languages

MDPI

Review

Automatic Speech Recognition in L2 Learning: A Review Based on PRISMA Methodology

Mireia Farrús [1,2]

[1] Centre de Llenguatge i Computació (CLiC), Universitat de Barcelona, 08007 Barcelona, Spain; mfarrus@ub.edu
[2] Institut de Recerca en Sistemes Complexos (UBICS), Universitat de Barcelona, 08028 Barcelona, Spain

Abstract: The language learning field is not exempt from benefiting from the most recent techniques that have revolutionised the field of speech technologies. L2 learning, especially when it comes to learning some of the most spoken languages in the world, is increasingly including more and more automated methods to assess linguistics aspects and provide feedback to learners, especially on pronunciation issues. On the one hand, only a few of these systems integrate automatic speech recognition as a helping tool for pronunciation assessment. On the other hand, most of the computer-assisted language pronunciation tools focus on the segmental level of the language, providing feedback on specific phonetic pronunciation, and disregarding the suprasegmental features based on intonation, among others. The current review, based on the PRISMA methodology for systematic reviews, overviews the existing tools for L2 learning, classifying them in terms of the assessment level, (grammatical, lexical, phonetic, and prosodic), and trying the explain why so few tools are nowadays dedicated to evaluate the intonational aspect. Moreover, the review also addresses the existing commercial systems, as well as the existing gap between those tools and the research developed in this area. Finally, the manuscript finishes with a discussion of the main findings and foresees future lines of research.

Keywords: automatic speech recognition; ASR; L2; prosody

Citation: Farrús, Mireia. 2023. Automatic Speech Recognition in L2 Learning: A Review Based on PRISMA Methodology. *Languages* 8: 242. https://doi.org/10.3390/languages8040242

Academic Editors: Paolo Mairano and Sandra Schwab

Received: 2 August 2023
Revised: 13 October 2023
Accepted: 17 October 2023
Published: 20 October 2023

1. Introduction

We learn languages for a wide range of reasons, either personal or professional, either in a voluntary or obliged (by the situation) way, because languages open the door to many opportunities of all types. However, we all know how difficult is to learn a second language (L2) after childhood, compared to the easiness of learning (language acquisition) in newborns (De Villiers and De Villiers 1978). Therefore, a great deal of effort has been put on second language learning, especially in the last decades, where globalisation has gained a path, expanding language needs and breaking linguistic barriers.

With the advent of new technologies, L2 has not been excluded from the information and communication technologies era. In this light, a wide range of systems have appeared to help users with the struggle of language learning. These are the so-called computer-assisted (or computer-aided) language learning (CALL) systems, or computer-assisted pronunciation tools (CAPT) (Tejedor García 2020), these are more specifically designed to help with the pronunciation aspect.

CALL and CAPT systems can be diverse, and one of the ways of making them more engaging is to base them on games or robots, especially for grown children (Magaña Redondo 2017; Belpaeme et al. 2018), performing collaborative activities (Robertson et al. 2018), or even through karaoke performances (Murad et al. 2018). Recently, some of these systems include or are based on natural language processing tools, such as text-to-speech, machine translation, or automatic speech recognition (ASR) (Yeh 2014). Most of the natural language processing systems focus on reading, writing, and listening skills. However,

speaking practice is crucial to achieving a good language command (van Doremalen 2014). Therefore, the use of ASR tools has increased over the last decade, since it is capable of recognising the speech of the user, both the content and the spoken characteristics such as low-level pronunciation (phonemes) or high-level speech characteristics (prosody). One of the main reasons for such recent inclusion is the dramatic improvement experienced in this field over the last two decades, mainly due to the revival of deep learning techniques in the field of speech technologies and the development of new algorithms. For an exhaustive and recent review of the ASR, see (Alharbi et al. 2021).

One of the crucial characteristics of ASR systems is their application to a large number of areas. Apart from the specific research on ASR as a standalone field, ASR can be found in biometric, medical, and education applications, among others. In the language learning field, especially in L2 learning, automatic speech recognition becomes a relevant tool for developing automatic assessment systems, based on the fact that: given a standard trained system, the better the pronunciation is, the better the outcome of the ASR will be in terms of word error rate (WER), which is the standard evaluation metric in this field. This hypothesis has been already used in several works with relevant results.

Most of the ASR for L2 have been focused on low-level phonetic pronunciation, both for being more understandable and for sounding more like a native speaker. However, to sound like a native speaker we cannot stick to the phonetic idiosyncrasy of the language. Instead, proper language learning goes beyond good phonetic pronunciation and lexical usage, because language learning relates to the good formation of the languages at each level. In this light, other linguistic levels such as prosody play a very important role when delivering a message, and prosody is also related to other linguistic structures (syntax, semantics, pragmatics, etc.), which should also be assessed in terms of language learning and its corresponding proficiency level, since word accent position, syntactic-prosodic boundaries, and especially rhythm, are underestimated features that help listeners to process syntactic, semantic and pragmatic content (Hönig 2016). To what extent segmental and suprasegmental features are difficult to teach and learn and how important are both of them with respect to language learning is a topic of debate (O'Brien 2020).

However, for several reasons (technical, historical, methodological, etc.) ASR systems have traditionally relied more on the spectral and phonetic levels of the language rather than on prosody. Prosodic characteristics are more difficult to model than other levels, which has always posed a challenge in this aspect (Rosenberg 2018). Nevertheless, the new deep learning paradigm is also changing the way how all the linguistic structures are modelled and how to model their intertwining, and the expansion of technologies is also reaching the education field, as we will see in the following sections.

The current article presents a review of the use of automatic speech recognition applications in the L2 learning field. In the first place, Section 2 presents the methodology carried out for the selection of the papers included in the review, following the PRISMA methodology for systematic reviews, which considers the inclusion and exclusion criteria, the search method, and the selection of articles. The review follows with a brief section (Section 3) dedicated to the main concepts of assessment pronunciation in L2. Section 4 reviews the main selected works regarding the use of ASR tools in L2 learning systems, by highlighting the L1 and L2 languages used in each one, and the linguistic level evaluated by the system. Although companies working in this field do not always disclose their progress due to trade secrets, and it might be difficult to find their research through PRISMA articles, I found it relevant to include, in Section 5, a brief overview of the main commercial systems that address pronunciation assessment, putting more emphasis on those assessing suprasegmental features. Finally, Section 6 draws up the main conclusions from the review and addresses the main remaining challenges.

