Computer Speech & Language 62 (2020) 101055

Contents lists available at ScienceDirect

Computer Speech & Language

journal homepage: www.elsevier.com/locate/csl

A survey on automatic speech recognition systems for Portuguese language and its variations

Thales Aguiar de Lima, Márjory Da Costa-Abreu*

Department of Informatic and Applied Mathematics, Federal University of Rio Grande do Norte (UFRN), Natal, Brazil

ARTICLE INFO

Article History:
Received 24 June 2019
Revised 9 August 2019
Accepted 30 November 2019
Available online 6 December 2019

Keywords:
Automatic speech recognition
Text-to-speech
Speech recognition
Portuguese
Survey

ABSTRACT

Communication has been an essential part of being human and living in society. There are several different languages and variations of them, so you can speak English in one place and not be able to communicate effectively with someone who speaks English with a different accent. There are several application areas where voice/speech data can be of importance, such as health, security, biometric analysis or education. However, most studies focus on English, Arabic or Asian languages, neglecting other relevant languages, such as Portuguese, which leaves their investigations wide open. Thus, it is crucial to understand the area, where the main focus is: what are the most used techniques for feature extraction and classification, and so on. This paper presents a survey on automatic speech recognition components for Portuguese-based language and its variations, as an understudied language. With a total of 101 papers from 2012 to 2018, the Portuguese-based automatic speech recognition field tendency will be explained, and several possible unexplored methods will be presented and discussed in a collaborative and overall way as our main contribution.

1. Introduction

For thousands of years, the ability to communicate is seen as one of the most important human interactions (Stacks and Salwen, 2014). By writing, drawing, making gestures, and speaking, the individual can share ideas, demonstrate their feelings, make allies, and even enemies. Communication impacts on relationships, cultural diversity and social backgrounds (Ruesch et al., 2017). While every kind of communication has a uniqueness, there is a random aspect that makes it part of the individual. Such characteristic cannot be dissociated from their social and cultural group (Muslim, 2007).

This scenario is also the case for oral communication, a natural human social phenomenon. The uniqueness of speech can be found on the 7097 living languages in the world, according to Eberhard et al. (2018). The same source defines a living language as “one that has at least one speaker for whom it is their first language”.

The great variety of currently spoken languages, however, does not mean that they are uniformly distributed over the world population. Since only 23 are spoken by more than half of the world population (Eberhard et al., 2018). The most popular spoken languages benefit from a significant number of resources for Automatic Speech Recognition (ASR), Natural Language Processing, and Computational Linguistics. An unpopular language, on the other hand, suffers from a lack of resources for research and development of dedicated technologies. Thus, the importance that the development of such techniques has for under-resourced languages is worth mentioning.

Some languages, however, are in-between popularity and resourceness. That is the case for the Portuguese language, the sixth most spoken language in the world by native speakers (Eberhard et al., 2018). However, it is far from having the same amount of

*Corresponding author.
E-mail addresses: thalesaguiar21@gmail.com (T. Aguiar de Lima), marjory@dimap.ufrn.br (M. Da Costa-Abreu).

https://doi.org/10.1016/j.csl.2019.101055
0885-2308/© 2019 Elsevier Ltd. All rights reserved.

resources as Mandarin, Spanish or English. This distinct position of the language instigated researchers to focus its efforts on it between 2012 and 2018.

Many breakthroughs in computer science research can be credited to the increasing number of computational resources developed or recently made available. ASR is on track with these advances and presented substantial improvements in classification tasks (Arel et al., 2010). However, under-resourced languages usually retrain popular acoustic models on a new target language. That is not only unreliable (Schultz and Kirchhoff, 2006) but also inconsiderate phonetic and grammatical differences of communication systems (Besacier et al., 2014).

This paper proposes to summarise and evaluate studies on ASR for the Portuguese language and its variations, as well as to find literature gaps and draw directions for future research. Therefore, creating a starting point for new scientists on the topic.

This paper is organised as follows. Section 2 will present the criteria used to select the works for this review, followed by Section 3 with an introduction to ASR. Next, the Section 4 overviews the resources available for ASR, introducing a list of Corpus in Section 4.1, and a set of tools and software explored in the search course in Section 4.2. Then a overview of feature vectors in Section 6, followed by some applications in Section 5. A categorisation for Speech-to-Text (STT) systems is given with respect to the approaches in Section 7. After, the Section 8 highlight the research gaps and possible directions, concluding in Section 9.

2. Our contribution search methodology

To limit the scope of this piece of research, we have decided to limit our web search starting in 2012, which produced a list of 663 papers. They were manually searched with the support tool Google Scholar, using a full-text search through all databases that this tool may reach. Also, a proper search was done in the following databases: Elsevier, IEEE, Springer, and ACM. The passphrases "speech recognition" and word "Portuguese" were included, while "scielo", "text-to-speech", "named entity", "emotion", and "sign language" were excluded. The reason to exclude works with some of these keywords was that they could lead to a context for speech recognition that diverges from the computational definition.

From the collection, we removed papers that focused on language processing, insufficiently described their results or methods, or presented conclusions unsupported by the results. Also, we only included journal or conference papers as scientific resources. Furthermore, the language variations (such as regional dialects or idioms) were not restricted, given that they would impact on a small number of samples for this survey. These restrictions screened the total of papers to 101 publishes in the search period.

Information for statistical measures was gathered for each paper. Among the collected information, the feature extraction and classification methods were highlighted, alongside datasets and software.

On the graph shown in Fig. 1, the distribution of papers over the selected time interval is displayed. It reveals a considerable drop in published works on the subject after 2014, which is another justified reason for us to understand the area and highlight the possible paths for researching.

This section presented the strategy used to collect the works and how the publications are distributed in the search period. In summary, this set of information allows this paper to be used as a starting point to initiate future research. Next, a short introduction to ASR systems is given.

3. Automatic speech recognition

An ASR system can be seen as a mathematical model that can make a speech-to-text conversion, generating text corresponding to the recognised pieces of speech input (Ghai and Singh, 2012).

Far from being trivial, such models may have to deal with a signal containing different kinds of noises, various speakers, and its particular characteristics. Examples of these characteristics are the rhythm or pace, accent, dialectical pronunciation, peculiar intonations, and even mispronunciations (Muslim, 2007).

The variety of languages led ASR systems to have mainly developed into different language-specific systems to cope with resource availability such as speech corpus (Besacier et al., 2014).

Even though ASR systems still comprise of language-specific models, recent research is showing that multi-language models are not only feasible but can also be used to bootstrap models for low or under-resourced languages. It is possible to argue that human languages are, in fact, not that different from each other when it comes to computational modelling (Tong et al., 2017; Vu et al., 2014). Despite this, it is of high importance to focus on specifics of the main spoken languages that are still under-investigated, such as Portuguese.

As already said, this survey focuses on researches that use Portuguese language and its variations, and since multi-lingual speech recognition is a trend on the field, it will be covered in this research as well. The researches that relate to speech recognition but could not be strictly classified as an speech recognition problem are also covered. For instance, disease diagnoses, assisting language acquisition and speech style detection, given that the proposed models are very related to ASR.

Works that focus on the state of the art of automatic speech recognition systems usually comprises several components, each contributing to the overall ability to incorporate knowledge from acoustic or phonetic modelling, language modelling, and lexical or pronunciation modelling. In fact, for a long time, the most common approach for proposing new models involved trying to take advantage of knowledge from different sources to improve the model efficiency (Chebotar and Waters, 2016).

On a somewhat different direction, big data approaches try to take advantage of the training data to propose or improve existing models (Kapralova et al., 2014). This approach has received a long-time negative criticism due to the idea that they propose blackbox models, making it difficult to interpret how the model works, and even giving descend to the very reliability to these models,

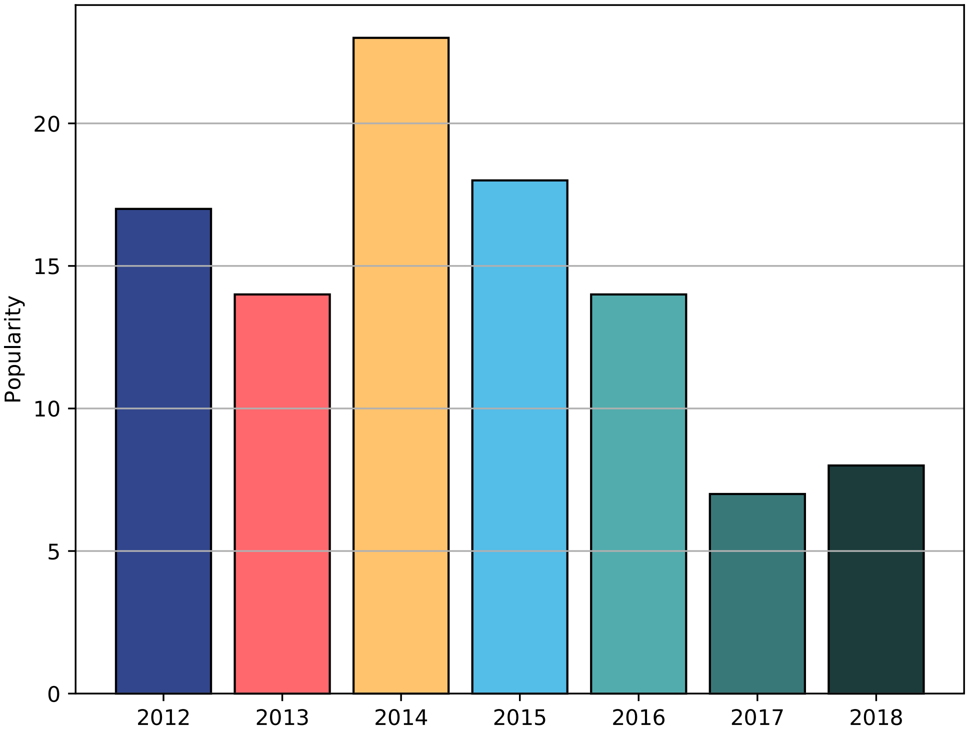

Fig. 1. Distribution of published works on Portuguese ASR between from January 01 of 2012 and December 31, of 2018.

since it is harder to predict when a model will fail (Torres-Huitzil and Girau, 2017). Even with these concerns, the performance improvement of recent deep learning models are undeniable when compared to already-established models (Zen et al., 2016).

The following sections will present some of the primary resources that are associated with automatic speech recognition.

4. Resources for automatic speech recognition

Since the research on Portuguese is limited, it has been challenging to develop technologies for such communication systems. Thus, this section presents the primary resources, described in the collected works, for Portuguese languages. Although, it is essential to remember that the resourceness for a language is defined according to the acoustic model and the amount of data required to prevent overfitting and other problems (Heigold et al., 2013). For instance, the same language could be considered as either well-resourced and under-resourced when applied to a classifier based on Deep Neural Network (DNN) and Multilayer Perceptron (MLP) or Support Vector Machines (SVM) approach, respectively, since the former needs far more data than the latter (Besacier et al., 2014). A vital collection ASR resources for Portuguese is given by (Neto et al., 2011), where several tools and corpora are mentioned, including the LapsMail corpus.

To better understand, the resources are divided into databases and toolkits. The next two sections will explore the main recent contributions for Portuguese automatic speech recognition.

4.1. Corpus

Public representative databases are essential for any research in speech recognition. In this area, when referring to a database, the authors use the terminology “Corpus” (or “Corpora”) (Robin, 2018). During this research, we did not found works that used a large set of structured sentences, that is, a text corpora. Therefore, all datasets presented in this section are speech corpora, a collection of voice recordings. Several papers have a data acquisition procedure, both as to its primary goal (Abad et al., 2013; Candeias et al., 2013; Cristoforetti et al., 2014; Deoldoto Paulino et al., 2018; Gretter, 2014; Hämäläinen et al., 2013; Lopes et al., 2012a; Proença et al., 2015; 2016; Raso and Mello, 2012; Rodríguez-Fuentes et al., 2016; Schultz et al., 2013) or as a requirement for feeding proposed models. Thus, this section will present the corpora and their features related to a speaker (gender, age, etc.), signal (noise, recording rate, etc.), and their availability.

