All recordings in phases 2 and 4 were presented to Google's online ASR service, and the results were obtained and registered in terms of P, from which PI, CR, and SCR were later calculated.

To reduce participant involvement (and inconvenience), for phase 3, an interface Android-compatible application (versions 4-10) previously developed by the research team was used for testing Google ASR against 22 words which were isolated in phase 2. The interface elements of this application were very similar to those of the web application used in phases 2 and 4. The screen prompted the voice input, and a subsequent text frame returned the recognition event in terms of a 5-step n-best list. Participants interacted with a mobile phone in which the application was installed and read the words one-by-one from a printed out list. After each reading event, a member of our team manually transcribed the n-best list results into their notes. Participants were seated on a comfortable chair, in an empty and quiet room.

Results

Phase 1. Corpus Compilation

In phase 1, the experts generated a list of 288 words susceptible to being used in contrastive minimal pairs. Only six vowels were covered at this stage. The initial corpus in Phase 1 presented the following distributions:

- Vowel /ɒ/: 43 items (15%), involved in 2 contrasts (/æ/, /ʌ/)
- Vowel /æ/: 77 items (27%), involved in 3 contrasts (/ʌ/, /ɒ/, /e/)
- Vowel /ʌ/: 47 items (16%), involved in 2 contrasts (/æ/, /ɒ/)
- Vowel /e/: 46 items (16%), involved in 2 contrasts (/æ/, /ɪ/)
- Vowel /i:/: 27 items (9%), involved in 1 contrast (/ɪ/)
- Vowel /ɪ/: 48 items (17%), involved in 2 contrasts (/e/, /i:/)

With an average presence of 48 items per vowel, the absolute number of items per vowel ranged between 27 (vowel /i:/) and 77 (vowel /æ/). This variability had to do with the number of potential contrasts explored by the experts who selected the primary corpus. Vowel /æ/ items were searched to guarantee the contrasts with /ʌ/, /e/, and /ɒ/ (e.g., sack, sec, suck, sock), while the exploration of vowel /i:/ was limited to its contrast with /ɪ/ (e.g., seat, sit). The uniformity of the set became evident when we considered the number of words that were involved in each contrast which were between 21 and 27 words (25.6 average). The primary corpus was therefore balanced in terms of the amount of testing opportunities it could provide for each of the contrasts, should all the items be integrated and used by an APA system.

Determining the Exploitation Potential of the Corpus (Phase 2, Intrinsic Conditions)

Prospective developers of a Google ASR-based APA will have to work only with those words that the ASR can recognize; successful recognition, on the other hand, must be predefined in terms of the n-best list (i.e., words recognized in position 1, position 2, and/or position 3, etc.). Table 3 presents the portion of words (out of 288) for each expert that attained a PI between 1 and 3. The PI values in Table 3 helped us then determine the exploitation potential of the corpus in three predictability conditions: total (Subset 1), high (Subset 2), and moderate (Subset 3) predictability. The two experts did not always get the same PI for every item. In extreme cases, words that attained total predictability in one of the experts' readings, turned out to be unpredictable in the other's. Homogeneous subsets were obtained by considering the intersections of both experts' readings, that is, those words that reach similar PI in the reading of both experts (e.g., 43 specific items, out of 288, reached PI = 1 in all three readings by both experts). By virtue of their shared total-to-moderate PI, these progressively embedding subsets contained materials that could be safely used in an APA protocol (based on the two experts' readings).

Table 3

Number of Words and Percentages in the Different Predictability Subsets (Determined by the Value of PI)

	PI = 1	**PI ≤ 2**	**PI ≤ 3**
Expert 1	98 (34.0%)	179 (62.2%)	228 (79.2%)
Expert 2	69 (24.0%)	147 (51.0%)	196 (68.1%)
	Subset 1	Subset 2	Subset 3
Both	43 (14.9%)	123 (42.7%)	176 (61.1%)
P-Words	14	79	142
Pairs	7	59	116

Out of 14 words (*P-Words*) that were able to be paired, subset 1 allowed for the construction of only seven minimal pairs to work with: back/beck, cat/cut, check/cheek, dean/done, drag/drug, rec/rock, trick/truck. However, these contained instances of all the targeted vowels. With four words, vowel /ʌ/ was the most represented, followed by /æ/ and /e/ with three words each, vowel /i:/ with two, and vowels /ɒ/ and /ɪ/ with one word each.

