Language Resources and Evaluation
https://doi.org/10.1007/s10579-023-09648-1

SURVEY

Human-inspired computational models for European Portuguese: a review

António Teixeira[1] · Samuel Silva[1]

Accepted: 14 February 2023

Abstract

This paper surveys human-inspired speech technologies developed for European Portuguese and the computational models they integrate and made them possible. In this regard, it covers systems for synthesis and recognition as well as information on the methods adopted for the speech production studies that were performed, in parallel, to support them. And, on doing so, it can also contribute to provide an entry point for those who work in the field but are not familiar with these particular areas, including: context, history, and comprehensive references. As the great majority of work in these areas for European Portuguese was done by the first author's research group, this paper can also be seen as a review of more than 25 years of research at University of Aveiro in these topics.

Keywords Human-inspired models · Portuguese · Articulatory speech synthesis · Audiovisual speech synthesis · Silent speech interface · Speech production

1 Introduction

Being speech one of the distinctive characteristics of humans, the interest in creating machines and systems capable of imitating our capabilities, for its production, recognition and understanding, has a long history (Schroeder, 1999). The use of speech in the interaction with computers, and other systems, can foster advantages at various levels, allowing, for example, the use of systems by people with disabilities, telephone access to information services, and keep our eyes and hands free for other tasks.

✉ António Teixeira
ajst@ua.pt

Samuel Silva
sss@ua.pt

[1] IEETA, Department of Electronics, Telecommunications and Informatics, University of Aveiro, Aveiro, Portugal

Published online: 03 May 2023

A diversified set of approaches were explored, over the years, ranging from methods closely inspired in humans' speech production and understanding—for which we adopt the designation Human-inspired models (HIMs)—to artificial approaches with scarce or null direct relation with human processes. Despite the good results obtained with artificial approaches, HIMs can complement them and have potential to inspire new solutions, as demonstrated, for example, in computational models for visual recognition inspired in the human visual cortex (e.g., Wang et al. 2015) or anthropomorphic robots (e.g., Duffy 2003). Also very important, HIMs contribute to to a better understanding of the speech production process, knowledge that can be used to improve current systems. These systems are also very useful as versatile informants in the creation of stimuli for perceptual experiments (e.g., Cooper 1962; Story 2019; Teixeira et al. 1998b), contributing to advance knowledge regarding the speech perception and comprehension process.

HIMs have been used in the production of speech from text (speech synthesis), the creation of representations inspired in human hearing and perception (e.g., the widely used Mel-frequency cepstral coefficients (MFCC) and cochleograms), and in speech and speaker recognition, among others.

Besides inspiration for such models, the human speech production process (Levelt, 1993), from the intention to articulation and its results (not only acoustic and including visible effects in the face and measurable electric signals related to brain and muscles), provides a rich set of information sources that can be used to replace or complement the acoustic speech signal in recognition tasks (Freitas et al., 2016). This type of use can be seen as a second form of HIMs. It includes, for example, Silent Speech Interfaces (SSI) (Freitas et al., 2016). While the first type of systems only considers the speech signal, this second type explores other sources of information related to the human speech production process.

Both types of HIMs systems require detailed information regarding the human speech production and perception processes, posing many challenges and constituting a fertile area for the application and development of computational models.

For European Portuguese, research related to Human-inspired computational models addressed essentially:

(1) Replication/simulation of the human speech production process, aiming at speech and audiovisual synthesis;
(2) Exploration of information from the different speech production phases to develop SSI;
(3) Computational models to obtain information from the speech production process to drive both synthesis and SSI.

The first and second topics relate directly to both roles humans and machines can assume in communication: being speakers and listeners.

This paper is a survey aiming to present the evolution and computational models that made possible the development of HIMs targeting synthesis and recognition of European Portuguese, as well as the speech production studies made in parallel to support them. It provides an entry point for those who work in the field but not familiar with these particular topics, including: context, history, and comprehensive references.