2. Methodology

The selection of the papers for the current review was performed under the PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analysis) framework for

reporting systematic reviews (Moher et al. 2009). The PRISMA statement includes a 27-item checklist and a flow diagram of four phases. Figure 1 shows the phase diagram with the several steps used in the current article. The paper selection process started with 515 articles, which ended up with 34:26 included in the original selection, and 8 extracted from the bibliography of such 26 papers. The review was performed by the author of the article and did not include a librarian.

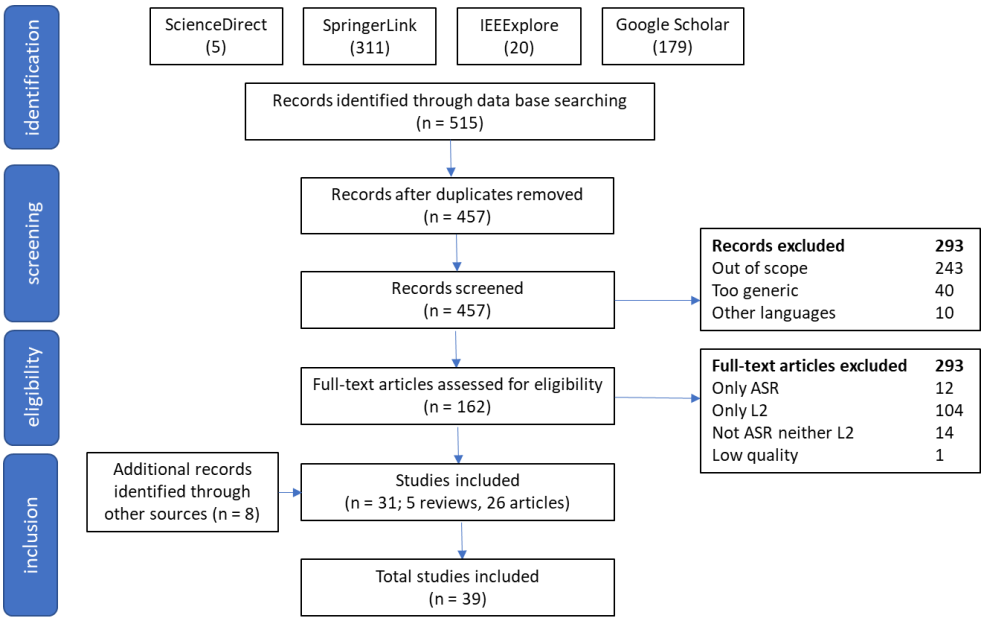

Figure 1. PRISMA flow chart.

2.1. Eligibility Criteria

Since this consists of a state-of-the-art review, I limited the publication period to 10 years, from 2013 to 2023, focusing mainly on those studies concerned with the use of automatic speech recognition (ASR) systems for L2 learning.

2.2. Exclusion Criteria

I excluded those studies in which the use of ASR does not aim directly at improving L2 learning. Those works related to L2 in which ASR or other speech technologies are not explicitly used in the L2 process were not included.

2.3. Search Method

The initial article selection was done through the following databases: SCOPUS, Web of Science, ScienceDirect, IEEE Xplore, and SpringerLink. The search was performed in July 2023. The main keywords used in the searches were: "automatic speech recognition", or "ASR" combined with "L2" and "second language learning". Additionally, keywords such as "prosody", "prosodic", "tone language", "tonal language" and commercial systems" were included in the searches. Only to account for a historical overview, the selection period for general keywords related to ASR and L2 was extended to 20 years.

An example of the search string is asr AND ("l2" OR "second language learning") AND ("prosody") OR "tone language").

2.4. Paper Selection

The first search yielded 515 results. In the first place, the duplicate papers were filtered, resulting in 457 articles. Then, a first screening by title yielded 161 results. The main reasons for excluding the studies were out of scope of the topic (243), too general topics on ASR or L2 or language learning in general (40), or papers written in other languages than English (10).

A second screening through the content of the abstracts and full texts left apart another set of papers, after taking the inclusion and exclusion criteria into account. The main exclusion factors were mainly due to the study being focused on the use of ASR systems but not explicitly on L2—sometimes only on learning in general—(12), or being focused on L2 but not specifically using ASR—although it was mentioned for any reason in the paper—(104). Some other papers were excluded because of the two reasons at the same time (14). One more paper was excluded because of the low quality of the paper in general, although it was addressing both L2 and ASR concepts. In the end, several remaining 26 papers were taken into consideration, plus 5 articles that were also considered for being reviews on the field of study, as well as eight more papers that were extracted from the bibliography of the selected papers.

3. Pronunciation Assessment in L2

Pronunciation helps to achieve intelligibility. Nevertheless, it is not the unique and core aspect of language learning, since this focuses basically on grammar aspects, but oral skills are also essential to achieve a notable language level if the learner wants to interact with other people. However, unlike grammar, oral aspects are difficult to practice in large collective classes or outside of class with appropriate feedback (Bashori et al. 2022; Cucchiarini and Strik 2019; Timpe-Laughlin et al. 2020). In this light, the use of automatic speech recognition (ASR) for L2 became popular about three decades ago, due to an increasing demand for practising oral skills, especially to practice and assess pronunciation assessment (Cucchiarini and Strik 2017). Therefore, most CALL and CAPT systems have been specifically designed to address and practise oral skills and pronunciation, respectively, and a wide range of them include the use of ASR technology.

What are the main aspects that should be assessed in L2 learning has been a topic of debate for many years. Although the standard evaluations of spoken language proficiency usually include pronunciation skills, the variation of aspects in the concept of pronunciation is wide (Pennington and Rogerson-Revell 2019; Danka 2018). Also, the pronunciation within the same language of study can suffer a lot of variation due to a wide range of factors: dialect, sociolect, idiolect, etc., which makes it difficult to compare a learner's pronunciation with the standard or reference native pronunciation, which actually does not exist as such. Moreover, an open debate that still remains unsolved in the design of learning models is the hierarchy of difficulty in L2 phonology (Munro 2021).

In this respect, a wide range of studies have been devoted to analysing the usefulness of ASR systems to assess pronunciation, as well as what oral features are relevant for a good and understandable language level. In (Yaneva 2021), for instance, two existing ASR systems (SpeechTexter and TalkTyper) were used to test the accuracy for Bulgarian as L2. The most common sources of errors, as stated by the author, seemed to come from those consonants not present in Latin languages, and also from accumulations of consonants. More importantly, stress was identified as a common mistake.