Subset 2 contained 79 items that could be combined into minimal pairs and trios, with a total of 59 minimal pairs constructed from this subset. With 21 words, vowel /æ/ was the most represented in this subset. This was followed by vowel /e/ with 16 instances; vowel /ɪ/ with 14; vowel /i:/ with 11; vowel /ʌ/ with 10; and vowel /ɒ/ with 7 instances.

Subset 3 contained 142 items that could be combined into pairs and trios. A total of 116 pairs were constructed from the subset. In this subset, vowel /æ/ was the most represented (34 instances); followed by /e/ and /ɪ/ (29 instances each); followed, in decreasing order, by /i:/ (19); /ʌ/ (18); and /ɒ/ (13).

Based on the results in phase 2, a fourth subset (subset 4) was also isolated which contained 22 words that had obtained PI values above 3 (PI > 3) in the readings of both experts (i.e., words at moderate total unpredictability levels, or problematic words). These are the words which were tested in phase 3 of the present study: bodge; blot; blub; clod; clomp; cluck; din; glom; hem; hut; kip; lass; much; muck; peck; Somme; stump; sung; than; tromp; tup; wan

Testing and Verification of Problematic Words (Phase 3, Intrinsic Condition)

Table 4 registers the average confusion rate (CR) and strict confusion rate (SCR) attained by each participant against subset 4. Only 12.4% of the total 858 reading events in phase 3 registered in the first position, and in 33.8% of the readings, the target word was not identified (NF). Expert 1 attained the lowest CR and SCR values (CR = 16.7% and SCR = 65.2%). The difference between the values obtained by Expert 1 and those of L05, L08, L11, and Expert 2 were statistically significant ($p < 0.05$, Chi-square test with distribution of the positions). Native speaker L08 attained the highest CR (53%) and SCR (97%), with statistically significant differences with respect to L09 and Expert 1 ($p < 0.05$, Chi-square test).

Table 5 presents the overall confusion rate of each of the words in subset 4 when read by the participating native speakers. There were clear differences between the words. The word 'glom' was never recognized by the system in any positions (CR =100%); at the other end, the word 'much' was always registered in the n-best list (CR = 0%), but never in the first position (SCR = 100%). Many words in subset 4 found a place in the n-best list in most readings (CR < 10%), but with SCR values above 30% in all cases, none was invariably found in first position in native readings.

Table 4

Participants' Confusion Rates

	PI							
SPK	**1**	**2**	**3**	**4**	**5**	**NF**	**CR**	**SCR**
L01	8	13	9	6	2	28	42.4%	87.9%
L02	7	14	5	8	6	26	39.4%	89.4%
L03	9	14	12	7	2	22	33.3%	86.4%
L04	10	18	9	4	6	19	28.8%	84.8%
L05	5	19	5	8	2	27	40.9%	92.4%
L06	9	15	13	3	4	22	33.3%	86.4%
L07	10	14	11	5	1	25	37.9%	84.8%
L08	2	8	10	7	4	35	53.0%	97.0%
L09	9	21	7	8	5	16	24.2%	86.4%
L10	7	16	14	8	3	18	27.3%	89.4%
L11	3	16	11	10	4	22	33.3%	95.5%
Expert 1	23	14	8	7	3	11	16.7%	65.2%
Expert 2	4	17	13	9	4	19	28.8%	93.9%
Total	**106 (12.4%)**	**199 (23.2%)**	**127 (14.8%)**	**90 (10.5%)**	**46 (5.4%)**	**290 (33.8%)**		

Note. SPK is the speaker identifier where L01 to L11 are native speakers of English and Expert 1 and Expert 2 are non-native expert speakers; columns 1 to 5 register the number of reading events where the target word appears in the corresponding position of the ASR-output list; NF counts the events where the ASR did not recognize a target word (PI = 6); CR stands for confusion rate and SCR for the strict confusion rate as defined in Metrics section. In the lower level, percentages are calculated against the total number of reading events performed (858).