1.1 Paper structure

Aligned with the three major topics identified above, this document is structured as follows: Sect. 2 reviews HIMs for speech synthesis, from articulatory models to recent articulatory-based approaches for audiovisual synthesis; Sect. 3 focuses on silent speech recognition, from initial explorations evolving into multimodal approaches and most recent experiments with radar technology; the computational models applied to speech production studies aiming to inform both the work on Human-inspired synthesis and SSI are presented in Sect. 4. Finally, the document concludes with a general discussion, in Sect. 5, along with overall conclusions and some ideas for future directions, in Sect. 6.

2 Human-inspired computational models for speech synthesis

Articulatory synthesis systems generate the speech signal by modeling the physical, anatomical, and physiological characteristics of the organs involved in human voice production (Teixeira et al., 2005). They model the production system instead of the signal or its acoustic characteristics.

Articulatory synthesizers usually include two subsystems: an anatomic-physiological model of the structures involved, the articulatory model, and a model of the production and propagation of sound in these structures, known as acoustic model. The later integrates sub-models to simulate: the creation of a source of periodic excitation, noise sources caused by the turbulent flow, propagation of the sound, and radiation at the lips and/or nostrils (Teixeira et al., 2005).

2.1 Initial models

The work in HIMs for simulation of speech production (a.k.a. articulatory speech synthesis) started in 1994/1995 and the first models were made public in Teixeira et al. (1997b) and Branco et al. (1997).

Both were capable of synthesizing European Portuguese vowels and diphthongs, including articulatory and acoustic models. They adopted the 2D articulatory model of Mermelstein (Mermelstein, 1973) with modifications by researchers of the Mind-Machine Research Center of University of Florida (Prado, 1991). They differ in terms of the acoustic model: Branco et al. (1997) adopts a digital filter while Teixeira et al. (1997b) uses a time-domain approach to perform analysis of transmission-line circuit, analogue of the vocal tract, and considers a non-interactive parametric model of the glottal flow derivative, known as LF-model, proposed by Fant et al. (1985).

These initial efforts included exploration of different acoustic models (time-domain and digital filters), assessing their limitations and advantages. After these two initial acoustic models (digital filters and time-domain approach) a frequency domain model was presented in Teixeira et al. (1998a) to allow the inclusion of frequency dependent losses, essential for adequate modeling of the nasal tract.

As nasality is a characteristic of Portuguese, that more differentiates it form other Latin languages such as Italian and Spanish (see for example Sampson (1999)), the initial modeling efforts also contemplated models for the nasal tract. A comprehensive nasal tract model was also developed adopting frequency domain (Teixeira et al., 1998a). The higher relevance of losses in propagation in the nasal tract was one of the main reasons for the development of the frequency-based acoustic model and its adoption in later developments. The obtained synthesizer was used mainly to synthesize and study nasal vowels (Teixeira et al., 1998b; Teixeira & Vaz, 2000a, b; Teixeira, 2000) and is presented in detail in Teixeira (2000).

2.2 Further developments

After 2000, the development focused on the extension of the synthesizer to other sounds besides vowels and nasals, and on the optimization of the computational implementations as well as user interaction.

Improvements to the computational models, increasing modularity and versatility resulted in a synthesizer for Windows, named `SAPWindows`[1] (Teixeira et al., 2002). Better visualization of the articulatory and acoustic information was also addressed (Silva et al., 2002).

The extension to other sounds started by addressing fricatives (Teixeira et al., 2003). The frication model added included: activation and deactivation of noise sources, calculation of the transfer functions from noise sources to the lips, convolution of noise sources volume velocity with the impulse response (obtained from the transfer function).

The improved synthesizer and all its models (vocal tract model; nasal tract model; interactive glottal source model; acoustic model (integrating the frication model) were presented in detail in Teixeira et al. (2005). It is a modular articulatory synthesizer for Portuguese, using object-oriented programming, and the adoption of a model-view-control architecture allows the addition of new models without major changes to the user interface. The implemented models comprise a flexible nasal tract area model and a hybrid acoustic model capable of dealing with asymmetric nasal tract configurations and frication noise sources. Synthesized speech has a quality ranging from fair to good. SAPWindows and the work regarding articulatory synthesis of Portuguese was also presented at the 2004 Workshop in Phonetics dedicated to the memory of inspiring inventor Farkas Kempelen (Teixeira et al., 2004).