Tables 1 presents the list of all the Portuguese corpora along with their utterance features to the day this paper has been published. Tables 2 and 3 displays the features of the most recent noisy corpora employed in the literature, while Tables 4 and 5 the clean corpora. Whereas Table 5 has monolingual datasets, Tables 3 and 4, on the other hand, have multilingual corpora. Their size is mixed because some authors measure them by hours, while others by the number of utterances. So, without loss of representation, when the size is in utterances and hours, it will be appended with *u* and *h*, respectively.

Table 1
List of Portuguese corpora along with their utterance features.

Papers	Corpus	Size	Noisy	Lang	kHz	Speakers	Gender	Age	Level	Open
(Gretter, 2014; Müller et al., 2016a; 2017)	Euronews	940h	✓	10	-	-	-	-	LVCSR	✘
(Abad et al., 2013)	ASP-I	1.004u	✓	1	8	8	3/5	52–78	Word	✘
(Abad et al., 2013)	ASP-II	850u	✓	1	16	18	6/10	19–62	Word	✘
(Candeias et al., 2013)	HESITA	27h	✓	1	16	-	-	-	LVSR	✓
(Cristoforetti et al., 2014; Matos et al., 2014; Ravanelli et al., 2015)	DIRHA	9.200u	✓	4	48	30	15/15	25–50	LVSR	✓
(Bonilla C et al., 2016; Gelly et al., 2016; Quintanilha et al., 2017; Zen et al., 2016)	Spoltech	2.540u	✘	1	44.1	477	-	-	*	✘
(Furtado et al., 2013)	Furtado	330u	✘	1	44.1	33	9/24	20–50	Word	✘
(Ghoshal et al., 2013; Lal and King, 2013; Lu et al., 2014; 2012; Mohan and Rose, 2015; Swietojanski et al., 2012; Tong et al., 2018; Vesel et al., 2012; Vu et al., 2014)	Global Phone	450h	✘	22	16	2.100	1050/1050	-	Word	✘
(Hämäläinen et al., 2012; Pellegrini et al., 2012)	MLDC	90h	✘	1	16	986	714/272	60–100	Word	✘
(Hämäläinen et al., 2014b; 2014c; 2014d)	CNG	21h	✘	1	22	510	286/224	3–10	Word	✘
(Lopes et al., 2012a)	EPCSC	3h	✘	1	-	101	56/55	5–6	Word	✘
(Lopez-Otero et al., 2012)	KALAKA-2	113h	✓	6	16	-	-	-	LVCSR	✘
(Oliveira et al., 2012b)	LapsMail	2.150u	✓	1	16	25	4/21	-	LVSR	✓
(Hämäläinen et al., 2014d; Pellegrini et al., 2013b)	BD-PUBLICO	20h	✘	1	-	100	50/50	19–30	Word	✓
(Raso and Mello, 2012)	CORALBR	21h	✘	1	-	362	183/179	18–60	LVSR	✓
(Proença et al., 2014)	PDSC	1.5h	✘	1	48	22	16/13	25–80	Word	✘
(Santos et al., 2015)	Biochaves	400u	✓	1	8	8	2/6	-	Word	✓
(Veiga et al., 2014)	Tecnovoz	22.627u	✘	1	-	368	-	-	Word	✘
(Hämäläinen et al., 2014a; 2014d; Pellegrini et al., 2013b; 2014)	EASR	44,033u	✘	1	16	778	203/575	60–100	Word	✘
(Proença et al., 2016)	LestRead	20h	✓	1	-	284	-	6–10	Word	✘
(Karmele et al., 2015)	AZTIAHO	10h	✓	8	16	40	20/20	20–98	LVCSR	✘
(Ribeiro et al., 2015)	AUDIMUS	75.5h	✘	-	32	-	-	-	LVCSR	✘
(Gonzalez-Dominguez et al., 2015; Heigold et al., 2013; Kapralova et al., 2014; Lei et al., 2013; Vanhoucke et al., 2013)	GoogleVoice	3034.5h	✓	34	-	-	-	-	LVCSR	✘
(Deoldoto Paulino et al., 2018)	BrSD	3,7h	✘	-	80	40/40	9–81	-	Word	✘

The majority of papers that collected data did not make it public nor detailed it enough. For instance, (Proença et al., 2014) compiled a series of speech audios from Parkinson's disease patients but did not provided any means for others to obtain this data. Similarly, some authors specified the dataset attributes in details; however, did not include any reference for validity (Silva et al., 2012). Other papers did not include enough information, leaving some rather essential features occult, represented by '-'.

Even though the number of corpora seems enough, the quantity of data is still scarce. Besides, there was no corpus composed of phoneme recordings in the period. Although some works make use of phonetic features (Hämäläinen et al., 2014c; Lopes et al., 2012a), but their utterances are recorded as words. Furthermore, the corpora with word recordings are more prevalent, followed by Large Vocabulary Speech Recognition (LVSR), and Large Vocabulary Continuous Speech Recognition (LVCSR). The last two being more challenging to process, while the Word and LVSR are harder to gather compared to LVCSR because they need to create several phonetically rich sentences (prompts) that subjects must read.

Global Phone is the most popular corpus, as shown in Fig. 2, it was explored in (Ghoshal et al., 2013; Lal and King, 2013; Lu et al., 2012; Mohan and Rose, 2015; Swietojanski et al., 2012; Vesel et al., 2012). This dataset is a project introduced by (Schultz, 2002), and further updated in (Schultz et al., 2013; Schultz and Schlippe, 2014), aiming to be a multilingual corpus and uniform across all languages. In the current state of this collection and this research, the project has 22 languages and more than 450 hours of transcribed data. For each language, there are approximately 100 speakers with relative balance about gender; thus, it has more than 2.100 native speakers.

The GlobalPhone data was recorded with basically the same instruments, and a similar environment to guarantee a high standardisation across languages. Due to this uniformity and the presence of multiple languages, this corpus was used in most of the works from 2012 to 2018. However, only the language models are available for download, while the speech, text, and dictionary data are distributed under a research and commercial license by two distributors[1].

[1] European Language Resources Association (ELRA) and Appen Buttler Hill Pty Ltd. (ABH).

Table 2
List of monolingual noisy Portuguese corpora along with their signal characteristics.

Papers	Corpus	Size	kHz
(Abad et al., 2013)	ASP-I	1.004u	8
(Abad et al., 2013)	ASP-II	850u	16
(Candeias et al., 2013)	HESITA	27h	16
(Oliveira et al., 2012b)	LapsMail	2.150u	16
(Santos et al., 2015)	Biochaves	400u	8
(Proença et al., 2016)	LestRead	20h	–

Table 3
List of multilingual noisy Portuguese corpora along with their signal characteristics.

Papers	Corpus	Size	Lang	kHz
(Gretter, 2014; Müller et al., 2016a; 2017)	Euronews	940h	10	-
(Cristoforetti et al., 2014; Matos et al., 2014; Ravanelli et al., 2015)	DIRHA	9.200u	4	48
(Lopez-Otero et al., 2012)	KALAKA-2	113h	6	16
(Karmele et al., 2015)	AZTIAHO	10h	8	16
(Gonzalez-Dominguez et al., 2015; Heigold et al., 2013; Kapralova et al., 2014; Lei et al., 2013; Vanhoucke et al., 2013)	GoogleVoice	3034.5h	34	-

Table 4
List of multilingual clean Portuguese corpora along with their signal characteristics.

Papers	Corpus	Size	Lang	kHz
(Ghoshal et al., 2013; Lal and King, 2013; Lu et al., 2014; 2012; Mohan and Rose, 2015; Swietojanski et al., 2012; Tong et al., 2018; Vesel et al., 2012; Vu et al., 2014)	Global Phone	450h	22	16

Around 52% of datasets have a type of noise on its audios or text, which allows researchers to investigate the noise robustness of their models. However, only 29% of the datasets have more than one language (5 noisy listed in Table 2 plus 1 listed in Table 3, from a total of 21 datasets). From these, only GlobalPhone has clean recording, making it challenging to research under both these conditions.

Another concerning characteristic of speech datasets is about the number, quantity, and age of speakers. Besides, knowing which tasks each volunteer had to perform is also of interest, since it may define the work as Word Recognition, Large Vocabulary Speech Recognition (LVSR), Large Vocabulary Continuous Speech Recognition (LVCSR), Phoneme Recognition, and others (Rabiner and Juang, 1993). All those features are listed in Table 7, for public datasets, and Table 6, for restrict.

Furthermore, the majority of corpus are not balanced concerning speaker gender, which can impact on ASR systems (Rabiner and Juang, 1993), as shown in Tables 7 and 6. The DIRHA (Cristoforetti et al., 2014) and BD-PUBLICO (Neto et al., 1997) are balanced, while LapsMail (Oliveira et al., 2012b) and EASR (Pellegrini et al., 2013b) are not.

Therefore, there is a considerable amount of datasets for ASR in Portuguese, but most of them are small and lacks on both balance and standardisation. Consequently, this is often used as the motivation of their acquisition since most of them were

Table 5
List of monolingual clean Portuguese corpora along with their signal characteristics.

Papers	Corpus	Size	kHz
(Bonilla C et al., 2016; Gelly et al., 2016; Quintanilha et al., 2017; Zen et al., 2016)	Spoltech	2.540u	44.1
(Furtado et al., 2013)	Furtado	330u	44.1
(Hämäläinen et al., 2012; Pellegrini et al., 2012)	MLDC	90h	16
(Hämäläinen et al., 2014b; 2014c; 2014d)	CNG	21h	22
(Lopes et al., 2012a)	EPCSC	3h	-
(Hämäläinen et al., 2014d; Pellegrini et al., 2013b)	BD-PUBLICO	20h	-
(Raso and Mello, 2012)	CORALBR	21h	-
(Proença et al., 2014)	PDSC	1.5h	48
(Veiga et al., 2014)	Tecnovoz	22.627u	-
(Hämäläinen et al., 2014a; 2014d; Pellegrini et al., 2013b; 2014)	EASR	44.033u	16
(Ribeiro et al., 2015)	AUDIMUS	75.5h	32
(Deoldoto Paulino et al., 2018)	BrSD	3.7h	-

Table 6
Restrict Portuguese corpora and its speakers characteristics. In the Gender column, the format X/Y is for Female/Male speakers.

Corpus	Speakers	Gender	Age	Level
Euronews	–	–	–	LVCSR
ASP-I	8	–	–	Word
ASP-II	18	–	–	Word
Spoltech	477	–	–	*
Furtado	33	9/24	20–50	Word
Global Phone	2100	–	–	Word
MLDC	986	714/272	60–100	Word
CNG	484	–	3–10	Word
EPCSCAT	101	56/55	5–6	Word
PDSC	22	16/13	25–80	Word
EASR	778	203/575	60–100	Word
LestRead	284	–	6–10	Word
AZTIAHO	40	20/20	20–98	LVCSR
AUDIMUS	–	–	–	LVCSR

collected aiming to a specific application. For example, children speech disorders (Lopes et al., 2012a) and voice commands from throat microphone (Ribeiro et al., 2018) were acquired in the Portuguese language. Thus, the data is still scarce, but very diverse with noise, age, medical, commands, and other features.

4.2. Tools and software

When developing a new model or working on annotating a new corpus, the decision to use existing tools may be essential to save time by taking advantage of experiences and improvements made by previous research. This section briefly mentions some tools used by researchers that appears in at least two of our selected papers. Thus, these tools and software are related to Portuguese language.

The HTK is one of the first freely available software to be used by the research community. It was made with (Cambridge, 1993) and written in C language. Besides, it provides recipes to build baseline systems with Hidden Markov Models (HMM). From our list of paper, the ones which use this toolkit are (Veiga et al., 2014; Veras et al., 2014; Zen et al., 2016). It can also be used for other applications rather than speech recognition.