At this point, it is relevant to note that most of the CAPT systems put their primary focus on the segmental phonetic level, ignoring the fact that prosody has also been shown essential for intelligibility and comprehensibility (Levis et al. 2022; Zielinski 2006; Frost and Picavet 2014). As we will see in the following section, there are only a few systems focusing on prosodic aspects in L2 assessment. One example is the *Supra Tutor* system (Lima 2020), an online pronunciation tool devoted to focusing on suprasegmental speech characteristics, with a demonstrated impact on the comprehensibility of international teaching assistants.

Another relevant study from (Hirai and Kovalyova 2023) explores the potential of five ASR applications (Google Docs' Voice Typing, Windows 10 Dictation, Apple's Dictation, the website service "Dictation.io," and the iOS application "Transcribe") to be used for L2 English learning with different L1 languages, concluding that, with an accuracy rate ranging between 50 and 70%, automatic speech recognition still remains a challenge compared to pronunciation assessment performed by human evaluators. However, ASR-based tools can still be used with remarkable success in the absence of human raters.

4. Assessment of Linguistic Levels in L2 through ASR

In the 90s, automatic speech recognition systems became popular as a tool to assess and improve L2 learning. Such assessment can be found in terms of different linguistic levels, depending on what aspect of the target language the learner wishes to improve. Since ASR-based CALL and CAPT systems address the acoustic characteristics of the learner's speech, these systems usually focus on pronunciation issues. Within the systems addressing pronunciation, most of them target the phonetic segmental level, while a few of them focus on suprasegmental features based on the prosody of speech, i.e., the elements of intonation, rhythm, and stress. Nevertheless, some CALL systems make use of ASR tools to assess language grammar and lexical aspects.

In what follows, I present a review of the recent findings and systems devoted to the use of ASR technology for L2 in three different levels: (a) grammar and lexical level, (b) phonetic level, and (c) prosodic level. Table 1 presents a summary of the main works appearing in the current section.

Table 1. Overview of the recent systems assessing several linguistic levels for L2 learning.

Reference	ASR/System	L1	L2	Level
(Mansour et al. 2019)	Kaldi-based	Arabic	English	grammar
(Ateeq and Hanani 2019)	Google, SLaTE2018	any	English	grammar
(Ling and Chen 2023)	"Speak and Translate" app	English	Chinese	lexical
(Tejedor García 2020)	Kaldi	any	English, Spanish	phonetic
(Wang and Young 2014)	iCASL	Taiwanese	English	lexical, phonetic
(Arkin et al. 2021)	Tsinghua University	Uyghur	Chinese	phonetic
(Guo et al. 2019)	ASR algorithm	Tibetan	Chinese	phonetic
(Bashori et al. 2022)	ASR-based websites	Indonesian	English	lexical, phonetic
(Guskaroska 2019)	Gboard, Siri, voice dictation on smartphones	Macedonian	English	phonetic
(Escudero et al. 2015)	Android ASR	Spanish	English	phonetic
(Tejedor-García et al. 2020)	Clash of Pronunciations (COP)	Spanish	English	phonetic
(van Doremalen et al. 2014)	bASSIsT, DISCO	any	Dutch	phonetics, morphology, syntax
(Pellegrini et al. 2013)	Daily-REAP	any	Portuguese	phonetic
(Liakin et al. 2015)	mobile ASR	any	French	phonetic
(Mirzaei et al. 2018)	Julius ASR	any	English	phonetic, lexical
(Johnson et al. 2016)	*not specified*	Portuguese	English	prosodic
(Liakin et al. 2017)	Nuance Dragon Dictation	English, Mandarin, Arabic, Spanish	French	phonetic, prosodic
(Demenko et al. 2010)	AzAR3.0	German, Polish, Slovak, Czech, Russian	German, Polish, Slovak, Czech, Russian	phonetic, prosodic

4.1. Grammar and Lexical Assessment

An ASR system converts the spoken utterance into a text string. Therefore, the output of an ASR can be used to assess both the lexical content and the grammaticality of the sentence. Several systems have been developed to fulfil this object, for instance, the English as L2 learning system for Arabic learners (Mansour et al. 2019), in which the spoken sentence from children is recorded and passed through an ASR, and then the output transcription is passed through a series of grammar checkers, such as a *Wh-question* word checker, a grammar checker returning the number of grammar errors in the given question, and a language checker based on machine learning techniques.

Other systems like (Ateeq and Hanani 2019) perform an automatic grammatical evaluation of English speech, by prompting the learner to a question in his native language

and analysing the spoken response in English, by means of several extracted features that allow a language grammar assessment and meaning errors. In a more original way, the system developed by (Ling and Chen 2023) uses an ASR translator—the "Speak and Translate" app included in iPhone and iPad, among others—in which the child pronounces the word in their L1 language and the system returns the same word in L2 so that the child can then utter the L2 word. The system was tested by Australian children learning Chinese.

4.2. Phonetic Assessment

The potential of ASR systems to identify pronunciation errors and to provide appropriate feedback to learners has been widely studied by many authors, as has been shown in previous sections. The majority of the systems addressing pronunciation issues focus on the phonetic aspects of the language, in segmental terms, without considering higher suprasegmental levels that rely on prosody, probably because the prosodic features may be more difficult to extract and assess.

The L1 and L2 languages target in the literature of CALL and CAPT systems is wide, although most of them focus on the most spoken languages in the world, that is, English, Chinese, and Spanish. (Tejedor García 2020), for instance, was developed at the University of Valladolid to perform a phonetic assessment. Using the Kaldi ASR system, the target was Spanish and American English learners. Their global technology also includes other speech tools such as a text-to-speech (TTS) system to give pronunciation feedback in an *exposure-perception-production* cycle. The iCASL system, described in (Wang and Young 2014), provides a list of words that are pronounced both accurately and inaccurately, for Taiwanese learners of English as L2.