Table 5

Average Confusion Rates of Subset 4

Word	CR	SCR	Mean	Mode
Glom	100.0%	100.0%	*	*
Sung	93.9%	100.0%	2.5	3
Tromp	81.8%	93.9%	1.7	2
Blub	75.8%	100.0%	4.9	5
Tup	69.7%	100.0%	3.6	3
Than	60.6%	97.0%	2.4	2
Cluck	57.6%	93.9%	3.1	2
Hem	42.4%	90.9%	2.7	2
Din	36.4%	100.0%	2.9	2
Hut	30.3%	93.9%	3.0	3
Blot	27.3%	75.8%	2.3	2
Bodge	24.2%	48.5%	1.8	1
Lass	18.2%	100.0%	2.8	2
Wan	15.2%	100.0%	3.9	4
Somme	15.2%	93.9%	2.9	3
Clomp	15.2%	100.0%	4.0	4
Muck	12.1%	51.5%	1.9	1
Peck	6.1%	81.8%	2.6	3
Kip	3.0%	93.9%	2.2	2
Clod	3.0%	45.5%	1.9	1
Stump	0.0%	100.0%	2.1	2
Much	0.0%	100.0%	2.0	2

Note. CR is the Confusion Rate, SCR is the Strict Confusion Rate, and Mean and Mode are the statistics referring to the position of the word in the n-best list returned by Google's ASR.

Tolerance of Foreign Pronunciation (Phase 4, Testing of the Derived Condition)

In phase 4, Google ASR was presented with a total of 3469 transferred readings of 288 words. Adopting a similar approach to that we deployed in phase 3, in the last phase of our study we could isolate four more subsets: subset A, subset B, subset C, subset D. Table 6 registers the recognized data from the transferred readings. We computed the number of the readings that were recognized (R-Readings) in four positions: position 1, positions 1 to 2, positions 1 to 3, and positions 1 to 5. Subset D reported on the recognition of transferred pronunciations in any positions within a 5-step n-best list.

Up to 18.2% of the readings obtained recognition of the target word in position 1. More than half of the words in the initial set (53.5%) achieved this level of recognition when consciously pronounced with transferred pronunciation. Table 6 shows the extent to which the number of words attaining recognition across larger ranges tends to increase; finally, up to 243 words (84.4% of the initial set) attained some sort of recognition by the ASR in any of the previously isolated positions (1 to 5) even though they were intentionally pronounced with non-English pronunciations.

Table 6

Transferred Readings

Column Header	PI	R-Readings	# Words
Subset A	PI = 1	576 (18.2%)	154 (53.5%)
Subset B	PI ≤ 2	809 (25.5%)	199 (69.1%)
Subset C	PI ≤ 3	948 (29.9%)	222 (77.1%)
Subset D	PI ≤ 5	1125 (35.5%)	243 (84.4%)

Note. R-Readings (Recognized-Readings) registers the amount of reading events that placed a target word on the n-best list. # Words registers the number of words in every subset. R-Readings percentages are based on the total number of reading events (3168 = 288 words * 11 speakers) and # Words percentages are based on the total number of words in the corpus (288).

Discussion

The initial premise stating that a word recognized by the system is a word that has been properly pronounced was proven wrong as the results show that many transferred pronunciations were properly recognized. However, we were able to verify the general satisfaction of the intrinsic condition in the case of Google ASR. The system recognized 61.1% of the properly pronounced English words presented to it by the two experts in phase 2 of the study, reflecting a predictability that ranged from total to moderate (Subset 3, Table 3). A relatively small number of English words could not be recognized in the first position of the n-best list (SCR near 100%). This might be because they were either among the words that are infrequent in English (e.g., 'glom'), or the ones that rarely appear in isolation (e.g., 'than'). The satisfaction of the initial condition by Google ASR may be related to the dissatisfaction of the derived condition. Phase 4 results showed that an alarming total of 84% of the readings attained some degree of recognition despite being pronounced with obvious non-English features (Table 6).