Sphinx, employed by Cox and Davies (2012); Oliveira et al. (2012b); Silva et al. (2012); Souza and Neto (2016), is another popular software package developed under Berkeley Software Distribution licence in Java programming language by researchers from Carnegie Mellon University (University, 1986) and maintained by a thriving community for decades, just like HTK.

The Praat (Boersma and Weenink, 1991) software is a GUI-based tool for visual analyses and manipulation of audio files. This software was tested in (Souza and Neto, 2016) to develop an automatic phonetic aligner for Brazilian Portuguese. In addition, this tool was also explored in Hämäläinen et al. (2014b); Karmele et al. (2015); Hämäläinen et al. (2014c); Mendoza et al. (2014); Proença et al. (2014). Furthermore, this is the most popular open-source tool for speech phone alignment and corpus annotation. However, this tool was most used as a feature-extraction software, as well as openSMILE, as it has some popular features available.

Julius (Akinobu, 1997) was initially developed with Kyoto University. It has several speech recognition features, such as forced-alignment tools, and almost real-time computing. Created initially for Japanese, it has expanded to accept other languages, while maintaining the goal of being a very efficient toolkit for LVCSR. As HTK, it is based on HMM models.

The Kaldi software (Povey et al., 2011), used by (Batista et al., 2018; Ghoshal et al., 2013; Gelly et al., 2016; Lee et al., 2016; Lu et al., 2014; Mohan and Rose, 2015; Ravanelli et al., 2015; Swietojanski et al., 2012; Tong et al., 2017; Vu et al., 2014), began to be

Table 7
Public Portuguese corpora and its speakers characteristics. In the Gender column, the format X/Y is for Female/Male speakers.

Corpus	Speakers	Gender	Age	Level
HESITA	-	-	-	LVSR
DIRHA	30	15/15	25–50	LVSR
LapsMail	25	4/21	-	LVSR
BD-PUBLICO	100	50/50	19–30	Word
CORALBR	362	183/179	18–60	LVSR
Biochaves	8	2/6	-	Word

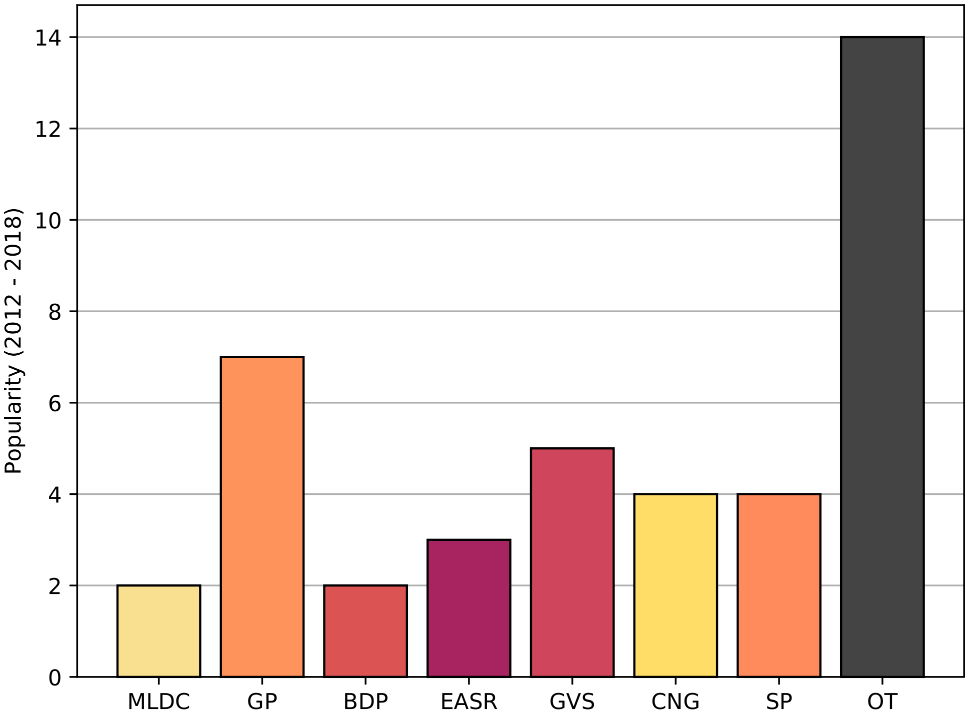

Fig. 2. The popularity of corpora from 2012 to 2018. Global Phone (GP); BD-PBLICO (BDP); European Portuguese Elderly Speech (EASR); Google Voice Search (GVS); Corpus of European Portuguese Childrens Speech (CNG); Spoltech (SP); Sum of popularity of other datasets (OT).

developed in 2009, on a workshop in Johns Hopkins University entitled “Low Development Cost, High-Quality Speech Recognition for New Languages and Domains” (Povey et al., 2009). This software is becoming very popular in the research community. In fact, it is the most popular among our collection of papers. Its model is based on a Subspace Gaussian Mixture Model (SGMM), coded on C++ programming language. Another speech-to-text tool that is worth mentioning is Voice Note (Lui et al., 2004).

Not only speech-to-text but also speech related applications have such tools and software. For speaker recognition systems, there is Alize, and BOSARIS (Brümmer and De Villiers, 2013), used in Pellegrini et al. (2014). For language modelling, the SRILM allows users to quickly generate language models for statistical classifiers.

Every tool presented in this section is free to use. Besides, Praat alongside Julius and openSMILE are open-source software. In regards to their operating system (OS) support, only HTK is available exclusively for Linux, while Praat has a Mac version and Sphinx can be compiled for this OS. However, some of the software above seem to be outdated since the last release for HTK was in 2016, Julius most recent log was in 2014, and Praat does not share this info. On the other hand, Sphinx and Kaldi are regularly updated on their respective repositories. Furthermore, C and C++ are the most popular languages for this software, even though Python and Java also have a presence. It is clear from what this section presented that the area is plentiful of tools when compared with the availability of datasets, and the use of other techniques for speech processing. This behaviour is justified by the fact that Portuguese-based solutions have not been the focus when exploring automatic speech recognition. The next section will explore the main areas that ASR for Portuguese has been applied.

5. Applications

The current state of ASR is allowing the development of systems for objectives other than converting a signal to text, such as treatment of speech dysfunctions (Abad et al., 2013; Rocha et al., 2017). Therefore, this section will present papers that used speech recognition techniques as their secondary goal. Most of the applications rely on the monolingual approach (more details in Section 7), such as an evaluation of children reading capabilities (Proença et al., 2015; 2016) further analysed in Proença et al. (2017). The second language learning, addressed in more details by (Pellegrini et al., 2013a), is also approached using both scored Elicited Oral Response (Cox and Davies, 2012) in conjunction to a test to evaluate proficiency in a language, i.e. Portuguese, and the Goodness of Pronunciation (Ribeiro et al., 2015). The latter is more of qualitative work, compared to the former.

The speech recognition is also applied as a more natural interface for humans to interact with computers. To establish a better interaction between children and technology, in (Alves et al., 2014), a speech-driven robot that executes a set of commands is constructed. Using voice commands, it is also possible to operate medical and graphics systems (Furtado et al., 2015; Rocha et al., 2016). A virtual assistant for the elderly was developed by (Teixeira et al., 2014) and (Lecouteux et al., 2018). The latter investigated the effectiveness of speech human-computer interaction. This system was further augmented by Hämäläinen et al. (2015), adding gesture and touch interfaces. Named as AALfred project, it aimed to facilitate the interactions of the elderly with technologies, allowing them to access the web and some desktop application through voice commands. Then, an automatic answering system for call centres was developed (Oliveira et al., 2012a), providing a better dialogue for clients when compared to traditional keyboard systems. Furthermore, a speech-to-speech translation system (Matsuda et al., 2013) able to transform some speech signals to other languages.

Other applications aimed to improve noise robustness of these systems, such as reverberating signals (Veras et al., 2014), or the Weiner Filter applied by Lima et al. (2015). Both achieving decent results for noise signals. While others tried to classify the speech style (Veiga et al., 2012) based on three classes: Lombard, prepared, and unprepared.

Several automatic transcription systems for TV shows, generally combining MLP with HMM to achieve a 21% Word Error Rate (WER) (Abad et al., 2012a). Similar applications were developed by Álvarez et al. (2016); Kapralova et al. (2014); Lopes et al. (2012b). Those transcriptions allowed an automatic generation of subtitles for distinct media contents, contributing to the digital inclusion of hearing impaired people. This procedure has permitted speech recognition researches to use bigger datasets. Since the data transcription is usually made by hand, making it difficult and costly for large amounts of data. Still dealing with multimedia, a keyword spotting system for audio (Veiga et al., 2014) and video (Eduardo and Eduardo, 2018) were developed to find specific words in speech signals. The use of topic related features in content filters were investigated in (Pereira et al., 2015). Those systems provided a mechanism to automatic summarise large amounts of digital content, based on user preferences.

In medical applications, several neurological issues affect the speech signal somehow (Logemann et al., 1978). Therefore, applications for the diagnose of Alzheimer's disease (Karmele et al., 2015) and Parkinson's disease (Proença et al., 2014) are explored, trying to create a more efficient and less troublesome early diagnose of this clinical state.

This section has presented some applications that made use of ASR solutions to achieve a different goal. Their areas are very distinct, ranging from medical software to automatic translation for multimedia contents. The following section will summarise and introduce the most popular features from 2012 to 2018, for Portuguese language.

6. Feature vectors

On the feature extraction techniques used between 2012 and 2018, there is a decent variety and a consistency when comparing the methods used for the Portuguese language for other languages. This is a crucial step for most ASR systems (Rabiner and Juang, 1993), as it can easily improve the system accuracy and robustness (Veras et al., 2014; Lima et al., 2015).

Through the research, several types of features were identified. Some of them are based on spectral analysis, which includes the use of Fast-Fourier Transform (FFT), filters, and some other transforms. That is the case for Mel-Frequency Cepstrum Coefficients (MFCC) (Mermelstein, 1976), Perceptual Linear Prediction (PLP), and Linear Predictive Codes (LPC). To split the signal, authors generally used 25ms frames with 10ms stride. Then a Hamming Window was applied on each frame, followed by a 512 point FFT. Those are a general spectral feature analysis steps. At this point, each feature can take a different direction, such as using a Mel Filter Bank (for MFCC) or Bark Filter Banks (for LPZ and PLP). Usually, the features vectors have 13 coefficients and can come extended with first and second derivatives, and sometimes with log-energy.

The MFCC in particular is quite trendy, being used for different purposes. It arises not only on speech-to-text, but also for language identification (Suresh and Thorat, 2018), keyword spotting (Veiga et al., 2014), Parkinson's diagnose (Proença et al., 2014), and other speech related research. Furthermore, the popular 13 coefficients are put into testing by Furtado et al. (2013); Silva et al. (2012), concluding that gains over this threshold are not significant.

Another spectral feature is PLP with log-RelAtive SpecTrAl, or PLP-RASTA, an extension of PLP that makes use of Relative Spectral Transform, proposed by (Hermansky, 1990). In our research, this feature was always used alongside Modulation Spectrogram (MSG), introduced by (Greenberg and Kingsbury, 1997). It is one more spectral feature that is closely related to MFCC, as it uses log-filters and the power spectrum. Both MSG and PLPRASTA were used with different purposes, on transcription (Abad et al., 2012a; 2012b; 2013; Álvarez et al., 2016), to estimate age and its impacts on ASR systems (Pellegrini et al., 2012; 2014), and nativeness estimation (Ribeiro et al., 2015).

The Bottleneck Features (BN) is a relatively new approach, proposed by Vu et al. (2012). It is based on MFFC but includes a feed-forward MLP that combines a Linear Discriminant Analysis (LDA) and a covariance transform to create a 42 dimension feature vector. The BN is a trendy feature for multilingual datasets, used by Müller et al. (2016a); Schultz et al. (2013); Shaik et al. (2015); Vesel et al. (2012); Vu et al. (2014), as it composes features from different languages and generate a new vector.