Uyghur learners of Mandarin Chinese can improve their language skills with the help of the Chinese automatic speech recognition system developed at the University of Tsinghua (Arkin et al. 2021), which is used to recognise specific phonemes, so that the learners obtain feedback in terms of possible phoneme error categories and possible erroneous pronunciation rules. Tibetan people can also learn Mandarin Chinese by means of the system developed in (Guo et al. 2019). Here, an ASR system is used to identify pronunciation errors. Similarly to (Tejedor García 2020), a speech synthesis system is adopted to correct pronunciation errors. An additional feature of this system is the capability of adjusting the speaking rate of the TTS system to increase comprehension accuracy.

A tool for kids to learn pronunciation skills is also presented by (Bashori et al. 2022), where Indonesian children can perform both vocabulary and pronunciation English tests using an ASR system before and after the task, clearly outperforming the control group, thus showing that learners can successfully learn vocabulary and pronunciation skills. Also for English, Macedonian speakers have at their disposal a tool that provides feedback for vowel pronunciation practice using Gboard, Siri, or any voice dictation on smartphones (Guskaroska 2019). Apart from the demonstrated improvement of pronunciation by using ASR, the relevance of this study relies on the analysis of learners' attitudes towards the use of ASR systems, showing that, despite some minor frustration episodes due to inaccurate feedback, the general message of the learners is that they enjoyed using the tool and found it practical as a learning environment.

As the works from (Guskaroska 2019; Ling and Chen 2023) show, smartphones containing an ASR tool are a good opportunity to integrate such technologies into daily life learning tasks. In line with the previous works, (Escudero et al. 2015) integrate Android ASR and TTS tools for Spanish learners of English as L2, showing that through such tools the learners are able to discern between three pronunciation levels, ranging from basic to native, by means of pronunciation challenges that are presented in the form of minimal pairs.

The use of minimal pairs is a frequent strategy to improve L2 pronunciation. The work found by (Tejedor-García et al. 2020) presents an L2 English tool for native Spanish learners through a mobile app learning game—called Clash of Pronunciations—that focuses on pronunciation training at the phonetic levels, mainly single speech sounds such as vowels,

using the minimal pairs approach, but disregarding any intonation or prosodic information. Moreover, the mobile application includes an automatic speech recognition system, as well as a speech synthesis system, and the corresponding study shows the great potential of games as motivating learning tools.

English and Chinese cope with the majority of L2 learning tools, followed by Spanish, as they are the most spoken languages in the world and thus are perceived as most "useful" to learn for academic and professional purposes, among others. However, many systems still focus on less spoken and less-resourced languages. (van Doremalen et al. 2014), for instance developed the DISCO system, which is based on the bASSIsT ASR system for L2 Dutch learning by providing feedback to learners based on several aspects of oral proficiency and grammar issues, including pronunciation, morphology, and syntax.

Most of the languages have some very specific phonetic issues, and many CAPT systems address such phonetic differences with respect to whatever the L1 is. In Spoken European Portuguese, for instance, the linguistic phenomena of vowel reduction have been shown to be difficult to understand for L2 learners. To properly address it, (Pellegrini et al. 2013) worked on the Daily-REAP (REAder-specific lexical Practice for improved reading comprehension) web platform, which used automatic speech recognition to generate learning material by identifying words pronounced in real speech utterances from broadcast news videos. Another project working on vowel characteristics is the one presented in (Liakin et al. 2015), which addresses French as L2 for elementary French students. The drawback of this study is that it does not specify what is the L1 targeted, which might be relevant for this study because it is aimed at improving the pronunciation of vowel /y/. However, it seems that students using an ASR obtained relevant improvements in pronunciation with respect to the control group and, once again, this work shows the usefulness of adopting a mobile ASR for language learning.

In another direction, it is worth highlighting the work from (Mirzaei et al. 2018), which takes the output of automatic speech recognition and analyses the errors committed to estimating the difficulties in L2 speech. Among the most common types of errors, the authors identify homophones, minimal pairs, negatives and breached boundaries. These findings can further be used to strengthen pronunciation learning of the phenomena that cause more errors.

4.3. Prosodic Assessment

Prosody is conveyed through three different elements: intonation, rhythm, and stress, which are perceived by listeners as changes in fundamental frequency, sound duration, and loudness, respectively (Adami 2007). Prosody takes an important role in delivering a message, and although intelligibility is highly related to the phonetic pronunciation of speech in L2, prosody also plays a crucial role in comprehension (Levis et al. 2022; Zielinski 2006), apart from presenting idiosyncratic characteristics for each language (Hirst and Di Cristo 1998). Nevertheless, very few works address the relevance of prosody in L2 considering their main elements related to intonation and stress, and when they do, the task is usually not performed through an ASR (e.g., (Yenkimaleki and van Heuven 2019; Kang and Johnson 2018; Escudero-Mancebo et al. 2021; Bataineh and Al-Qadi 2014; De Iacovo et al. 2021)).

Among the recent studies that include prosodic information as analysis and feedback for L2 learning using ASR, it is worth mentioning the work from (Johnson et al. 2016) addressed to Brazilian Portuguese learners of English as L2. In this study, an ASR is trained to recognise phones—instead of words—, and then suprasegmental (prosodic) features are computed from the ASR output in terms of phones to extract prominent syllable detection and tone choice classification. This way, an automatic score of the English proficiency of continuous speech can be estimated.

Liakin et al. (2017) is another worth mentioning work that investigates the use of mobile TTS and ASR tools to improve pronunciation skills in French as L2. This work focuses on the pronunciation of the vowel /y/ (as in (Liakin et al. 2015), but here the

authors extend the analysis to resyllabification and liaison as suprasegmental features by making use of the "shadowing" technique—also called "echoing"—and advanced language learning technique in which participants repeat speech immediately after hearing it, used by learners to improve their pronunciation and intonation. (Zhang et al. 2020) go further and present WithYou, an automated adapted speech tutoring system in real time that uses context-dependent speech recognition to evaluate shadowing, so that the tool can adjust the playback and difficulty of the spoken utterance when learners fail to make the shadowing task much easier and customised.

A Pronunciation Trainer was developed by (Demenko et al. 2010) to integrate both phonetic and prosodic levels of learning in a CAPT system. Here, the AzAR3.0 software (Jokisch et al. 2005; Cylwik et al. 2008) includes an ASR tool that helps with individualised and corrective feedback. Here, a prosody test over a set of around 60 sentences is performed to identify the errors that are easier to detect and that are of special relevance for good comprehension, that is, for instance, a non-native-like vowel duration at the phonetic (segmental) level, or an erroneous stress placement at the prosodic (suprasegmental) level. One of the main advantages of AzAR3.0 is the wide range of L1 and L2 languages covered: it includes German as a target language for native speakers of Slovak, Czech, Polish, and Russian, and vice versa.