As subsets 1 to 3 became more and more inclusive, we saw that the satisfaction of the intrinsic condition increased (a total of 14.9% of subset 1 words in comparison to 61.1% of subset 3 words of the initial corpus). Even more recognition was attained for phase 4 subsets A to C (subset A included 54.1% of the initial corpus, reaching to 77% in subset C). Interestingly, it was noticed that while the increasing recognition of the properly pronounced words increasingly satisfied the intrinsic condition, the extensive recognition of the transferred pronunciations expressed an equally extensive dissatisfaction of the derived condition. In other words, recognition defines success for an ASR, but it may also depict failure for an accent-reduction APA system. Our results show that the dissatisfaction of the derived condition amply eclipsed the satisfaction attained by Google ASR at the intrinsic condition.

Furthermore, our study uncovered a lack of correlation between native pronunciation and the metrics used in this study (PI, CR, SCR). Experts and native informants often got different measures for the same word even when all the readings inputted in phases 2 and 3 met pronunciation standards. Consequently, it seems that the n-best list generated by Google ASR should not be used as a device for granting discriminatory pronunciation scores. It could be seen that perfect pronunciations did not always register in position 1, while increasingly accented pronunciations were not guaranteed to register in positions 2 to 5. It was not always true, either, that unrecognized pronunciations were necessarily non-native. Finally, it was not true that non-native pronunciations went unrecognized or even got lower positions in the n-best list. Therefore, it can be safely claimed that any automatic scoring procedure based on the n-best list will fail to generate a truthful and reliable assessment.

While it might be expected that future versions of Google's ASR may increasingly satisfy the intrinsic condition, we should not expect its efficacy within an accent-reduction APA system to improve with it.

Rather on the contrary, the intrinsic condition is incompatible with this extraneous and artificially imposed derived condition. This can be explained by the language module integrated in the ASR. In the average ASR, an acoustic model generates a number of options which are later filtered through the sieve of a linguistic model. This linguistic model incorporates criteria other than the purely acoustic (e.g., collocational probabilities, word/chunk frequency, etc.). As a result, when we present the word 'din' to the ASR system, the possibility that anybody might be really uttering a one-word sentence with content such as 'din' in order to communicate a full message is so low that the ASR system pushes it down the n-best list, giving priority to a word such as 'then'. This is due to a compromise between the features detected in the input—alveolar nasal at the end, a front vowel which is not totally close—and a calculation of probability which is based on context, collocational likelihood, and frequency. The result, in terms of assessment, might be the following: we ask our student to read 'then,' but the student, due to an interference of their L1, produces an expression with the features of 'din,' and the ASR recognizes 'then' perfectly, not because of the acoustic properties of the sound, but by virtue of probability-based processing. It is precisely this probability-based process that should be absent from an accent-reduction APA system if the assessment of the pronunciation of isolated short words is the goal.

It might be argued that what an accent-reduction APA system needs is a specific-purpose ASR. Such a system might carry out assessment tasks more satisfactorily (i.e., by limiting target words beforehand, and predefining the set of possible input interpretations). Our opinion, based on the results of the present study, is that efficient accent-reduction APA must abandon the orbit of recognition altogether. In this way, it would mirror traditional human assessment, where a human agent asks the student to pronounce a particular, and therefore expected, word. Recognition plays no role in this scenario. Where an ASR system is trying to determine what the user might be saying, an APA system should be delivered from such tension. The system should be expecting the user to pronounce a particular word; the user would produce it, and then, the system should simply measure the distance between the inputted expression and a stored model. Finally, a scoring module should interpret this distance and generate a score dependent on the acoustic properties of the input.