Other features worth mentioning, that are present on published works from 2012 to 2018, are: Low-Rank Matrix Factorisation (LRMF), Spectral-Temporal Receptive Fields (STRF), Zero-Crossing with Peak Amplitudes (ZCPA), Elementary Acoustic Events (EAE), a, extension of BN that uses deep learning called Deep Bottleneck Features (DBF), Fractal Dimensions (FD), and Heteroscedastic Linear Discriminant Analysis (HLDA).

Th is section summarised the features employed between 2012 and 2018. The following section will present the main contributions in Automatic Speech Recognition for Portuguese (ASRPT) from 2012 to 2018. They will be categorised and organised in a better way, following the characteristics of their acoustic models and the way information is shared.

7. Automatic speech recognition approaches for portuguese

The traditional categorisation of ASR approaches, given by Rabiner and Juang (1993) and further used in Ghai and Singh (2012) has three classes: Acoustic-Phonetic (AP), Pattern Recognition (PR), and Artificial Intelligence (AI). However, the field has evolved, and new classification procedures, as well as extraction methods, have been used. For instance, the Discrete Wavelet Packet Decomposition (Hibare and Vibhute, 2014) and Spectral-Temporal Receptive Fields (Wang et al., 2014). As the AI field grows, embracing new definitions and techniques such as PR (Nasrabadi, 2007), previous categorisations become obsolete or ambiguous.

A classification based on the acoustic model was given by Schultz and Kirchhoff (2006) as Monolingual, and Multilingual. The former is defined as a system specialised into one language, while the latter is a system able to recognise many languages. The multilingual models are also subdivided, being them a set of monolingual models (language dependent) or a set of classifier with shared knowledge (language independent).

Therefore, the following sections will present a novel categorisation of the field based on the novel methods for ASRPT. Three main approaches were identified: monolingual, skew-lingual, and cross-lingual. All based on the acoustic model strategy of sharing parameters between its units. Fig. 3 display each strategy. These units can be a perceptron for MLPs or a complete acoustic model for some skew and cross-lingual models.

7.1. Monolingual

The Monolingual classifiers are a classical approach for acoustic models; they consist of training and testing the model with the same language (Schultz and Kirchhoff, 2006). This strategy has some advantages, such as overall better accuracy, more simple compared to Multilingual and Cross-Language, and the only data required is about the target language. The better results are expected, and can be visualised in Table 8, since the training and testing data are in the same domain, helping the technique to create a more accurate model. Most of the acoustic models in this category rely on classical extraction and classification methods, such as MFCC and MLP (Abad et al., 2012b; Bonilla C et al., 2016; Chebotar and Waters, 2016), for which the field has robust and stable software available (as shown in Section 4).

Some works still address the speech recognition on Portuguese languages with this approach, for example in Silva and Serra (2014b,a) the author uses a hybrid model based on a Fuzzy Inference System and Genetic Algorithm trained on a corpus of the Portuguese language. It is also possible to find new methods, such as Convolutional Neural Network (CNN) (Santos et al., 2015), and modifications on Linear Discriminant Analysis (Saeidi et al., 2016) to include noise and uncertainties in observations, therefore improving the system performance on noisy environments. Another work that addresses ASR with this approach is Silva and Barbosa (2015a), which gave the same contribution in Silva and Barbosa (2015b,c); Silva and Batista (2015a, 2015b, 2015c), with an SVM. Furthermore, an investigation of the DNN ability to learn window features, up to 400ms, is given by Vanhoucke et al. (2013). Other researchers that follow this line are (Beserra et al., 2015) with an investigation of the impact of Chaotic Particle Optimisation (CPO) in the Gaussian Mixture Model (GMM) accuracy, and the HMM baseline system built by (Oliveira et al., 2012b). Besides, the identification of vocal ageing (Mendoza et al., 2014) with a combination of Artificial Neural Networks (ANN) and SVM is also classified as Monolingual, as well as the ensemble of ANNs (Ribeiro et al., 2018) trained on a single language dataset. Then, in a different direction is (Silva et al., 2012), which evaluates the efficiency and performance impacts of well-known parameters for ASR.

While the advantages of this approach seem attractive, it also has some drawbacks. The main challenge of developing language technologies for low-resource languages (e.g. Portuguese) is the difficulty to find a suitable amount of data to train the models while preventing problems, such as overfitting (Heigold et al., 2013). Due to this issue, and the fact that monolingual techniques have to be trained on the target language, there is a performance reduction (Besacier et al., 2014) when these models are applied to under-resourced languages.

The Monolingual researches showed two main similarities. First, the necessity to collect data for the target language, given that a description of the data gathering procedure is usually seen in most publications. Besides, it is also common that works on

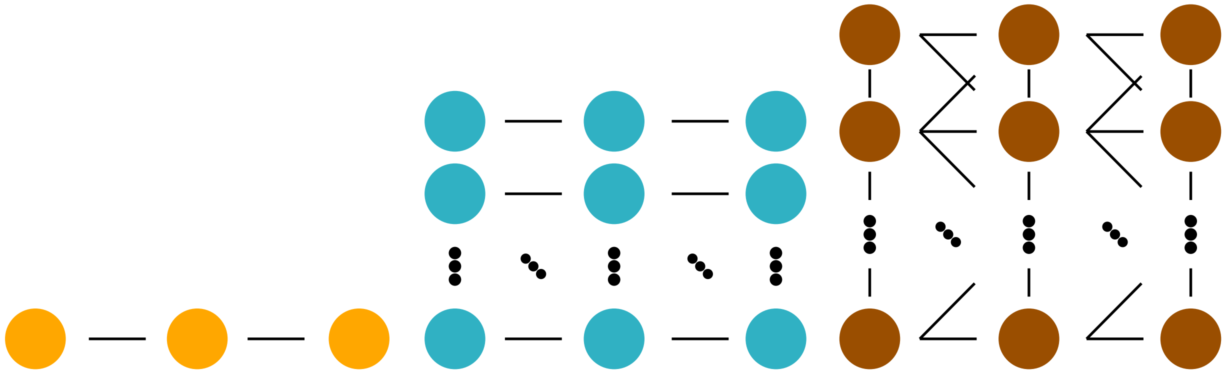

Fig. 3. Parameter sharing strategy for monolingual (yellow), skew-lingual (blue), and cross-lingual (brown) acoustic model. The lines (-, \, and /) represent the parameter connections, and the coloured circles are acoustic model units; such as a perceptron or a complete classifier. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

Table 8
Accuracy of works between 2012 and 2018.

Accuracy	f_i	F_i	Papers
50⊢70	1	1	(Hämäläinen et al., 2014c; Matos et al., 2014; Veras et al., 2014)
70⊢80	7	8	(Hämäläinen et al., 2014d; Lopes et al., 2012b; Lu et al., 2012; Mendoza et al., 2014; Pellegrini et al., 2012; 2013b; Vesel et al., 2012)
80⊢90	10	18	(Abad et al., 2012a; 2012b; Beserra et al., 2015; Chebotar and Waters, 2016; Furtado et al., 2013; Gretter, 2014; Heigold et al., 2013; Kapralova et al., 2014; Lal and King, 2013; Lei et al., 2013; Lu et al., 2014; Mohan and Rose, 2015; Pellegrini et al., 2013a; Swietojanski et al., 2012; Vanhoucke et al., 2013; Vu et al., 2014; Wang et al., 2014)
90⊢100	10	28	(Ghoshal et al., 2013; Hämäläinen et al., 2014c; 2014d; Lee et al., 2016; Oliveira et al., 2012b; 2012a; Santos et al., 2015; Silva and Serra, 2014b; 2014a; Veiga et al., 2014; Vu et al., 2014; Silva et al., 2012)
Total	28	-	-

this line are addressing something else than STT, using it as an intermediate to achieve its goals. This is the case in Santos et al. (2015), aiming to reduce the impact of noise in acoustic models, investigating the effect of biological factors in speech (Pellegrini et al., 2012; 2013b), evaluating the performance of ASR systems for the elderly (Pellegrini et al., 2014), speech dysfunction (Hämäläinen et al., 2014c), the identification of different intonations of Brazilian Portuguese questions (Cantero and Font-Rotchés, 2013), voice-controlled vehicles (Alves et al., 2014), and the prevention of user intrusion for speaker identification systems (Goncalves et al., 2017). Therefore, the Monolingual approach is still useful for several applications for ASR, because their goals differ from just a speech-to-text conversion, and the need for a large amount of data is alleviated since the domain has a limited set of words.

7.2. Skew-lingual

The multilingual approach emerged from the lack of a more generic speech recogniser, enabling the same system to convert speech from *n*-languages for their respective text format. However, this paper proposes a Skew-lingual, following a geometry definition of lines that do not intersect and are not parallel. This association emphasises the fact that each unit of the model does not share knowledge, working independently. The central aspect of this method is that it is composed of specific language acoustic model sets. Since each unit is language dependent, then its quantity defines the number of languages it can recognise.

Because a skew-lingual model has multiple independent classifiers, STT task is usually preceded by a language identification (Schultz and Kirchhoff, 2006). This process will allow the model to choose which classifier can better recognise the current utterance. There are also works focusing on language identification, such as classification of closely related languages and dialects (Gelly et al., 2016), DNN performance for language classification (Gonzalez-Dominguez et al., 2015; Lee et al., 2016; Snyder et al., 2018), including Portuguese, and even GMM models (Suresh and Thorat, 2018). Besides, an application of Fishervoice extraction with SVM was investigated in (Lopez-Otero et al., 2012).

The employment of ensemble networks (Chebotar and Waters, 2016) and SVM (Lopez-Otero et al., 2012) is also another option. The former, in particular, is an excellent example of a skew-lingual acoustic model. The architecture consisted of six networks; each of them classified the inputs as a language unit. After that, an ensemble network decided which of the previous classifiers were more likely to be correct. Therefore, the last network had the same function as a language identification procedure.

However, some issues are inherent to this approach. One of them is scalability, which increased with the number of languages. Even though we have seen substantial success in recent models, we are still not sure about the feasibility of a universal model that can process any language with the same results that we see on language-specific one.

Catastrophic failure is also an aspect that should not be underrated, as it was observed by several authors when the model deals with conditions that could not be modelled by the data or the inherent constraints of models (Torres-Huitzil and Girau, 2017). Given the somewhat random occurrence of such problems, we may see an increasing number of events for skew-lingual models, because this class of models is a more significant challenge than a language-specific one.

The literature does not have a large number of skew-lingual researchers. Due to its flaws and small gains over mono and cross-lingual methods, these techniques are not popular.

This section illustrated the advantages and disadvantages of Skew-lingual models and exemplified its characteristics with applications and researches. Next, an introduction to cross-lingual procedures will be presented.

7.3. Cross-lingual

This approach depends on the assumption that different languages have a typical structure, a mapping function that can correctly associate one to another (Schultz and Kirchhoff, 2006). From that, it is possible to train an acoustic model with a corpus that contains *n*-languages, while the presence of a target language is optional. However, the amount of data for the target language is usually low, and actually, none in some cases (Vesel et al., 2012), as well as (Müller et al., 2016a) for classification on an African idiom. The model was trained on 10 languages, including Portuguese, and further improved in (Müller et al., 2017) with a

Connectionist Temporal Classification (which gave the same contribution (Müller et al., 2016b)). Despite the number of languages that a model can recognise, there are those that combine knowledge between its units but focuses on single language classification, such as (Lei et al., 2013).

Cross-lingual systems are being further investigated, in some cases outperforming the previous classes (Ghoshal et al., 2013). Its ability to recognise a language for which the model was not presented during the training is essential for low-resource languages (Schultz and Kirchhoff, 2006). This method alleviates the need of a large amount of data for language technologies, enabling the development of more sophisticated acoustic models, such as DNNs, and CNN's while preventing common learning problems (Heigold et al., 2013; Shaik et al., 2015).