5. Commercial Systems

Language learning application based on automatic speech recognition is not new. One of the first systems using ASR technology for L2 learning dates back to the late 90s with the development of the Fluency Pronunciation Trainer,[1] (Eskenazi 1999) which was based on the Carnegie speech recognition software to detect L2 pronunciation errors. Considering that the earliest ASR technologies were developed only some decades back in the 50s, the integration of such new speech recognition technology into learning tools that require a good quality output shows their positive success.

Further development of these technologies led to other commercial products based on ASR such as Rosetta Stone (2013)[2], Tell Me More (2013), or EduSpeak (2010), which offer pronunciation exercises to practise both L2 speaking and listening and receive automatic feedback. Usually, the production exercises are based on a quality threshold that the learner must achieve to successfully pass the corresponding exercise. This kind of systems usually even allow a degree of customisation so that the learner can specify the age and gender to receive better feedback. The most common mode of providing feedback is to mark those words that have been correctly or incorrectly pronounced. However, the transparency of these systems is low, in the sense that it is not possible to know which parameters (either phonetic or prosodic) are being taken into account to provide the pronunciation feedback. Therefore, although the learner can repeat the words or sentences several times, it is actually difficult to know where the real focus must be put.

The issue of feedback to the learner has been pointed out by some authors. (McCrocklin 2019), for instance, examines the perception and engagement of ASR-generated transcripts for L2 pronunciation, showing that the involved participants made use of the transcripts to identify individual words with errors. In addition, participants were able to figure out what the cause of the errors was, mainly from segmental and articulatory features, so that they could rehearse the specific parts to improve the production. Thus, this work shows that the learners can guess what kind of feedback the system is providing, despite the low transparency. Nevertheless, this requires a minimum language level from the learner.

The number of systems utilising prosodic features in the learning and feedback process is scarce. The ShadowTalk mobile app (Mrozek 2020), for instance, provides L2 English learners with custom-made recordings. They are then used to help the learners to develop and improve suprasegmental features such as natural rhythm, syllable stress, prominence, and intonation. However, no ASR tool is used.

ELSA[3] (English Language Speech Assistant, (Anguera and Van 2016)) is probably one of the few works using ASR tools and providing feedback on prosody. This application

integrates the Kaldi ASR software, and not only contains pronunciation exercises at the phonetic level, marking phonetic hints to fix existing errors, but also intonation exercises, where users can practice sentence intonation and rhythm, as well as word syllable stress. ELSA also contains conversation exercises that allow the learners to practice real-life conversation and to obtain feedback on both the phonetic and prosodic pronunciation at the word level.

All in all, from a brief review of commercial L2 systems using automatic speech recognition, it transpires that: (1) little or no prosodic information is usually used in L2 commercial applications, or they do not provide enough information to figure out whether segmental or prosodic features are being used in the feedback process. And from those systems providing suprasegmental information, very few of them include an ASR tool. To overcome this, more research in improving algorithms to better calculate prosodic features is needed (Johnson et al. 2016). (2) A significant gap exists between what is being done in research and universities or research centres, and companies developing L2 learning tools (Kochem et al. 2022). (3) A lot of room for improvement still exists in the usability issues, both from the learners' side and the developers teachers' point of view, since they feel discomfort with such new technical learning innovations, especially with respect to their implementation in class.

6. Conclusions

In our globalised era, where moving between countries is increasingly frequent, learning languages has become indispensable. At the same time, when not acquired as a native language, languages are difficult to learn, especially if the learner is not immersed in the country where the language is spoken, and if the time to dedicate to it is scarce. Moreover, learning grammar can be more or less self-taught, but when it comes to oral skills, practicing with other speakers (especially native ones) is essential, which is not always easy in normal classes with several students and dedicated a limited time per week.

This is one of the reasons to see how more and more CALL and CAPT systems have appeared during the last two or three decades to allow learners to practice speaking skills out of class and in an economic way. Therefore, while some of these systems address grammatical aspects, most of them focus on pronunciation issues. Among those systems focusing on pronunciation, most of them rely on phonetic features, while few of them address suprasegmental prosodic features. Moreover, we should not overlook the role that artificial intelligence is taking in the ASR field in this new era. ASR systems such as Whisper and wav2vec2.0 offer free models that are being used for many applications such as automatic assessment of oral reading accuracy for reading diagnosis (Molenaar et al. 2023), or the early detection of cognitive decline by using ASR-based transcriptions (Gómez-Zaragozá et al. 2023).

While the range and diversity of existing CALL and CAPT are wide, the current review has limited the focus on those using automatic speech recognition tools. ASR tool have evolved dramatically during the past decades with the explosion of new and more sophisticated deep learning techniques, so the improvements achieved in these tools convert them into a good option to assess language aspects. The main challenge of such ASR tools remains in using them to assess pronunciation at the phoneme level since they provide very detailed phonetic information, or even grammatical issues when the transcription provided is of enough quality. However, ASR systems have traditionally relied on phonetic modelling ignoring some essential prosodic information, which is highly important in disambiguating semantic or syntactic information to assess grammatical aspects, or to provide native-like pronunciation assessment. The main reason is that suprasegmental characteristics are more difficult to model than segmental phonetic features. Therefore, as pointed out by (Johnson et al. 2016), it is important to develop proper algorithms to better compute prosodic characteristics, putting special emphasis on silent pause detection, filled pause detection, tone unit detection, syllabification, prominent syllable detection, and tone choice classification, which have been shown to comprise useful information for

the assessment of learners' oral skills. This is not to say that CALL and CAPT systems are not considering prosody at all for their analyses to provide learning feedback, but that most of the learning tools including prosody are not making use of an ASR tool. One of the challenges is thus to leverage the dramatic improvement of ASR technologies for developing improved learning technologies, since in the end, recognising the way and the content spoken by the learners is essential to obtain a good assessment of their language skills. Moreover, the inclusion of suprasegmental features in language assessment becomes of high relevance when dealing with tone languages, which represent between 60–70% of the total languages in the world.