More recently, in the scope of project HERON II, and building on the modular approach of SAPWindows, initial work was performed on the integration of a

[1] SAP stands for "Articulatory Synthesizer for Portuguese" ("Sintetizador Articulatório para o Português", in Portuguese).

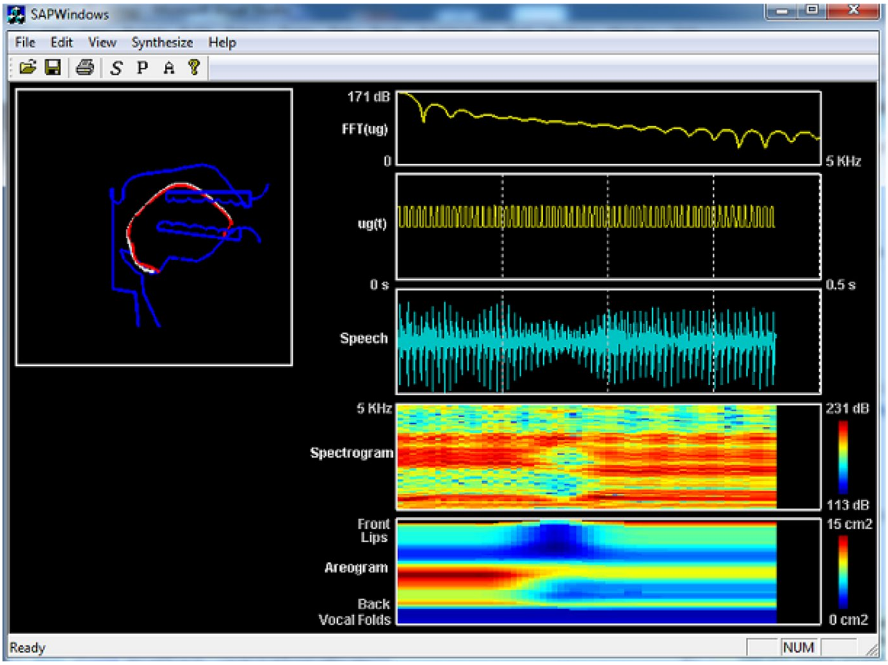

Fig. 1 Graphical user interface of SAPWindows, an articulatory synthesizer for European Portuguese, showing the results for the synthesis of the European Portuguese word “Ela” (she)

3D vocal tract model developed by Birkholz (2013a) and this led to first experiments with the modelling of lateral sounds (/l/). Laterals are a class of sounds characterized by the lowering of the sides of the tongue to allow the passage of air and the adoption of a 3D vocal tract model is an important step to study them. Figure 1 shows SAPWindows during the synthesis of European Portuguese word "Ela" (she). At left is the representation of the articulatory model, showing position of the articulators and track walls; At the right are presented, from the top, glottal wave information (signal and its FFT), synthesized speech wave, spectrogram of the synthesized speech and, an areogram (representation of the vocal tract areas over time), time-aligned with the spectrogram. The effects of the constriction during production of the fricative are noticeable in the speech signal, spectrogram and areogram.

2.3 Articulatory-based text-to-speech

The articulatory synthesizer mentioned above produces speech from the articulators’ trajectories over time. To create a text-to-speech (TTS) system capable of processing text as input, new models are needed to convert from textual input to the articulators’ trajectories.

As these models aim at producing information on the evolution of articulation, over time, a suitable theoretical framework is needed, and most of the research for

Portuguese adopts the articulatory phonology (AP) framework (Saltzman & Munhall, 1989; Teixeira et al., 2008b), having as basic unit of speech the articulatory gesture. Gestures are essentially instructions to achieve the formation (and release) of a constriction at some place in the vocal tract (for example, lips closure). The different sounds of a language are a combination of gestures (or created by a single gesture).