There are several examples of this technique in the selected papers. For instance, an acoustic model trained with Spanish, Portuguese, and Swiss had a 24% WER for German as target language (Mohan and Rose, 2015). Besides, the literature provided several applications of DNNs. One of them is a study if merging phoneme classes from distinct languages can provide a classification procedure superior to pretained DNNs (Vu et al., 2014). Several state-of-the-art classifiers are in this category, such as SGMM adaptation (Lu et al., 2012; 2014) and Connectionist Temporal Classification (Quintanilha et al., 2017; Tong et al., 2018; Watanabe et al., 2017). Another cross-lingual approach is given in Lu et al. (2012, 2014) with an SGMM adapted with Maximum a posteriori.

Other cross-lingual research with good outcomes is (Lal and King, 2013), where a tandem-feature approach is investigated, motivated by a more straightforward parameter sharing strategy. In a tandem-feature, the acoustic model units are lined up one behind the other and facing the same direction (English Dictionary, 2018). This way, the knowledge of source languages are mixed in an attempt to classify the target language.

This class of models are recent, and less research is available for it. Even though cross-lingual results and applications, they are promising. In fact, for low-resource languages, these models are critical. This necessity is due to the ability to recognise languages that were not present in training. Here, an introduction for cross-lingual acoustic models was given with some clarification based on its recent publications.

Next, a general overview of field directions and gaps and the authors conclusions will be presented.

8. Research directions and gaps in Portuguese automatic speech recognition

This section aims to present suggestions about the field directions and gaps based on the essential works selected in this study and the information collected about them. In Section 4, we added information about the popularity of datasets in the period. Thus, following this reasoning, Fig. 4 presents the same idea for classification methods employed on any ASR sub-module, while Fig. 5 presents the extraction methods. In the next sections, an overview of future research on classification and extraction methods for ASR is presented.

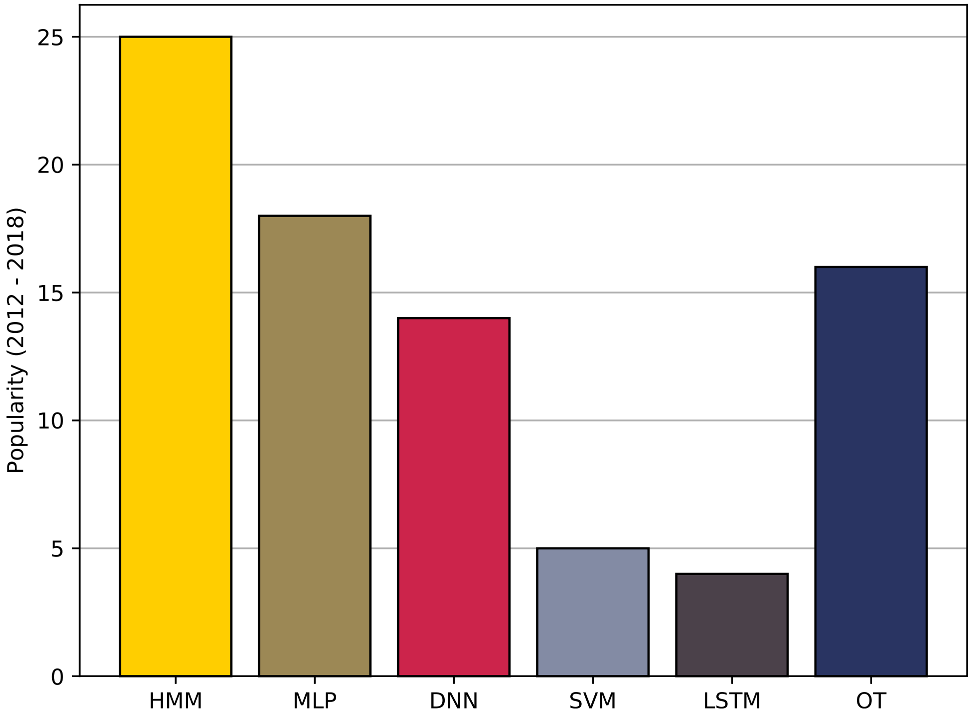

Fig. 4. Popularity of classifiers between 2012 and 2018. Support Vector Machine (SVM); Hidden Markov Model (HMM); Multilayer Perceptron (MLP); Restricted Boltzmann Machine (RBM); Deep Neural Network (DNN); Sum of classifiers with one usage (OT).

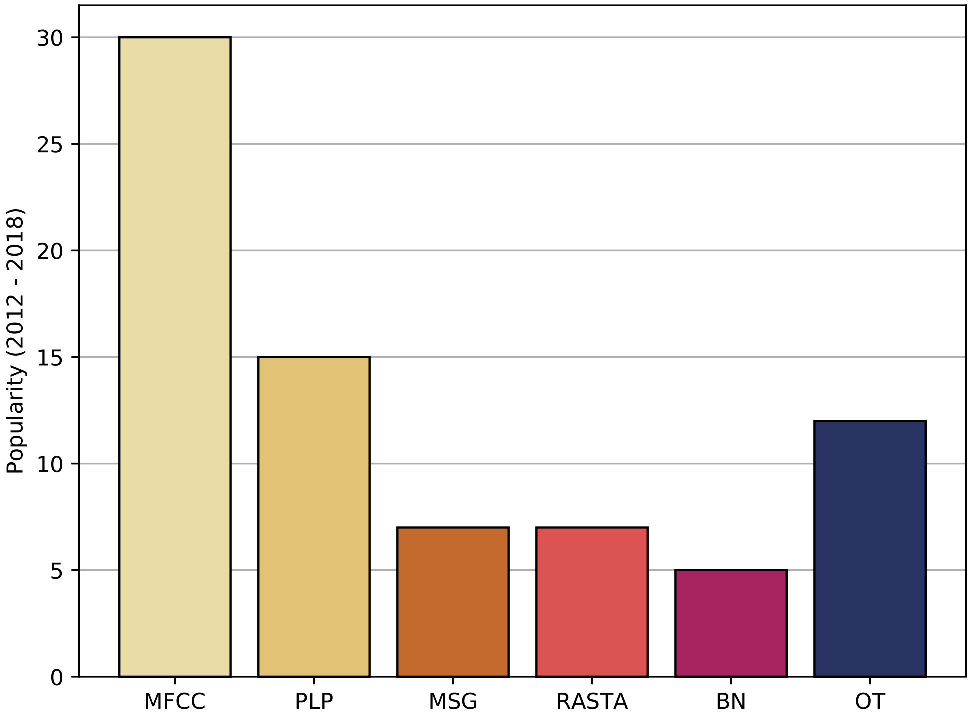

Fig. 5. The popularity of feature extraction techniques from 2012 to 2018. Mel-Frequency Cepstrum Coefficients (MFCC); Perceptual Linear Prediction (PLP); Modulation SpectroGram (MSG); log-RelAtive SpecTrAl (RASTA); (SMILE); Sum of other extraction methods used (OT).

8.1. *Classification methods*

The most recent classification procedures were, on its totality, artificial intelligence methods. The chart in Fig. 4 present the HMM with the highest popularity. Its presence in the area, for an extended period, was usually high (Besacier et al., 2014). However, most of it is now overlapped because this model is commonly used in conjunction with MLPs.

Very few papers use a pure HMM (Hämäläinen et al., 2014c; Oliveira et al., 2012b; Veiga et al., 2012), since mixtures of ANN and HMMs is providing better results (Abad et al., 2012a; Ghoshal et al., 2013; Mohan and Rose, 2015; Pellegrini et al., 2013b). This combination is also the fact for Restricted Boltzmann Machines (RBM), which is mostly used for DNN parameters initialisation, such as pretraining (Mohan and Rose, 2015; Swietojanski et al., 2012).

The popularity of DNN is close to MLPs. This pattern shows growth in deep learning approaches for ASR. It is also notable that DNN popularity is still smaller than other techniques. Since this approach requires a large amount of data, it is normal that low resource languages do not use it very often. While most of the resources for Portuguese remain with restricted access, the techniques that do not rely on large amounts of data (SVN, MLP) will increase in popularity.

However, some gaps can still be spotted. The AI field has several methods for classification. Although, for Portuguese and its variations, the diversity of techniques is not notable. A total of eight (SVM, HMM, MLP, SGMM, DNN, ANFIS-3, CPSO, and CNN) classifiers were used from 2012 to 2018. Even though these methods already presented good results for English variations, it is known that distinct languages can give different structures, phonemes, and features (Schultz and Kirchhoff, 2006). Therefore, it is important to try other methods, such as ANFIS variations or CNN, or even classifiers, that failed to provide good outcomes for well-resourced languages.

One more flaw identified is the absence of clustering techniques and more combinations of classifiers with evolutionary algorithms. The clustering methods are usually applied on vector quantisation for ASR (Rabiner and Juang, 1993). While they had no presence at all between 2012 and 2018, some evolutionary algorithms have been highlighted (Silva et al., 2016; Beserra et al., 2015; Silva and Serra, 2014b; 2014a). While evolutionary techniques are investigated in both (Silva and Serra, 2014b) and (Beserra et al., 2015). The first paper show results of replacing the Least Square Estimation by a Genetic Algorithm on the learning procedure of an ANFIS. The second introduces a combination of GMMs and a CPO, compare it with a pure GMM, and the authors conclude that CPO has improved over GMM. Therefore, these hybridisations are promising, but they require a more in-depth investigation, given their small number of researches.

Finally, Table 8 shows the accuracy of the selected papers. Note that a total of 28 papers are present on the table; this is because applications and dataset creation were excluded from it. Table 8 shows that most of the works have accuracy above 80%, while only one is below 70%. These results represent some maturity on the field, with decent outcomes even though it is under-resourced. Also, a hybrid method (Santos et al., 2015) is among the most accurate classifiers, reinforcing that these methods are promising.

8.2. Features

This section will present the most popular extraction methods for ASR from 2012 to 2018, besides an overview of researches on this topic. Thus, in Fig. 5, the chart represents the popularity in the field for these methods.

The MFCC is by far the most popular, along with PLP and it is likely to keep that way. Both of them have presented a more generic aspect, being employed on distinct tasks other than STT. Usually, they can provide good results on other applications rather than STT, such as language recognition (Müller et al., 2017) and age estimation (Pellegrini et al., 2014). In particular, MFCC present a general better accuracy, as well as simplicity over other methods (Furtado et al., 2013).

However, it is important to carefully determine the best features for a given problem. As other features, explored between 2012 and 2018, also have an influence. During this period, a total of 16 features were explored. Compared to the number of distinct classifiers, the features had a more in depth exploration, being used in distinct applications and approaches.

Finally, the field is very diverse with respect to features. However, the availability of dataset may have restrict the use of some methods. Next section will describe the authors conclusions.

9. Conclusion

This paper presented a review of ASRPT from the period between 2012−2018. The information about classifiers, extraction methods, and datasets can be used to start new researches for that language. The gathered information from papers made it possible to see the area tendency and gaps for every Portuguese ASR component. Besides, they can be used to investigate the field further and explore new methods.

The researches of speech recognition for Portuguese language and its variations has shown an expected small number of investigations. Although, their accuracy falls behind the researches for more popular languages. This decay, as shown in this review, because of the lack of data for developing language technologies. So, even though some papers aim to collect speech data, most of them do not publish it freely. Such hidden information difficulted the use of more sophisticated techniques, such as DNNs and CNN’s, which would require a large amount of data. Furthermore, the studies on this subject have shown a diversity of reasons to use ASR, while at the same time did not explore many classification techniques. Therefore, this area of research still has room to exploration on several components of ASR.

Finally, this paper may also be used to introduce automatic speech recognition for anyone with little information about it, especially for scientists interested in ASR.

Acknowledgments

We would like to acknowledge the short contribution of Eduardo Coelho de Lima in the early stages of this work. This work was supported by the Brazilian Coordination for the Improvement of Higher Education Personnel (CAPES).

References

Abad, A., Astudillo, R.F., Trancoso, I., 2012a. The L2F spoken web search system for Mediaeval. In: Proceedings of the MediaEval. Citeseer.