Another important point in L2 learning systems is the link between research and applications, and the usability aspects. As stated by (O'Brien et al. 2019), "some researchers and teachers express discomfort with new technological innovations: some implementations are difficult to use, and others are seen as unwelcome replacements for instructors". This reflects that usability is a crucial aspect if we want to take real advantage of the findings and developments of the CALL and CAPT tools in research. So, testing in real environments is a must in a practical field such as language learning. Moreover, as pointed out by (Kochem et al. 2022), "a persistent gap still exists between ASR-equipped software available to participants in research studies and what is available to university and classroom teachers and student", which strengthens the thought that commercial systems are still disconnected from research and vice versa. Once again, similarly to usability, companies, and research centres should work more closely to fill such an availability gap. Works like (Zhang et al. 2020) show that tailoring the learning tasks to the learners made them more suitable and successful for the purpose of the tool.

There is a final point I would like to comment on with respect to the availability of L2 learning systems, which is their target language. As we have seen in the present review, most of the L2 tools are developed to learn the most spoken languages; i.e., mainly English and Chinese. There are other target languages, of course, like French, Portuguese, Spanish, or German, among others. The common pattern is always some of the 10 most spoken languages in the world, and Indo-European languages, apart from Chinese. Cases like Mandarin Chinese as L2 for Tibetan and Uyghur actually reflect the minorisation of the L1 language with respect to the dominant language in the region. While this is of course understandable, because linguistic needs in both the professional and academic fields are highly conditioned by the number of speakers that want to learn, we should not look at the other side when it comes to protecting minority, minorised, and endangered languages. The development of computer-assisted tools needs linguistic resources, which are nowadays at the hands of the top ten most spoken languages, leaving aside the other existing (approximately) 6000 languages in the world. Therefore, both ASR and L2 learning research should also be committed to preserving the linguistic rights of the entire population, creating resources to allow people to learn less spoken languages, and even fostering learning for them (Besacier et al. 2014). If necessary, by making knowledge, resources, and tools open source, free, and publicly available. In the end, there is a lot of knowledge that can be transferred between languages, both in the linguistic and computational fields, and promoting language learning for every language will enrich all of us as speakers, whether it be our mother tongue or not.

Funding: This research was partially funded by the Generalitat de Catalunya under the CLIC-UB 2021 SGR 00313 grant, and by MCIN/AEI/10.13039/501100011033 and the European Union under grant PID2021-124361OB-C33 (FairTransNLP-Language project).

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: No new data were created under the development of this study.

Conflicts of Interest: The author declares no conflict of interest.

Notes

[1] http://www.lti.cs.cmu.edu/Research/Fluency/ (accessed on 18 July 2023).

[2] http://www.rosettastone.com (accessed on 20 July 2023).

[3] https://elsaspeak.com/en/about-us (acessed on 10 July 2023).

References

Adami, André Gustavo. 2007. Modeling Prosodic Differences for Speaker Recognition. *Speech Communication* 49: 277–91. [CrossRef]

Alharbi, Sadeen, Muna Alrazgan, Alanoud Alrashed, Turkiayh Alnomasi, Raghad Almojel, Rimah Alharbi, Saja Alharbi, Sahar Alturki, Fatimah Alshehri, and Maha Almojil. 2021. Automatic Speech Recognition: Systematic Literature Review. *IEEE Access* 9: 131858–76. [CrossRef]

Anguera, Xavier, and Vu Van. 2016. English Language Speech Assistant. In *Interspeech*. San Francisco: International Speech Communication Association. Available online: http://kaldi-asr.org (accessed on 15 June 2023).

Arkin, Gulnur, Askar Hamdulla, and Mijit Ablimit. 2021. Analysis of Phonemes and Tones Confusion Rules Obtained by ASR. *Wireless Networks* 27: 3471–81. [CrossRef]

Ateeq, Mohammad, and Abualsoud Hanani. 2019. Speech-Based L2 Call System for English Foreign Speakers. In *Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*. Berlin/Heidelberg: Springer, pp. 43–53. [CrossRef]

Bashori, Muzakki, Roeland van Hout, Helmer Strik, and Catia Cucchiarini. 2022. 'Look, I Can Speak Correctly': Learning Vocabulary and Pronunciation through Websites Equipped with Automatic Speech Recognition Technology. *Computer Assisted Language Learning* 2022: 1–29. [CrossRef]

Bataineh, Ahmad, and Nasir Al-Qadi. 2014. The Effect of Using Authentic Videos on English Major Students' Prosodic Competence. *Journal of Education and Practice* 5: 157–72.

Belpaeme, Tony, Paul Vogt, Rianne van den Berghe, Kirsten Bergmann, Tilbe Göksun, Mirjam de Haas, Junko Kanero, James Kennedy, Aylin C. Küntay, Ora Oudgenoeg-Paz, and et al. 2018. Guidelines for Designing Social Robots as Second Language Tutors. *International Journal of Social Robotics* 10: 325–41. [CrossRef]

Besacier, Laurent, Etienne Barnard, Alexey Karpov, and Tanja Schultz. 2014. Automatic Speech Recognition for Under-Resourced Languages: A Survey. *Speech Communication* 56: 85–100. [CrossRef]

Cucchiarini, Catia, and Helmer Strik. 2017. Automatic Speech Recognition for Second Language Pronunciation Training. In *The Routledge Handbook of Contemporary English Pronunciation*. Abingdon and New York: Routledge, pp. 556–69. [CrossRef]

Cucchiarini, Catia, and Helmer Strik. 2019. Second Language Learners' Spoken Discourse: Practice and Corrective Feedback Through Automatic Speech Recognition. In *Computer-Assisted Language Learning*. Hershey: IGI Global, pp. 787–810. [CrossRef]

Cylwik, N., G. Demenko, O. Jokisch, R. Jäckel, M. Rusko, R. Hoffmann, A. Ronzhin, D. Hirschfeld, U. Koloska, and L. Hanisch. 2008. The Use of CALL in Acquiring Foreign Language Pronunciation and Prosody-General Specifications for Euronounce Project. *Speech and Language Technology* 11: 123–29.

Danka, Sandor. 2018. Current Debates in the Theory and Teaching of English L2 Pronunciation. *The New English Teacher* 12: 59. Available online: http://www.assumptionjournal.au.edu/index.php/newEnglishTeacher/article/view/3093 (accessed on 6 June 2023).