Formally, gestures are specified using a set of tract variables (e.g. lips and tongue tip) and the location and degree of constrictions in the vocal tract.

The articulatory phonology approach has been incorporated into a computational model (Saltzman and Munhall, 1989; Rubin et al., 1996) with three major components:

Linguistic gestural model analyses the input, converts it into a set of discrete, concurrently active gestures, and specifies a gestural score;

Task dynamic model calculates the articulators trajectories given the gestural score;

Vocal tract model takes articulators trajectories as input and calculates the tract shape, area function, transfer function, and speech waveform and generates an acoustic signal.

For European Portuguese, several contributions were made regarding the Linguistic Gestural model, as presented next.

2.3.1 Linguistic processing

As preparatory work for the creation of a system capable of synthesizing speech from textual input, models for syllabification, grapheme-to-phone (G2P) and for the gestural organization of European Portuguese were developed.

2.3.1.1 Syllabification- Coordination of speech gestures depends on syllable structure. The syllable information is essential for the correct functioning of the Gestural Model (Teixeira et al., 2008). Motivated by this need, several techniques were explored for integration of syllabification into an articulatory-based TTS system[2].

Oliveira et al. (2005b) explored finite state orthography-based syllabification, following the proposal of Bouma (2003) for Dutch. The process consisted in the composition of three transducers in sequence: First marking the nucleus, second inserting syllable boundaries, and last removing nucleus marks. Before this three step process proposed by Bouma, a transducer adds a mark representing an empty nucleus using a list of forbidden consonant sequences (e.g., pneu “tire” is transformed in pVneu).

The same authors, in Oliveira et al. (2005a, b), present an algorithm for European Portuguese syllabification, based on the proposal of Mateus and d’Andrade (2000).

[2] The work in European Portuguese for Portuguese is much more vast. Here we focus on work directly motivated or used in human-inspired systems.

Two variants of the algorithm were developed: one to process directly the graphemes, other to process the output of G2P (phones).

This Linguistics-based syllabification algorithm was evaluated with two small test sets obtaining accuracies of 99.57% and 98.85% when input consisted of word graphemes. The first test set consisted of 2076 common words, corresponding to a fraction of the Fundamental Portuguese corpus (Nascimento et al., 1987)

The second, including longer and more complex words, consisted of 1303 words randomly selected from the Público corpus created by Project Linguateca (Linguateca, 2008).

2.3.1.2 Grapheme-to-phoneme- After an initial proposal of G2P based on transducers (Oliveira et al., 2004) and the parallel creation of datasets with a reasonable number of syllabified words (more than 10 kWords), Machine Learning based approaches were considered.

Teixeira et al. (2006) describes European Portuguese G2P modules based on memory based learning (MBL), transformation based learning (TBL) and their combination with an existing rule-based system.

Combinations included parallel processing by all three systems and a majority voting scheme (“winner-take-all” method); and sequential processing (MBL followed by TBL).

The best results were obtained with the parallel approach, with a best phone error rate (PER) of 2.66%. The single system with overall better performance was based on MBL (best PER 3.76%). Results for all the approaches to G2P explored improved by inclusion of syllable information.

2.3.1.3 Gestural model- After conversion of the textual input to a phonetic transcription (including details about the syllables), the next step in this type of TTS systems is the specification of the gestural composition of each Portuguese phoneme. This information, coded in the Phonemes to Gestures dictionary (Nam et al., 2006), is the crucial part to create the Gestural Model for a specific language.

The Portuguese gestural dictionary was created by an iterative process (Teixeira et al., 2008b; Oliveira, 2009): (1) a first approximation was estimated from phonetic literature and from articulatory data in general; (2) target values for location and constriction degree were estimated from Magnetic Resonance Imaging (MRI) data; (3) test words, CV and VCV sequences were synthesized and their intelligibility and quality assessed.