The implementation of such a system would involve a deep and extensive understanding of acoustic phonetics, since the functionality of a given expression is not granted by the totality of its acoustic features (statistically processed), but by a few indicators in each case (e.g., the formant values of a vowel, the formant transitions before and after plosive, the voice onset times, the frequency-range of frictions, the duration and intensity of fricative noise, etc.). Whatever the approach, it seems clear to us that an accent-reduction APA system for short words in isolation should reduce combinatory logic in favor of acute acoustic measurement. Combinatory logic, such as the logic that emerges from phonotactic considerations, might not have to be totally eradicated; while not all pronunciation oddities are logical, in dealing with specific groups of learners (e.g., L1-Spanish learners of English), the recurrence of specific and expected pronunciation difficulties might play a role in automatic assessment.

There were a number of limitations to this study. First, the present study relied on a variety of devices and tools (web applications, smartphones, computers, etc.) that were not always working perfectly. Therefore, in phase 4, each participant used their own personal equipment. Although the process mirrored a realistic setting, where students would access CAPT tools through their own personal devices, variable control in this sense would have been stricter under different circumstances. Given the robustness of Google ASR and having verified all the audio materials which were collected in phase 4, we are confident that this limitation did not invalidate our conclusions.

Interestingly, in the phase 3 results, Google ASR recognized Expert 1, who was a non-native speaker of English, better than all the other participants. This might be because our expert was acclimated to the Google ASR system. However, this hypothesis could not be explored properly under the present design and should be left for future research.

Conclusions

The current study suggests that the very nature of an ASR system such as Google ASR limits its subsidiary use within an accent-reduction APA system and prospectively prevents it. In reaching ever higher levels of recognition (intrinsic condition), it also reduces a derived ability to filter out non-standard pronunciations.

Google ASR satisfied the intrinsic condition (RQ1). However, its operationality within an accent-reduction APA system can be questioned because its analysis of probability (n-best list) could not reliably support a scoring system. Furthermore, the system did not satisfy the derived condition imposed upon it, allowing the recognition by default of non-standard and even exotic pronunciations, a recognition that is based more on linguistic co-textual probability and frequency than on the objective acoustic properties of the input (RQ2). It was revealed that the heuristic mechanisms which lead to a satisfaction of the intrinsic condition are those which are responsible for the dissatisfaction of the derived condition (RQ3). Our study showed that Google ASR cannot be used to generate reliable accent-reduction APA (RQ4) as it cannot evaluate the input according to the acoustic reality presented to it. It is recommended that future APA systems for short words in isolation reduce contextual, probabilistic, and by-default recognitions in favor of a precise acoustic measurement and an objective consideration of the acoustic distances between stored models and recorded inputs.

References

Adda-Decker, M., & Lamel, L. (2000). The use of lexica in automatic speech recognition. In F. Van Eynde and D. Gibbon (Eds.), *Lexicon development for speech and language processing* (pp. 235–266) Springer, Dordrecht. https://doi.org/10.1007/978-94-010-9458-0_8

Avery, P., & Ehrlich, S. (1995). *Teaching American English pronunciation*. Oxford University Press.

Becker, K., & Nguyen, P. (2017). Technology-enhanced language learning for specialized domains: Practical applications and mobility. *Language Learning & Technology, 21*(3), 67–71. https://doi.org/10125/44633

Bellegarda, J. R. (2004). Statistical language model adaptation: Review and perspectives. *Speech Communication, 42*(1), 93–108. https://doi.org/10.1016/j.specom.2003.08.002

Cámara-Arenas, E. (2013). *Curso de pronunciación de la lengua inglesa para hispano-hablantes: A native cardinality method*. Universidad de Valladolid.

Celce-Murcia, M., & Goodwin, J. M. (2014). *Teaching pronunciation: A reference for teachers of English to speakers of other languages*. Cambridge University Press.