De Iacovo, Valentina, Marco Palena, and Antonio Romano. 2021. Evaluating Prosodic Cues in Italian: The Use of a Telegram Chatbot as a CALL Tool for Italian L2 Learners. In *Speaker Individuality in Phonetics and Speech Sciences*. Edited by Camilla Bernardasci, Dalila Dipino, Davide Garassino, Stefano Negrinelli, Elisa Pellegrino and Stephan Schmid. Milan: Officinaventuno, pp. 283–98. [CrossRef]

De Villiers, Jill G., and Peter A. De Villiers. 1978. *Language Acquisition*. Cambridge: Harvard University Press. Available online: https://www.hup.harvard.edu/catalog.php?isbn=9780674509313 (accessed on 28 June 2023).

Demenko, Grazyna, Agnieszka Wagner, and Natalia Cylwik. 2010. The Use of Speech Technology in Foreign Language Pronunciation Training. *Archives of Acoustics* 35: 309–29. [CrossRef]

Escudero, David, Enrique Cámara, Cristian Tejedor, César González, and Valentín Cardeñoso. 2015. Implementation and Test of a Serious Game Based on Minimal Pairs for Pronunciation Training. Paper presented at the Workshop on Speech and Language Technology in Education (SLATE), Leipzig, Germany, September 4–5. Available online: https://uvadoc.uva.es/handle/10324/27533 (accessed on 19 June 2023).

Escudero-Mancebo, David, Valentín Cardeñoso-Payo, Mario Corrales-Astorgano, César González Ferreras, Valle Flóres-Lucas, Lourdes Aguilar, Yolanda Martín-De-San-Pablo, and Alfonso Rodríguez-De-Rojas. 2021. Incorporation of a Module for Automatic Prediction of Oral Productions Quality in a Learning Video Game. Paper presented at the IberSPEECH, Valladolid, Spain, March 24–25; pp. 123–26. [CrossRef]

Eskenazi, Maxine. 1999. Using Automatic Speech Processing for Foreign Language Pronunciation Tutor: Some Issues and a Prototype. *Language Learning & Technology* 2: 62–76.

Frost, Dan, and Francis Picavet. 2014. Putting Prosody First—Some Practical Solutions to a Perennial Problem: The Innovalangues Project. *Research in Language* 12: 233–43. [CrossRef]

Gómez-Zaragozá, Lucía, Simone Wills, Cristian Tejedor-Garcia, Javier Marín-Morales, Mariano Alcañiz, and Helmer Strik. 2023. Alzheimer Disease Classification through ASR-Based Transcriptions: Exploring the Impact of Punctuation and Pauses. In *Proceedings of the Interspeech*. Dublin: International Speech Communication Association, pp. 2403–7. [CrossRef]

Guo, Weitong, Hongwu Yang, and Zhenye Gan. 2019. Improving Mandarin Chinese Learning in Tibetan Second-Language Learning by Artificial Intelligent Speech Technology. Paper presented at the International Joint Conference on Information, Media, and Engineering, IJCIME 2019, Osaka, Japan, December 17–19; pp. 368–72. [CrossRef]

Guskaroska, Agata. 2019. *ASR as a Tool for Providing Feedback for Vowel Pronunciation Practice*. Ames: Iowa State University.

Hirai, Akiyo, and Angelina Kovalyova. 2023. Using Speech-to-Text Applications for Assessing English Language Learners' Pronunciation: A Comparison with Human Raters. *English Language Education* 31: 337–55. [CrossRef]

Hirst, Daniel, and Albert Di Cristo. 1998. *Intonation Systems: A Survey of Twenty Languages*. Edited by Daniel Hirst and Albert Di Cristo. Cambridge: Cambridge University Press (CUP).

Hönig, Florian Thomas. 2016. *Automatic Assessment of Prosody in Second Language Learning*. Erlangen: Friedrich-Alexander-Universität.

Johnson, David O., Okim Kang, and Romy Ghanem. 2016. Improved Automatic English Proficiency Rating of Unconstrained Speech with Multiple Corpora. *International Journal of Speech Technology* 19: 755–68. [CrossRef]

Jokisch, Oliver, Uwe Koloska, Diane Hirschfeld, and Rüdiger Hoffmann. 2005. Pronunciation Learning and Foreign Accent Reduction by an Audiovisual Feedback System. In *Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*. Berlin/Heidelberg: Springer, pp. 419–25. [CrossRef]

Kang, Okim, and David Johnson. 2018. The Roles of Suprasegmental Features in Predicting English Oral Proficiency with an Automated System. *Language Assessment Quarterly* 15: 150–68. [CrossRef]

Kochem, Tim, Jeanne Beck, and Erik Goodale. 2022. The Use of ASR-Equipped Software in the Teaching of Suprasegmental Features of Pronunciation: A Critical Review. *CALICO Journal* 39: 306–25. [CrossRef]

Levis, John, Tracey M. Derwing, and Sinem Sonsaat-Hegelheimer. 2022. Second Language Pronunciation: Bridging the Gap between Research and Teaching.

Liakin, Denis, Walcir Cardoso, and Natallia Liakina. 2015. Learning L2 Pronunciation with a Mobile Speech Recognizer: French /Y/. *CALICO Journal* 32: 1–25. [CrossRef]

Liakin, Denis, Walcir Cardoso, and Natallia Liakina. 2017. Mobilizing Instruction in a Second-Language Context: Learners' Perceptions of Two Speech Technologies. *Languages* 2: 11. [CrossRef]

Lima, Edna F. 2020. The Supra Tutor Improving Speaker Comprehensibility through a Fully Online Pronunciation Course. *Journal of Second Language Pronunciation* 6: 39–67. [CrossRef]

Ling, Li, and Weiying Chen. 2023. Integrating an ASR-Based Translator into Individualized L2 Vocabulary Learning for Young Children. *Education and Information Technologies* 28: 1231–49. [CrossRef]

Magaña Redondo, Juan José. 2017. *Audio Trainer Play: Design of a Gamified App for the Development of Audio Skills in a Secondary School Context*. Madrid: Universidad Nacional de Educación a Distancia. Facultad de Filología.

Mansour, Eman, Rand Sandouka, Dima Jaber, and Abualsoud Hanani. 2019. Speech-Based Automatic Assessment of Question Making Skill in L2 Language. In *Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*. Berlin/Heidelberg: Springer, pp. 317–26. [CrossRef]

McCrocklin, Shannon. 2019. Dictation Programs for Second Language Pronunciation Learning: Perceptions of the Transcript, Strategy Use and Improvement. *Konińskie Studia Językowe* 7: 137–57.