Abad, A., Meinedo, H., Trancoso, I., Neto, J., 2012b. Transcription of Multi-variety Portuguese Media Contents. In: Caseli, H., Villavicencio, A., Teixeira, A., Perdigão, F. (Eds.), Computational Processing of the Portuguese Language. Springer Berlin Heidelberg, Berlin, Heidelberg, pp. 409–420.

Abad, A., Pompili, A., Costa, A., Trancoso, I., Fonseca, J., Leal, G., Farrajota, L., Martins, I.P., 2013. Automatic word naming recognition for an on-line aphasia treatment system. Comput. Speech Lang. 27 (6), 1235–1248. https://doi.org/10.1016/j.csl.2012.10.003. Special Issue on Speech and Language Processing for Assistive Technology.

Akinobu, L., 1997. JULIUS Website. http://julius.osdn.jp/en_index.php/. Accessed: 2018-11-30.

Álvarez, A., Mendes, C., Raffaelli, M., Luís, T., Paulo, S., Piccinini, N., Arzelus, H., Neto, J., Aliprandi, C., del Pozo, A., 2016. Automating live and batch subtitling of multimedia contents for several European languages. Multimed. Tools Appl. 75 (18), 10823–10853. https://doi.org/10.1007/s11042-015-2794-z.

Alves, S.F.R., Silva, I.N., Ranieri, C.M., Filho, H.F., 2014. Assisted robot navigation based on speech recognition and synthesis. In: Proceedings of the 5th ISSNIP-IEEE Biosignals and Biorobotics Conference: Biosignals and Robotics for Better and Safer Living (BRC), pp. 1–5. https://doi.org/10.1109/BRC.2014.6881003.

Arel, I., Rose, D.C., Karnowski, T.P., 2010. Deep machine learning - a new frontier in artificial intelligence research [Research Frontier]. IEEE Comput. Intell. Mag. 5 (4), 13–18. https://doi.org/10.1109/MCI.2010.938364.

Batista, C., Dias, A.L., Neto, N.S., 2018. Baseline acoustic models for Brazilian Portuguese using Kaldi tools. In: Proceedings of the IberSPEECH. ISCA, pp. 77–81. https://doi.org/10.21437/iberspeech.2018-17.

Besacier, L., Barnard, E., Karpov, A., Schultz, T., 2014. Automatic speech recognition for under-resourced languages: a survey. Speech Commun. 56, 85–100. https://doi.org/10.1016/j.specom.2013.07.008.

Beserra, A.A.V., Silva, W.L.S., d. O. Serra, G.L., 2015. A GMM/CPSO speech recognition system. In: Proceedings of the IEEE 24th International Symposium on Industrial Electronics (ISIE), pp. 26–31. https://doi.org/10.1109/ISIE.2015.7281438.

Boersma, P., Weenink, D., 1991. PRAAT Website. http://www.fon.hum.uva.nl/praat/ Accessed: 2018-11-30.

Bonilla C, D.A., Nedjah, N., de Macedo Mourelle, L., 2016. Online pattern recognition for portuguese phonemes using multi-layer perceptron combined with recurrent non-linear autoregressive neural networks with exogenous inputs. In: Proceedings of the IEEE Latin American Conference on Computational Intelligence (LA-CCI), pp. 1–6. https://doi.org/10.1109/LA-CCI.2016.7885705.

Brümmer, N., De Villiers, E., 2013. The bosaris toolkit: theory, algorithms and code for surviving the new dcf. arXiv preprint arXiv:1304.2865.

Cambridge, U., 1993. HTK Speech Recognition Toolkit. http://htk.eng.cam.ac.uk/ Accessed: 2018-11-30.

Candeias, S., Celorico, D., Proença, J., Veiga, A., Perdigão, F., 2013. HESITA (tions) in Portuguese: a database. In: Proceedings of the Sixth Workshop on Disfluency in Spontaneous Speech.

Cantero, F.J., Font-Rotchés, D., 2013. The intonation of absolute questions of Brazilian portuguese. Linguist. Lit. Stud. 1 (3), 142–149. https://doi.org/10.13189/lls.2013.010302.

Chebotar, Y., Waters, A., 2016. Distilling knowledge from ensembles of neural networks for speech recognition.. In: Proceedings of the Interspeech, pp. 3439–3443.

Cox, T., Davies, R.S., 2012. Using automatic speech recognition technology with elicited oral response testing. Calico J. 29 (4), 601–618.

Cristoforetti, L., Ravanelli, M., Omologo, M., Sosi, A., Abad, A., Hagmüller, M., Maragos, P., 2014. The DIRHA simulated corpus.. In: Proceedings of the LREC, pp. 2629–2634.

Deoldoto Paulino, M.A., Maldonado e Gomes da Costa, Y., Souza Britto Junior, A., Renan Svaigen, A., Ruiz Aylon, L.B., Soares de Oliveira, L.E., 2018. A brazilian speech database. In: Proceedings of the IEEE 30th International Conference on Tools with Artificial Intelligence (ICTAI), pp. 234–241. https://doi.org/10.1109/ICTAI.2018.00044.

English Dictionary, O., 2018. Tandem. Oxford University Press Accessed 29 October 2018.

Eberhard, D. M., Simons, G. F., Fennig, C. D., 2018. Ethnologue: Languages of the World. Dallas, Texas: SIL International.

Eduardo, S., Eduardo, B., 2018. A framework for automatic topic segmentation in video lectures. Anais Estendidos do Simpsio Brasileiro de Sistemas Multimdia e Web (WebMedia) 31–36.

Furtado, L., Miranda, B., Neto, N., Meiguins, B., 2015. Ivorpheus - a proposal for interaction by voice commands in three-dimensional environments of information visualization. In: Proceedings of the IEEE International Conference on Computer and Information Technology; Ubiquitous Computing and Communications; Dependable, Autonomic and Secure Computing; Pervasive Intelligence and Computing, pp. 878–883. https://doi.org/10.1109/CIT/IUCC/DASC/PICOM.2015.131.

Furtado, S., Vinícius, M.A.S., de Gustavo, E.A.P.A.B., 2013. A comparative study between MFCC and LSF coefficients in automatic recognition of isolated digits pronounced in Portuguese and English. Acta Sci. Technol. 35 (4). https://doi.org/10.4025/actascitechnol.v35i4.19825.

Gelly, G., Gauvain, J.-L., Lamel, L., Laurent, A., Le, V. B., Messaoudi, A., 2016. Language recognition for dialects and closely related languages. Odyssey, Bilbao, Spain.

Ghai, W., Singh, N., 2012. Literature review on automatic speech recognition. Int. J. Comput. Appl. 41 (8), 42–50.

Ghoshal, A., Swietojanski, P., Renals, S., 2013. Multilingual training of deep neural networks. In: Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing, pp. 7319–7323. https://doi.org/10.1109/ICASSP.2013.6639084.

Goncalves, A.R., Violato, R.P.V., Korshunov, P., Marcel, S., Simoes, F.O., 2017. On the generalization of fused systems in voice presentation attack detection. In: Proceedings of the International Conference of the Biometrics Special Interest Group (BIOSIG), pp. 1–5. https://doi.org/10.23919/BIOSIG.2017.8053516.

Gonzalez-Dominguez, J., Lopez-Moreno, I., Moreno, P.J., Gonzalez-Rodriguez, J., 2015. Frame-by-frame language identification in short utterances using deep neural networks. Neural Netw. 64, 49–58. https://doi.org/10.1016/j.neunet.2014.08.006. Special Issue on Deep Learning of Representations.

Greenberg, S., Kingsbury, B.E.D., 1997. The modulation spectrogram: in pursuit of an invariant representation of speech. In: Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing, 3, pp. 1647–1650. https://doi.org/10.1109/ICASSP.1997.598826.

Gretter, R., 2014. Euronews: a multilingual benchmark for ASR and LID. In: Proceedings of the Fifteenth Annual Conference of the International Speech Communication Association.

Hämäläinen, A., Avelar, J., Rodrigues, S., Dias, M.S., Kolesinski, A., Fegyó, T., Németh, G., Csobánka, P., Lan, K., Hewson, D., 2014a. The EASR Corpora of European Portuguese, French, Hungarian and polish elderly speech. In: Proceedings of the LREC, pp. 1458–1464.

Hämäläinen, A., Candeias, S., Cho, H., Meinedo, H., Abad, A., Pellegrini, T., Tjalve, M., Trancoso, I., Sales Dias, M., 2014b. Correlating ASR errors with developmental changes in speech production: a study of 3-10-year-old European Portuguese children's speech. In: Proceedings of the Workshop on Child Computer Interaction - WOCCI 2014, pp. 1–7. Singapore, Singapore.

Hämäläinen, A., Cho, H., Candeias, S., Pellegrini, T., Abad, A., Tjalve, M., Trancoso, I., Dias, M.S., 2014c. Automatically recognising European Portuguese children's speech. In: Baptista, J., Mamede, N., Candeias, S., Paraboni, I., Pardo, T.A.S., Volpe Nunes, M.d. G. (Eds.), Proceedings of the Computational Processing of the Portuguese Language. Springer International Publishing, Cham, pp. 1–11.

Hämäläinen, A., Meinedo, H., Tjalve, M., Pellegrini, T., Trancoso, I., Dias, M.S., 2014d. Improving speech recognition through automatic selection of age group – specific acoustic models. In: Baptista, J., Mamede, N., Candeias, S., Paraboni, I., Pardo, T.A.S., Volpe Nunes, M.d. G. (Eds.), Proceedings of the Computational Processing of the Portuguese Language. Springer International Publishing, Cham, pp. 12–23.

Hämäläinen, A., Pinto, F., Dias, M., Júdice, A., Freitas, J., Pires, C., Teixeira, V., Calado, A., Braga, D., 2012. The first european portuguese elderly speech corpus. In: Proceedings of the IberSPEECH, 10.

Hämäläinen, A., Rodrigues, S., Júdice, A., Silva, S.M., Calado, A., Pinto, F.M., Dias, M.S., 2013. The CNG Corpus of European Portuguese Children's Speech. In: Habernal, I., Matoušek, V. (Eds.), Proceedings of the Text, Speech, and Dialogue. Springer Berlin Heidelberg, Berlin, Heidelberg, pp. 544–551.

Hämäläinen, A., Teixeira, A., Almeida, N., Meinedo, H., Fegyó, T., Dias, M.S., 2015. Multilingual speech recognition for the elderly: the aalfred personal life assistant. Procedia Comput. Sci. 67, 283–292. https://doi.org/10.1016/j.procs.2015.09.272. Proceedings of the 6th International Conference on Software Development and Technologies for Enhancing Accessibility and Fighting Info-exclusion.

Heigold, G., Vanhoucke, V., Senior, A., Nguyen, P., Ranzato, M., Devin, M., Dean, J., 2013. Multilingual acoustic models using distributed deep neural networks. In: Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing, pp. 8619–8623. https://doi.org/10.1109/ICASSP.2013.6639348.

Hermansky, H., 1990. Perceptual linear predictive (PLP) analysis of speech. J. Acoust. Soc. Am. 87 (4), 1738–1752. https://doi.org/10.1121/1.399423.

Hibare, R., Vibhute, A., 2014. Feature extraction techniques in speech processing: a survey. Int. J. Comput. Appl. 107 (5), 1–8.

Kapralova, O., Alex, J., Weinstein, E., Moreno, P.J., Siohan, O., 2014. A big data approach to acoustic model training corpus selection. In: Proceedings of the Fifteenth Annual Conference of the International Speech Communication Association.

Karmele, L.-d.-I., Solé-Casals, J., Eguiraun, H., Alonso, J.B., Travieso, C.M., Aitzol, E., Nora, B., Miriam, E.-T., Pablo, M.-L., Blanca, B., 2015. Feature selection for spontaneous speech analysis to aid in alzheimer's disease diagnosis: a fractal dimension approach. Comput. Speech Lang. 30 (1), 43–60. https://doi.org/10.1016/j.csl.2014.08.002.