Chen, N. F., & Li, H. (2016). Computer-assisted pronunciation training: From pronunciation scoring towards spoken language learning. In *2016 Asia-Pacific Signal and Information Processing Association annual summit and conference* (pp. 1–7). IEEE. https://doi.org/10.1109/APSIPA.2016.7820782

Cheng, J. (2018). Real-time scoring of an oral reading assessment on mobile devices. In *Proceedings of Interspeech 2018* (pp. 1621–1625). International Speech Communication Association (ISCA). https://doi.org/10.21437/Interspeech.2018-34

Coniam, D. (1999). Voice recognition software accuracy with second language speakers of English. *System*, *27*(1), 49–64. https://doi.org/10.1016/S0346-251X(98)00049-9

Crossley, S., & McNamara, D. (2013). Applications of text analysis tools for spoken response grading. *Language Learning & Technology*, *17*(2), 171–192. https://doi.org/10125/44329

Cucchiarini, C., & Strik, H. (2018). Second language learners' spoken discourse: Practice and corrective feedback through automatic speech recognition. In Information Resources Management Association (Ed.), *Smart technologies: Breakthroughs in research and practice* (pp. 367–389). IGI Global. https://doi.org/10.4018/978-1-5225-2589-9.ch016

Deng, L., Hinton, G., & Kingsbury, B. (2013). New types of deep neural network learning for speech recognition and related applications: An overview. In *2013 IEEE international conference on acoustics, speech and signal processing* (pp. 8599–8603). IEEE. https://doi.org/10.1109/ICASSP.2013.6639344

Derwing, T., Munro, M., & Carbonaro, M. (2000). Does popular speech recognition software work with ESL speech? *TESOL Quarterly*, *34*(3), 592–603. https://doi.org/10.2307/3587748

Google Cloud. (n.d). *Cloud speech-to-text*. Google. https://cloud.google.com/speech-to-text

Hasan, R., Hussein, H., Lazaridis, P., Khwandah, S., Ritter, M., & Eibl, M. (2017). Improvement of speech recognition results by a combination of systems. In *2017 23rd international conference on automation and computing* (pp. 1–4). IEEE. https://doi.org/10.23919/IConAC.2017.8082082

Hermans, F., & Sloep, P. (2018). Teaching English pronunciation beyond intelligibility. *International Journal of Language Studies*, *12*(1), 107–124.

Huang, X., Acero, A., & Hon, H. W. (2001). *Spoken language processing: A guide to theory, algorithm, and system development.* Prentice Hall PTR.

Jenkins, J. (2002). A sociolinguistically based, empirically researched pronunciation syllabus for English as an international language. *Applied Linguistics*, *23*(1), 83–103. https://doi.org/10.1093/applin/23.1.83

Këpuska, V., & Bohouta, G. (2017). Comparing speech recognition systems (Microsoft API, Google API and CMU Sphinx). *International Journal of Engineering Research and Applications*, *7*(3), 20–24. https://doi.org/10.9790/9622-0703022024

Kukulska-Hulme, A. (2012). Mobile-assisted language learning. *Blackwell Publishing Limited.* https://doi.org/10.1002/9781405198431.wbeal0768

Levis, J. (2005). Changing contexts and shifting paradigms in pronunciation teaching. *TESOL Quarterly*, *39*(3), 369–377. https://doi.org/10.2307/3588485

Li, K., Qian, X., & Meng, H. (2017). Mispronunciation detection and diagnosis in L2 English speech using multidistribution deep neural networks. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, *25*(1), 193–207. https://doi.org/10.1109/TASLP.2016.2621675

Liakin, D., Cardoso, W., & Liakina, N. (2015). Learning L2 pronunciation with a mobile speech recognizer: French /y/. *CALICO Journal*, *32*(1), 1–25. https://doi.org/10.1558/cj.v32i1.25962

Luo, B. (2016). Evaluating a computer-assisted pronunciation training (CAPT) technique for efficient classroom instruction. *Computer Assisted Language Learn*ing, *29*(3), 451–476. https://doi.org/10.1080/09588221.2014.963123

McCrocklin, S. M. (2016). Pronunciation learner autonomy: The potential of automatic speech recognition. *System*, *57*, 25–42. https://doi.org/10.1016/j.system.2015.12.013