2.3.1.4 Complete articulatory-based text-to-speech- Resulting from integration of all the models presented in this section, the first articulatory-based TTS system for European Portuguese was presented by Teixeira et al. (2008b) and its general architecture is presented in Fig. 2.

The system has three major parts:

(1) *Linguistic Processing* deriving information regarding phones, stress and syllable structure for the words to be synthesized;

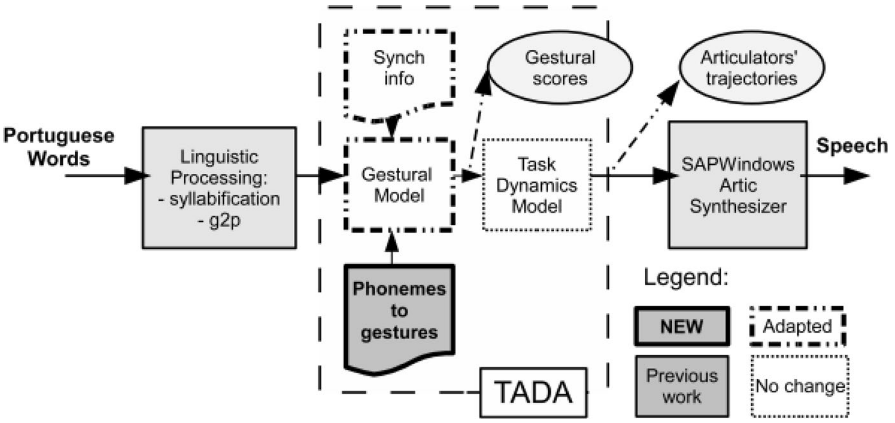

Fig. 2 General architecture of the articulatory-based TTS system developed for European Portuguese. It integrates an adaptation to Portuguese of the TAsk dynamics application (TADA). From Teixeira et al. (2008b)

(2) *The TAsk dynamics application (TADA)* (Nam et al., 2006) *adapted for European Portuguese* creating a Gestural Score, by combining the information received from the Linguistic Processing with a Phonemes to Gestures dictionary and an intergestural coupling model (Synch info in the figure), and feeding it to the Task Dynamics Model, that produces articulators' trajectories;

(3) *Articulatory synthesizers* converting articulators' trajectories to synthetic speech.

Some of the parts, as the Linguistic Processing, were entirely developed by the authors; others, as the Gestural Model, are adaptations to the original TADA system.

2.4 Computational methods to derive parameters for the models

In parallel with modeling efforts and to provide the needed parameters to configure and drive the synthesis, several computational methods have been explored.

Initially, due to the nonexistence of detailed information regarding European Portuguese articulation, processes to obtain articulatory descriptions for European Portuguese vowels from features extracted from speech (essentially formant frequencies) were investigated. In Teixeira et al. (1997a) and Teixeira (2000) optimization of tract configurations from formants using Simulated Annealing was explored. Several error measures were used, including weighted Euclidean and Bark scale. The number of formants used in the optimization process was varied between 1 and 4. During optimization, the configurations were limited to those that were anatomically plausible. Resulting configurations were "as predicted by articulatory phonetic descriptions" (Teixeira et al., 1997a).

With the work on direct derivation of the vocal tract configuration and articulation, described in Sect. 4, and the new challenges brought by the complete articulatory-based TTS, particularly the Gestural Model, this line of research was resumed more recently. The work considering Real-Time Magnetic Resonance Imaging (RT-MRI) of the vocal tract (see Sect. 4), along with the methods proposed for a more