Lal, P., King, S., 2013. Cross-lingual automatic speech recognition using tandem features. IEEE Audio Speech Lang. Process. 21 (12), 2506–2515. https://doi.org/10.1109/TASL.2013.2277932.

Lecouteux, B., Vacher, M., Portet, F., 2018. Distant speech processing for smart home: comparison of ASR approaches in scattered microphone network for voice command. Int. J. Speech Technol. 21 (3), 601–618. https://doi.org/10.1007/s10772-018-9520-y.

Lee, K., Li, H., Deng, L., Hautamäki, V., Rao, W., Xiao, X., Larcher, A., Sun, H., Nguyen, T., Wang, G., et al., 2016. The 2015 NIST language recognition evaluation: the shared view of L2R, Fantastic4 and SingaMS. In: Proceedings of the Interspeech, 2016, pp. 3211–3215.

Lei, X., Lin, H., Heigold, G., 2013. Deep neural networks with auxiliary Gaussian mixture models for real-time speech recognition. In: Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing, pp. 7634–7638. https://doi.org/10.1109/ICASSP.2013.6639148.

Lima, Í.A., Alencar, M.S., Lopes, W.T.A., Madeiro, F., 2015. Evaluation of optimal and sub-optimal speech noise reduction wiener filters. In: Proceedings of the International Workshop on Telecommunications (IWT), pp. 1–5. https://doi.org/10.1109/IWT.2015.7224581.

Logemann, J.A., Fisher, H.B., Boshes, B., Blonsky, E.R., 1978. Frequency and cooccurrence of vocal tract dysfunctions in the speech of a large sample of parkinson patients. J. Speech Hear. Disord. 43 (1), 47–57. https://doi.org/10.1044/jshd.4301.47.

Lopes, C., Veiga, A., Perdigão, F., 2012. A European Portuguese children speech database for computer aided speech therapy. In: Caseli, H. (Ed.), Proceedings of the PROPOR. Springer, pp. 368–374.

Lopes, J., Eskenazi, M., Trancoso, I., 2012. Incorporating asr information in spoken dialog system confidence score. In: Caseli, H., Villavicencio, A., Teixeira, A., Perdigão, F. (Eds.), Proceedings of the Computational Processing of the Portuguese Language. Springer Berlin Heidelberg, Berlin, Heidelberg, pp. 403–408.

Lopez-Otero, P., Docio-Fernandez, L., Garcia-Mateo, C., 2012. A Fishervoice-SVM language identification system. In: Caseli, H., Villavicencio, A., Teixeira, A., Perdigão, F. (Eds.), Proceedings of the Computational Processing of the Portuguese Language. Springer Berlin Heidelberg, Berlin, Heidelberg, pp. 381–391.

Lu, L., Ghoshal, A., Renals, S., 2012. Maximum a posteriori adaptation of subspace Gaussian mixture models for cross-lingual speech recognition. In: Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 4877–4880. https://doi.org/10.1109/ICASSP.2012.6289012.

Lu, L., Ghoshal, A., Renals, S., 2014. Cross-lingual subspace gaussian mixture models for low-resource speech recognition. IEEE/ACM Trans. Audio, Speech and Lang. Proc. 22 (1), 17–27. https://doi.org/10.1109/TASL.2013.2281575.

Lui, C. E., Blum, J., Parks, M. J., Paulson, K. P., 2004. Method and system for embedding voice notes. URL: https://voicenote.in/US Patent 6,720,980.

Matos, M., Abad, A., Astudillo, R., Trancoso, I., 2014. Recognition of distant voice commands for home applications in Portuguese. In: Navarro Mesa, J.L., Ortega, A., Teixeira, A., Hernndez Prez, E., Quintana Morales, P., Ravelo Garca, A., Guerra Moreno, I., Toledano, D.T. (Eds.), Proceedings of the Advances in Speech and Language Technologies for Iberian Languages. Springer International Publishing, pp. 178–188.

Matsuda, S., Hu, X., Shiga, Y., Kashioka, H., Hori, C., Yasuda, K., Okuma, H., Uchiyama, M., Sumita, E., Kawai, H., Nakamura, S., 2013. Multilingual Speech-to-Speech Translation System: VoiceTra. In: Proceedings of the IEEE 14th International Conference on Mobile Data Management, 2, pp. 229–233. https://doi.org/10.1109/MDM.2013.99.

Mendoza, L.A.F., Cataldo, E., Vellasco, M.M.B.R., Silva, M.A., Apolinrio, J.A., 2014. Classification of vocal aging using parameters extracted from the glottal signal. J. Voice 28 (5), 532–537. https://doi.org/10.1016/j.jvoice.2014.02.001.

Mermelstein, P., 1976. Distance measures for speech recognition, psychological and instrumental. Pattern Recognit. Artif. Intell. 116, 374–388.

Mohan, A., Rose, R., 2015. Multi-lingual speech recognition with low-rank multi-task deep neural networks. In: Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 4994–4998. https://doi.org/10.1109/ICASSP.2015.7178921.

Müller, M., Stüker, S., Waibel, A., 2016. Language adaptive DNNS for improved low resource speech recognition.. In: Proceedings of the INTERSPEECH, pp. 3878–3882.

Müller, M., Stüker, S., Waibel, A., 2016. Language feature vectors for resource constraint speech recognition.. In: Proceedings of the ITG Symposium on Speech Communication, pp. 1–5.

Müller, M., Stüker, S., Waibel, A., 2017. Language Adaptive Multilingual CTC Speech Recognition. In: Karpov, A., Potapova, R., Mporas, I. (Eds.), Proceedings of the Speech and Computer. Springer International Publishing, pp. 473–482.

Muslim, E.M., 2007. An introduction to computational linguistics advantages & disadvantages. J. Coll. Basic Educ. 10 (51), 29–40.

Nasrabadi, N.M., 2007. Pattern recognition and machine learning. J. Electron Imaging 16 (4), 49–90.

Neto, J.P., Martins, C.A., Meinedo, H., Almeida, L.B., 1997. The design of a large vocabulary speech corpus for Portuguese. In: Proceedings of the Fifth European Conference on Speech Communication and Technology.

Neto, N., Patrick, C., Klautau, A., Trancoso, I., 2011. Free tools and resources for brazilian portuguese speech recognition. J. Braz. Comput. Soc. 17 (1), 53–68. https://doi.org/10.1007/s13173-010-0023-1.

Oliveira, A.L.C., Silva, E.S., Macedo, H.T., Matos, L.N., 2012. Brazilian Portuguese speech-driven answering system. In: Proceedings of the 6th Euro American Conference on Telematics and Information Systems. ACM, New York, NY, USA, pp. 277–284. https://doi.org/10.1145/2261605.2261647.

Oliveira, R., Batista, P., Neto, N., Klautau, A., 2012. Baseline acoustic models for Brazilian Portuguese using CMU sphinx tools. In: Caseli, H., Villavicencio, A., Teixeira, A., Perdigão, F. (Eds.), Proceedings of the Computational Processing of the Portuguese Language. Springer Berlin Heidelberg, Berlin, Heidelberg, pp. 375–380.

Pellegrini, T., Correia, R., Trancoso, I., Baptista, J., Mamede, N., Eskenazi, M., 2013. ASR-Based exercises for listening comprehension practice in european portuguese. Comput. Speech Lang. 27 (5), 1127–1142. https://doi.org/10.1016/j.csl.2013.02.004.

Pellegrini, T., Hämäläinen, A., de Mareüil, P.B., Tjalve, M., Trancoso, I., Candeias, S., Dias, M.S., Braga, D., 2013. A corpus-based study of elderly and young speakers of European Portuguese: acoustic correlates and their impact on speech recognition performance.. In: Proceedings of the INTERSPEECH, pp. 852–856.

Pellegrini, T., Hedayati, V., Trancoso, I., Hämäläinen, A., Dias, M.S., 2014. Speaker age estimation for elderly speech recognition in European Portuguese. In: Proceedings of the Fifteenth Annual Conference of the International Speech Communication Association.

Pellegrini, T., Trancoso, I., Hämäläinen, A., Calado, A., Dias, M.S., Braga, D., 2012. Impact of Age in ASR for the elderly: preliminary experiments in European Portuguese. In: Torre Toledano, D., Ortega Giménez, A., Teixeira, A., González Rodríguez, J., Hernández Gómez, L., San Segundo Hernández, R., Ramos Castro, D. (Eds.), Proceedings of the Advances in Speech and Language Technologies for Iberian Languages. Springer Berlin Heidelberg, Berlin, Heidelberg, pp. 139–147.

Pereira, M.H.R., de Souza, C.L., Pdua, F.L.C., Silva, G.D., de Assis, G.T., Pereira, A.C.M., 2015. SAPTE: a multimedia information system to support the discourse analysis and information retrieval of television programs. Multimedia Tools Appl. 74 (23), 10923–10963. https://doi.org/10.1007/s11042-014-2311-9.

Povey, D., Ghoshal, A., Boulianne, G., Burget, L., Glembek, O., Goel, N., Hannemann, M., Motlicek, P., Qian, Y., Schwarz, P., Silovsky, J., Stemmer, G., Vesely, K., 2011. The Kaldi speech recognition toolkit. In: Proceedings of the IEEE Workshop on Automatic Speech Recognition and Understanding. IEEE Signal Processing Society. IEEE Catalog No.: CFP11SRW-USB.

Povey, D., Goelb, N., Burgetc, L., Agarwald, M., Akyazie, P., Kaif, F., Ghoshalg, A., Glembekc, O., Karafiátc, M., Rastrowh, A., et al., 2009. Low development cost, high quality speech recognition for new languages and domains: report from 2009 Johns Hopkins/CLSP Summer Workshop. Technical Report. Johns Hopkins University.

Proença, J., Celorico, D., Candeias, S., Lopes, C., Perdigão, F., 2015. Children's reading aloud performance: a database and automatic detection of disfluencies. In: Proceedings of the Sixteenth Annual Conference of the International Speech Communication Association.

Proença, J., Celorico, D., Lopes, C., Dias, M.S., Tjalve, M., Stolcke, A., Candeias, S., Perdigão, F., 2016. Design and analysis of a database to evaluate children's reading aloud performance. In: Silva, J., Ribeiro, R., Quaresma, P., Adami, A., Branco, A. (Eds.), Proceedings of the Computational Processing of the Portuguese Language. Springer International Publishing, Cham, pp. 385–395.

Proença, J., Lopes, C., Tjalve, M., Stolcke, A., Candeias, S., Perdigão, F., 2017. Detection of mispronunciations and disfluencies in children reading aloud. In: Proceedings of the INTERSPEECH, pp. 1437–1441.

Proença, J., Veiga, A., Candeias, S., Lemos, J., Januário, C., Perdigão, F., 2014. Characterizing Parkinson's disease speech by acoustic and phonetic features. In: Baptista, J., Mamede, N., Candeias, S., Paraboni, I., Pardo, T.A.S., Volpe Nunes, M.d. G. (Eds.), Proceedings of the Computational Processing of the Portuguese Language. Springer International Publishing, Cham, pp. 24–35.

Quintanilha, I.M., Biscainho, L.W.P., Netto, S.L., 2017. Towards an end-to-end speech recognizer for portuguese using deep neural networks. In: Proceedings of the XXXV Simpósio Brasileiro de Telecomunicaç oes e Processamento de Sinais, Sao Pedro, Brazil.

Rabiner, L., Juang, B.-H., 1993. Fundamentals of Speech Recognition. Prentice-Hall, Inc., Upper Saddle River, NJ, USA.

Raso, T., Mello, H., 2012. The C-ORAL-BRASIL I: reference corpus for informal spoken Brazilian Portuguese. In: Caseli, H., Villavicencio, A., Teixeira, A., Perdigão, F. (Eds.), Proceedings of the Computational Processing of the Portuguese Language. Springer Berlin Heidelberg, Berlin, Heidelberg, pp. 362–367.