McCrocklin, S., & Edalatishams, I. (2020). Revisiting popular speech recognition software for ESL speech. *TESOL Quarterly*, *54*(4), 1086–1097. https://doi.org/10.1002/tesq.3006

Meeker, M. (2017). *Internet trends 2017*. Kleiner Perkins. https://www.bondcap.com/report/it17

Mirzaei, M. S., Meshgi, K., & Kawahara, T. (2016). Automatic speech recognition errors as a predictor of L2 listening difficulties. In D. Brunato, F. Dell'Orletta, G. Venturi, T. François, & P. Blache (Eds.), *Proceedings of the workshop on computational linguistics for linguistic complexity* (pp. 192–201). The COLING 2016 Organizing Committee. https://www.aclweb.org/anthology/W16-4122

Munro, M. J., & Derwing, T. M. (2015). A prospectus for pronunciation research in the 21st century: A point of view. *Journal of Second Language Pronunciation*, *1*(1), 11–42. https://doi.org/10.1075/jslp.1.1.01mun

Neri, A., Cucchiarini, C., Strik, H., & Boves, L. (2010). The pedagogy-technology interface in computer-assisted pronunciation training. *Computer Assisted Language Learning*, *15*(5), 441–467. https://doi.org/10.1076/call.15.5.441.13473

O'Brien, M., Derwing, T. M., Cucchiarini, C., & Hardison, D. M. (2018). Directions for the future of technology in pronunciation research and teaching. *Journal of Second Language Pronunciation*, *4*(2), 182–207. https://doi.org/10.1075/jslp.17001.obr

Rahimi, M. (2015). *Handbook of research on individual differences in computer-assisted language learning.* IGI Global.

Seed, G., & Xu, J. (2018). Integrating technology with language assessment: Automated speaking assessment. In E.G. Eugenio (Ed.), *Learning and assessment: Making the connections – Proceedings of the ALTE 6th international conference, 3-5 May 2017* (pp. 286–291). Association of Language Testers in Europe (ALTE).

Tejedor-García, C., Escudero-Mancebo, D., Cámara-Arenas, E., González-Ferreras, C., & Cardeñoso-Payo, V. (2020). Assessing pronunciation improvement in students of English using a controlled computer-assisted pronunciation tool. *IEEE Transactions on Learning Technologies*, *13*(2), 269–282. https://doi.org/10.1109/TLT.2020.2980261

Thomson, R. I. (2012). Improving L2 listeners' perception of English vowels: A computer-mediated approach. *Language Learning*, *62*(4), 1231–1258. https://doi.org/10.1111/j.1467-9922.2012.00724.x

Thomson, R. I., & Derwing, T. M. (2015). The effectiveness of L2 pronunciation instruction: A narrative review. *Applied Linguistics*, *36*(3), 326–344. https://doi.org/10.1093/applin/amu076

Yarra, C., Srinivasan, A., Gottimukkala, S., & Ghosh, P. K. (2019). SPIRE-fluent: A self-learning app for tutoring oral fluency to second language English learners. In *Proceedings of Interspeech 2019* (pp. 968–969). International Speech Communication Association (ISCA).

About the Authors

Enrique Cámara-Arenas teaches English phonetics at the University of Valladolid. His research interests are English orthography and grapho-phonemics, EFL methodology, and the teaching of English pronunciation.

E-mail: enrique.camara@uva.es

Cristian Tejedor-García works as a postdoc at the Centre for Language and Speech Technology (CLST), Radboud University Nijmegen (the Netherlands). His research fields are automatic speech recognition, human-computer interaction, and technology in education.

E-mail: cristian.tejedorgarcia@ru.nl

Cecilia Judith Tomas-Vázquez is a University of Valladolid English Studies graduate, where she also earned a Master's degree in Secondary Education Teaching, specializing in English as a Second Language.

E-mail: ceciliatomasvazquez@gmail.com

David Escudero-Mancebo is an associate professor in the Department of Computer Science of the University of Valladolid. He is an expert in multimedia technologies and has led several projects on the application of speech technology to pronunciation training.

E-mail: descuder@infor.uva.es