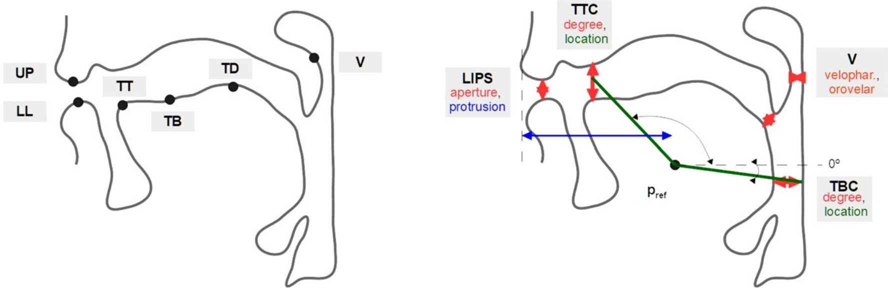

Fig. 3 Two different approaches to represent the vocal tract configuration considered for the research on the data-driven determination of critical articulators: on the left, a set of landmarks is considered, resembling the position of EMA pellets; on the right, a set of variables, aligned with the Task Dynamics framework, considering constrictions and their locations and degrees

systematic processing and analysis of the images to extract articulatory data created the conditions to explore novel paths for a phonological description of European Portuguese sounds. Instead of relying on a more human-based analysis of articulatory data, the aim was to propose a data-driven method that could automatically determine, for each sound, the set of critical articulators involved in its production, essential information for the European Portuguese Gestural Model.

First, Silva and Teixeira (2017b) showed that the statistical method proposed by Jackson and Singampalli (2009), for Electromagnetic Articulography (EMA) data, could also be applied to RT-MRI data using vocal tract contours extracted from 14 Hz RT-MRI data. To this effect, the vocal tract configuration, for each sound, was defined by the position of landmarks (over the lips, tongue and velar contours) simulating the pellets in EMA (see Fig. 3, left panel). However, and even though a static analysis was considered, selecting one frame as representative of each sound's tract configuration (e.g., central frame for vowels, closed lips for bilabials), the low frame-rate of the data made the choice of this key frame questionable.

More recently, Silva et al. (2019) explored higher framerate (50 Hz) RT-MRI data considering a larger amount of data per speaker and showing a positive impact of both these aspects on the outcomes, with the unsupervised phonological descriptions for European Portuguese sounds, showing strong agreement with earlier descriptions performed manually (Oliveira, 2009).

While these first works established the viability of the method, it was important to move beyond a tract representation based on landmarks (mimicking EMA), towards a consolidation of this research with the work on articulatory synthesis. To this end, Silva et al. (2020a, b) explored a representation of the tract considering variables aligned with the Task Dynamics framework, adopting the concept of constrictions (characterized by their location and degree). This allowed for a more compact representation of the vocal tract (i.e., less variables involved) and a more direct relation with existing Articulatory Phonology descriptions. Figure 3 shows both representations adopted for the research in data-driven critical articulator determination.

2.5 Audiovisual modeling

Besides the acoustic speech signal, speaking also produces visible effects, used by interlocutors to improve message decoding, particularly in noisy conditions. Motivated by the limitations of state-of-the-art audiovisual speech methods in providing natural coarticulation, a new line of research emerged, articulatory-based audiovisual synthesis (AVS), proposed by Silva and Teixeira (2017a).

In Silva and Teixeira (2017a) a conceptual framework for creating the conditions to foster a more structured evolution of the research in articulatory-based AVS was proposed. Aligned with this conceptual framework, the first proof-of-concept was presented by Silva et al. (2016) for the AVS module relying on: (1) the TADA (Nam et al., 2006; Saltzman & Munhall, 1989)) to compute articulators (e.g., lips, tongue) movement over time; (2) an articulatory synthesizer; (3) a realistic face model driven by the same articulators' trajectories. The fact that the movements of the articulators controlling the visual output and the speech signal are produced from the same information is an advantage since they are intrinsically synchronized. Currently, this AVS is being evolved in several aspects, including the use of deep learning approaches (Bi-LSTM) to improve the previous work on grapheme-to-phoneme, enabling an audiovisual synthesizer able to deal with a wide range of input texts.

2.6 Comparison with state-of-the-art

The articulatory synthesizers developed for Portuguese and a major part of its constituent models were at state-of-the-art level or contributed to the advancement of it when they were presented. This was particularly true for simulation/synthesis of nasals. For example, the nasal model continues to be, to the best of the authors' knowledge, the most comprehensive in the literature. As the last improvements of some of the models presented here were made a few years ago, even though they continue to be state of the art for Portuguese, they are being surpassed by recent work in aspects such as the modeling of other classes of sounds, use of 3D models, and new control models to derive articulators' trajectories (e.g. Kröger and Birkholz 2009; Stone et al. 2020).