Ravanelli, M., Cristoforetti, L., Gretter, R., Pellin, M., Sosi, A., Omologo, M., 2015. The dirha-english corpus and related tasks for distant-speech recognition in domestic environments. In: Proceedings of the IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU), pp. 275–282. https://doi.org/10.1109/ASRU.2015.7404805.

Ribeiro, E., Ferreira, J., Olcoz, J., Abad, A., Moniz, H., Batista, F., Trancoso, I., 2015. Combining multiple approaches to predict the degree of nativeness. In: Proceedings of the Interspeech.

Ribeiro, F.C., Carvalho, R.T.S., Cortez, P.C., De Albuquerque, V.H.C., Filho, P.P.R., 2018. Binary neural networks for classification of voice commands from throat microphone. IEEE Access 6, 70130–70144. https://doi.org/10.1109/ACCESS.2018.2881199.

Robin, 2018. What is corpus?World of Computing Accessed 29 October 2018.

Rocha, R.S., Ferreira, P., Dutra, I., Correia, R., Salvini, R., Burnside, E., 2016. A speech-to-text interface for mammoclass. In: Proceedings of the IEEE 29th International Symposium on Computer-Based Medical Systems (CBMS), pp. 1–6. https://doi.org/10.1109/CBMS.2016.25.

Rocha, T., Marques, A., Brito, J.P., Cardoso, L., Martins, P., Barroso, J., 2017. Web application for the training of the correct pronunciation of words in portuguese for people with speech and language disorders preliminary usability study. In: Proceedings of the 12th Iberian Conference on Information Systems and Technologies (CISTI), pp. 1–7. https://doi.org/10.23919/CISTI.2017.7975889.

Rodríguez-Fuentes, L.J., Penagarikano, M., Varona, A., Diez, M., Bordel, G., 2016. Kalaka-3: a database for the assessment of spoken language recognition technology on youtube audios. Lang. Resour. Eval. 50 (2), 221–243. https://doi.org/10.1007/s10579-015-9324-5.

Ruesch, J., Bateson, G., Pinsker, E.C., Combs, G., 2017. Communication: The Social Matrix of Psychiatry. Routledge.

Saeidi, R., Astudillo, R.F., Kolossa, D., 2016. Uncertain lda: including observation uncertainties in discriminative transforms. IEEE Trans. Pattern Anal. Mach. Intell. 38 (7), 1479–1488. https://doi.org/10.1109/TPAMI.2015.2481420.

Santos, R.M., Matos, L.N., Macedo, H.T., Montalvão, J., 2015. Speech recognition in noisy environments with convolutional neural networks. In: Proceedings of the Brazilian Conference on Intelligent Systems (BRACIS), pp. 175–179. https://doi.org/10.1109/BRACIS.2015.44.

Schultz, T., 2002. GlobalPhone: a multilingual speech and text database developed at Karlsruhe University. In: Proceedings of the Seventh International Conference on Spoken Language Processing.

Schultz, T., Kirchhoff, K., 2006. Multilingual Speech Processing. Elsevier.

Schultz, T., Schlippe, T., 2014. GlobalPhone: Pronunciation Dictionaries in 20 Languages. In: Proceedings of the LREC, pp. 337–341.

Schultz, T., Vu, N.T., Schlippe, T., 2013. GlobalPhone: A multilingual text amp; speech database in 20 languages. In: Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing, pp. 8126–8130. https://doi.org/10.1109/ICASSP.2013.6639248.

Shaik, M.A.B., Tüske, Z., Tahir, M.A., Nußbaum-Thom, M., Schlüter, R., Ney, H., 2015. Improvements in RWTH LVCSR evaluation systems for polish, portuguese, english, urdu, and arabic. In: Proceedings of the Sixteenth Annual Conference of the International Speech Communication Association.

Silva, D.D.C., Vasconcelos, C.R., Neto, B.G.A., Fechine, J.M., 2012. Evaluation of the impact in reducing the number of parameters for continuous speech recognition for Brazilian Portuguese. In: Proceedings of the ISSNIP Biosignals and Biorobotics Conference: Biosignals and Robotics for Better and Safer Living (BRC), pp. 1–6. https://doi.org/10.1109/BRC.2012.6222182.

Silva, W., Serra, G., 2014a. Intelligent genetic fuzzy inference system for speech recognition: an approach from low order feature based on discrete cosine transform. J. Control, Autom. Electr. Syst. 25 (6), 689–698. https://doi.org/10.1007/s40313-014-0148-0.

Silva, W., Serra, G., 2014b. A novel intelligent system for speech recognition. In: Proceedings of the International Joint Conference on Neural Networks (IJCNN), pp. 3599–3604. https://doi.org/10.1109/IJCNN.2014.6889833.

Silva, W.L.S., Barbosa, F.G., 2015a. Automatic voice recognition system based on multiple support vector machines and mel-frequency cepstral coefficients. In: Proceedings of the 11th International Conference on Natural Computation (ICNC), pp. 665–670. https://doi.org/10.1109/ICNC.2015.7378069.

Silva, W.L.S., Barbosa, F.G., 2015b. Multiple support vector machines and MFCCS application on voice based biometric authentication systems. In: Proceedings of the IEEE International Conference on Digital Signal Processing (DSP), pp. 712–716. https://doi.org/10.1109/ICDSP.2015.7251968.

Silva, W.L.S., Barbosa, F.G., 2015c. Support vector machines, mel-frequency cepstral coefficients and the discrete cosine transform applied on voice based biometric authentication. In: Proceedings of the SAI Intelligent Systems Conference (IntelliSys), pp. 1032–1039. https://doi.org/10.1109/IntelliSys.2015.7361270.

Silva, W.L.S., Batista, G.C., 2015a. Application of support vector machines and two dimensional discrete cosine transform in speech automatic recognition. In: Proceedings of the SAI Intelligent Systems Conference (IntelliSys), pp. 687–691. https://doi.org/10.1109/IntelliSys.2015.7361215.

Silva, W.L.S., Batista, G.C., 2015b. Application of support vector machines to recognize speech patterns of numeric digits. In: Proceedings of the 11th International Conference on Natural Computation (ICNC), pp. 831–836. https://doi.org/10.1109/ICNC.2015.7378099.

Silva, W.L.S., Batista, G.C., 2015. Using Support Vector Machines and two dimensional discrete cosine transform in speech automatic recognition. In: Proceedings of the International Joint Conference on Neural Networks (IJCNN). IEEE, pp. 1–5. https://doi.org/10.1109/IJCNN.2015.7280407.

Silva, W.L.S., Batista, G.C., Menezes, A.G., 2016. Automatic speech recognition using support vector machine and particle swarm optimization. In: Proceedings of the IEEE Symposium Series on Computational Intelligence (SSCI), pp. 1–6. https://doi.org/10.1109/SSCI.2016.7850125.

Snyder, D., Garcia-Romero, D., McCree, A., Sell, G., Povey, D., Khudanpur, S., 2018. Spoken language recognition using x-vectors. Odyssey: The Speaker and Language Recognition Workshop, Les Sables dOlonne.

Souza, G., Neto, N., 2016. An automatic phonetic aligner for brazilian portuguese with a praat interface. In: Silva, J., Ribeiro, R., Quaresma, P., Adami, A., Branco, A. (Eds.), Proceedings of the Computational Processing of the Portuguese Language. Springer International Publishing, Cham, pp. 374–384.

Stacks, D.W., Salwen, M.B., 2014. An Integrated Approach to Communication Theory and Research. Routledge.

Suresh, M. J. S., Thorat, S., 2018. Language identification system using MFCC and SDC feature. LANGUAGE.

Swietojanski, P., Ghoshal, A., Renals, S., 2012. Unsupervised cross-lingual knowledge transfer in DNN-based LVCSR. In: Proceedings of the IEEE Spoken Language Technology Workshop (SLT), pp. 246–251. https://doi.org/10.1109/SLT.2012.6424230.

Teixeira, A., Hmlinen, A., Avelar, J., Almeida, N., Nmeth, G., Fegy, T., Zaink, C., Csap, T., Tth, B., Oliveira, A., Dias, M.S., 2014. Speech-centric multimodal interaction for easy-to-access online services a personal life assistant for the elderly. Procedia Comput. Sci. 27, 389–397. https://doi.org/10.1016/j.procs.2014.02.043. 5th International Conference on Software Development and Technologies for Enhancing Accessibility and Fighting Info-exclusion, DSAI 2013

Tong, S., Garner, P.N., Bourlard, H., 2017. An investigation of deep neural networks for multilingual speech recognition training and adaptation. In: Proceedings of the Interspeech. ISCA. https://doi.org/10.21437/interspeech.2017-1242.

Tong, S., Garner, P.N., Bourlard, H., 2018. Cross-lingual adaptation of a CTC-based multilingual acoustic model. Speech Commun. 104, 39–46. https://doi.org/10.1016/j.specom.2018.09.001.

Torres-Huitzil, C., Girau, B., 2017. Fault and error tolerance in neural networks: a review. IEEE Access 5, 17322–17341.

University, C. M., 1986. SPHINX Website. https://cmusphinx.github.io/Accessed: 2018-11-30.

Vanhoucke, V., Devin, M., Heigold, G., 2013. Multiframe deep neural networks for acoustic modeling. In: Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing, pp. 7582–7585. https://doi.org/10.1109/ICASSP.2013.6639137.

Veiga, A., Candeias, S., Celorico, D., Proença, J., Perdigão, F., 2012. Towards Automatic Classification of Speech Styles. In: Caseli, H., Villavicencio, A., Teixeira, A., Perdigão, F. (Eds.), Proceedings of the Computational Processing of the Portuguese Language. Springer Berlin Heidelberg, Berlin, Heidelberg, pp. 421–426.

Veiga, A., Lopes, C., Sá, L., Perdigão, F., 2014. Acoustic Similarity Scores for Keyword Spotting. In: Baptista, J., Mamede, N., Candeias, S., Paraboni, I., Pardo, T.A.S., Volpe Nunes, M.d. G. (Eds.), Proceedings of the Computational Processing of the Portuguese Language. Springer International Publishing, Cham, pp. 48–58.

Veras, J.C.S., Prego, T.d.M., de Lima, A.A., Ferreira, T.N., Netto, S.L., 2014. Speech quality enhancement based on spectral subtraction. In: Proceedings of REVERB Challenge Workshop, p1, 7.

Vesel, K., Karafit, M., Grzl, F., Janda, M., Egorova, E., 2012. The language-independent bottleneck features. In: Proceedings of the IEEE Spoken Language Technology Workshop (SLT), pp. 336–341. https://doi.org/10.1109/SLT.2012.6424246.

Vu, N.T., Breiter, W., Metze, F., Schultz, T., 2012. Initialization schemes for multilayer perceptron training and their impact on asr performance using multilingual data. In: Proceedings of the Thirteenth Annual Conference of the International Speech Communication Association.

Vu, N.T., Imseng, D., Povey, D., Motlicek, P., Schultz, T., Bourlard, H., 2014. Multilingual deep neural network based acoustic modeling for rapid language adaptation. In: Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 7639–7643. https://doi.org/10.1109/ICASSP.2014.6855086.

Wang, J., Lin, C., Chen, E., Chang, P., 2014. Spectral-temporal receptive fields and MFCC balanced feature extraction for noisy speech recognition. In: Proceedings of the Signal and Information Processing Association Annual Summit and Conference (APSIPA), 2014 Asia-Pacific, pp. 1–4. https://doi.org/10.1109/APSIPA.2014.7041624.

Watanabe, S., Hori, T., Hershey, J.R., 2017. Language independent end-to-end architecture for joint language identification and speech recognition. In: Proceedings of the IEEE Automatic Speech Recognition and Understanding Workshop (ASRU), pp. 265–271. https://doi.org/10.1109/ASRU.2017.8268945.

Zen, H., Agiomyrgiannakis, Y., Egberts, N., Henderson, F., Szczepaniak, P., 2016. Fast, compact, and high quality lstm-rnn based statistical parametric speech synthesizers for mobile devices. arXiv preprint arXiv:1606.06061.