Mirzaei, Maryam Sadat, Kourosh Meshgi, and Tatsuya Kawahara. 2018. Exploiting Automatic Speech Recognition Errors to Enhance Partial and Synchronized Caption for Facilitating Second Language Listening. *Computer Speech & Language* 49: 17–36. [CrossRef]

Moher, David, Alessandro Liberati, Jennifer Tetzlaff, Douglas G Altman, and PRISMA Group. 2009. Preferred Reporting Items for Systematic Reviews and Meta-Analyses: The PRISMA Statement. *PLoS Medicine* 6: e1000097. [CrossRef]

Molenaar, Bo, Cristian Tejedor-Garcia, Catia Cucchiarini, and Helmer Strik. 2023. Automatic Assessment of Oral Reading Accuracy for Reading Diagnostics. Paper presented at the Interspeech, Dublin, Ireland, August 20–24; pp. 5232–36. [CrossRef]

Mrozek, Patryk Mikołaj. 2020. *ShadowTalk: A Prosody-Training Mobile App for English as a Second or Foreign Language Students*. Long Beach: California State University.

Munro, Murray J. 2021. On the Difficulty of Defining 'Difficult' in Second-Language Vowel Acquisition. *Frontiers in Communication* 6: 639398. [CrossRef]

Murad, Dania, Riwu Wang, Douglas Turnbull, and Ye Wang. 2018. SLIONS: A Karaoke Application to Enhance Foreign Language Learning. In Paper presented at the MM 2018—Proceedings of the 2018 ACM Multimedia Conference, Seoul, Republic of Korea, October 22–26; Volume 18, pp. 1679–87. [CrossRef]

O'Brien, Mary Grantham, Tracey M. Derwing, Catia Cucchiarini, Debra M. Hardison, Hansjörg Mixdorff, Ron I. Thomson, Helmer Strik, John M. Levis, Murray J. Munro, Jennifer A. Foote, and et al. 2019. Directions for the Future of Technology in Pronunciation Research and Teaching. *Journal of Second Language Pronunciation* 4: 182–207. [CrossRef]

O'Brien, Mary Grantham. 2020. Ease and Difficulty in L2 Pronunciation Teaching: A Mini-Review. *Frontiers in Communication* 5: 626985. [CrossRef]

Pellegrini, Thomas, Rui Correia, Isabel Trancoso, Jorge Baptista, Nuno Mamede, and Maxine Eskenazi. 2013. ASR-Based Exercises for Listening Comprehension Practice in European Portuguese. *Computer Speech & Language* 27: 1127–42. [CrossRef]

Pennington, Martha C., and Pamela Rogerson-Revell. 2019. Assessing Pronunciation. *English Pronunciation Training and Research*, 287–342. [CrossRef]

Robertson, Sean, Cosmin Munteanu, and Gerald Penn. 2018. Designing Pronunciation Learning Tools: The Case for Interactivity against over-Engineering. Paper presented at the Conference on Human Factors in Computing Systems, Montreal, QC, Canada, April 21–26.

Rosenberg, Andrew. 2018. Speech, Prosody, and Machines: Nine Challenges for Prosody Research. Paper presented at the 9th International Conference on Speech Prosody 2018, Poznan, Poland, June 13–16; pp. 784–93. [CrossRef]

Tejedor García, Cristian. 2020. Design and Evaluation of Mobile Computer-Assisted Pronunciation Training Tools for Second Language Learning. Ph.D. thesis, Universidad de Valladolid, Valladolid, Spain. [CrossRef]

Tejedor-García, Cristian, David Escudero-Mancebo, Valentín Cardeñoso-Payo, and César González-Ferreras. 2020. Using Challenges to Enhance a Learning Game for Pronunciation Training of English as a Second Language. *IEEE Access* 8: 74250–66. [CrossRef]

Timpe-Laughlin, Veronika, Tetyana Sydorenko, and Phoebe Daurio. 2020. Using Spoken Dialogue Technology for L2 Speaking Practice: What Do Teachers Think? *Computer Assisted Language Learning* 35: 1194–217. [CrossRef]

van Doremalen, Joost, Lou Boves, Catia Cucchiarini, and Helmer Strik. 2014. Implementation of an ASR-Enabled CALL System for Practicing Pronunciation and Grammar: The BASSIsT System. In *Developing Automatic Speech Recognition-Enabled Language Learning Applications: From Theory to Practice*. Edited by Joost van Doremalen. Nijmegen: Radboud University Nijmegen, pp. 71–94.

van Doremalen, Joost. 2014. *Developing Automatic Speech Recognition-Enabled Language Learning Applications: From Theory to Practice*. Nijmegen: Radboud Universiteit Nijmegen.

Wang, Yi-Hsuan, and Shelley Shwu-Ching Young. 2014. A Study of the Design and Implementation of the ASR-Based ICASL System with Corrective Feedback to Facilitate English Learning. *Educational Technology & Society* 17: 219–33.

Yaneva, Alexandrina. 2021. *Speech Technologies Applied to Second Language Learning. A Use Case on Bulgarian*. Barcelona: Universitat Pompeu Fabra. Available online: http://repositori.upf.edu/handle/10230/48854 (accessed on 20 July 2023).

Yeh, Rosa. 2014. *Effective Strategies for Using Text-to-Speech, Speech-to-Text, and Machine-Translation Technology for Teaching Chinese: A Multiple-Case Study*. Prescott Valley: Northcentral University.

Yenkimaleki, Mahmood, and Vincent J. van Heuven. 2019. The Relative Contribution of Computer Assisted Prosody Training vs. Instructor Based Prosody Teaching in Developing Speaking Skills by Interpreter Trainees: An Experimental Study. *Speech Communication* 107: 48–57. [CrossRef]

Zhang, Xinlei, Takashi Miyaki, and Jun Rekimoto. 2020. WithYou: Automated Adaptive Speech Tutoring with Context-Dependent Speech Recognition. Paper presented at the 2020 CHI Conference on Human Factors in Computing Systems, Honolulu, HI, USA, April 25–30.

Zielinski, Beth. 2006. The Intelligibility Cocktail: An Interaction between Speaker and Listener Ingredients. *Prospect* 21: 22–45. Available online: https://researchers.mq.edu.au/en/publications/the-intelligibility-cocktail-an-interaction-between-speaker-and-l (accessed on 19 July 2023).