Regarding the complete TTS system, despite its limitations, is one of a few articulatory-based complete systems. To best of our knowledge there is only another complete TTS system, developed by Haskins Laboratories for English (Nam et al., 2001, 2006). The extension of the TTS system to Audiovisual synthesis was proposed initially by the authors for Portuguese and, to best of our knowledge, has not been done for other languages.

Finally, one distinguishing characteristic of the work being carried out for the determination of critical articulators concerns its alignment with HIMs, in the sense that we are not just seeking for data-driven approaches that are blind to the speech production process. Instead, we seek to do so without losing sight of the the variables aligned with the articulatory phonology framework and this is, to the best of our knowledge, novel. This has two main advantages: it keeps the results interpretable in light of the speech-related anatomy, and provides outcomes that can, in the future, support further developments on on articulatory synthesis, particularly in the definition of gestural scores.

Despite not being a mainstream research topic, there have been some continuous evolution in articulatory speech synthesis. Due to the entailed high complexity, as mentioned before, only very few complete systems capable of performing articulatory synthesis continue to exist. A notable example of the current state-of-the-art is the articulatory synthesizer VocalTractLab (Birkholz, 2013b). The quality of the speech synthesized by this type of synthesizers has continually improved, in recent years, but it is not yet competitive with other speech synthesis methods, such as unit-selection or neural end-to-end systems (Krug et al., 2021), as can be confirmed by examples made available online by Birkholz (2022). One of the main current lines of research to improve quality includes the exploration of more realistic tract and acoustic models, going beyond classic one-dimensional (1D) models and plane wave propagation within the vocal tract (e.g., Freixes et al. 2019; Blandin et al. 2022).

3 Silent speech interfaces

Silent Speech Interfaces (SSI) use sensor data from elements of the human speech production process (e.g., muscular and articulators activity, or visible changes in the face) to perform speech recognition in the absence of an intelligible acoustic signal, providing human-computer interface (HCI) modalities usable in high-background-noise environments, such as a crowded restaurant or by speech-impaired individuals (Denby et al., 2010)

3.1 Initial exploratory work

Until 2011, SSI research has been mainly done by groups from the USA, UK, Germany, France and Japan, which have focused their experiments on their respective languages (e.g., Calliess & Schultz, 2006; Hueber, 2008; Tran et al., 2010). There was no published work for European Portuguese in the area of SSIs.

Work on SSI for Portuguese started by visual speech recognition (VSR) and Doppler experiments. One of the first VSR systems (Freitas et al., 2011) consisted of a pipeline of feature extraction (several were considered: scale invariant feature transform (SIFT) (Lowe, 2004), principal component analysis (PCA)-SIFT (Ke & Sukthankar, 2004), speeded up robust features (SURF) (Bay et al., 2006) and fast invariant to rotation and scale transform (FIRST) (Bastos & Dias, 2009)), feature calibration, post-processing to remove outliers, and classification based on the distance obtained by applying dynamic time warping (DTW).

The process consisted of: (1) random selection of K observations from each word in the selected corpus to be used as the reference pattern (training); (2) comparison of test examples with the representative examples, selecting the word with the minimum distance, (3) repetition of the process N times.

This first system was assessed with a database of 112 videos containing 14 repetitions by a single speaker of 8 European Portuguese words. The 8 words were not randomly selected and represent four pairs of words containing oral and nasal vowels (e.g. *cato/canto*) and sequences of nasal consonant followed by nasal or oral vowel (e.g. *mato/manto*). Considering N = 20 and K varying from 1 to 10, the best result was achieved for K = 9, having an average Word Error Rate (WER[3]) of 8.63% and 2.5% WER for the best repetition.

Almost at the same time, the first SSI experiments considering Doppler-based sensing were presented for Portuguese (Freitas et al., 2011). This technique is based on the emission of an ultrasound pure tone towards the speaker's face that is received by an ultrasound sensor tuned to the transmitted frequency. The reflected signal contains frequency shifts caused by speaker's face movements due to the Doppler Effect[4]. Based on the analysis of the Doppler signal, patterns associated with the movements of the facial muscles, lips, tongue, and jaw can be extracted (Toth et al., 2010).

3.2 Unimodal approaches

This initial exploratory work was followed by more complete experiments aiming to achieve SSI at state-of-the-art level for Portuguese and, if possible, contributing to its advance. Being a challenge, a less studied aspect of SSI (Denby et al., 2010) and a distinctive characteristic of Portuguese, these studies gave particular attention to nasality, always favoring less invasive solutions. Experiments contemplated surface electromyography (sEMG) (Freitas et al., 2012a), and further exploration of Doppler-based sensing (Freitas et al., 2012b) and VSR (Abreu, 2014).

3.2.1 Surface electromyography

The first experiments with sEMG sensors in SSI for Portuguese, presented by Freitas et al. (2012a), adopted four pairs of electrodes positioned to sense activity of several muscles (tongue, anterior belly of the digastric, zygomaticus major, lower orbicularis oris, and levator anguli oris). From the signal of these electrodes several frame-based temporal and spectral features were extracted (including mean value, power value, zero-crossing rate) and the dimensionality of the feature set was reduced by applying PCA, in order to obtain 32 coefficients per 30 ms frame. For classification, the Dynamic Time Warping (DTW) technique was used to find an optimal match between the observations. DTW was chosen due to the small number of data samples available given the complexity of the acquisition protocol and also because it addresses well the temporal alignment of time varying

[3] Word Error Rate (WER) is given by the number of incorrect word classifications divided by the total number of observations considered for testing.

[4] The Doppler Effect is the change in frequency of an emitted wave perceived by a listener moving relative to the source of the wave. Considering a source T that emits a wave with frequency f_0 that is reflected by the moving object—in our case the speaker's face –, the reflected signal is then $f = f_0 \frac{v_s + v}{v_s - v}$, with v being the velocity of the moving object based on the transmitter T and v_s is the velocity of the sound in the medium.

signals of different duration, precisely the case. The evaluation was performed with two corpora—defined to cover nasality: PT-EMG-A consisting of 8 different European Portuguese words, 4 words that are part of a minimal pair where the presence or absence of nasality in one of its phones is the only difference, and 4 digits; PT-EMG-B consisting also of 8 European Portuguese words representing four minimal pairs containing oral and nasal vowels (e.g. *cato/canto*) and sequences of a nasal consonant followed by nasal or oral vowel (e.g. *mato/manto*). Detailed information regarding these corpora is presented in Table 1. Both corpora were recorded by a single speaker on a unique recording session (no electrodes repositioning was considered). More than 10 repetitions were recorded for each word (12 for PT-EMG-A and 15 for PT-EMG-B).

The best results obtained, similar to state-of-the-art at the time (Schultz & Wand, 2010), are presented in Table 2.

The best results obtained for EP, with PT-EMG-A corpus, had "slightly worst accuracy when compared with the latest state-of-the-art results for EMG-based recognition" (Freitas et al., 2012a). Representative of the state-of-the-art at the time, Schultz and Wand (2010) report that their "final system achieves 10% word error rate for the best-recognized speaker on a 101-word vocabulary". The results with PT-EMG-B were much worse, showing clearly the limitation of sEMG-based SSI for languages with distinctive use of nasality, such as Portuguese.

3.2.2 Ultrasound Doppler

For the Doppler-based SSI (Freitas et al., 2012b) a dedicated circuit board was developed, allowing simultaneous acquisition of speech and doppler signals. The acquired signal is zero-meaned, filtered to suppress the carrier, differentiated and split into 50 ms frames prior to feature extraction. Feature extraction entails the calculation of the Discrete Fourier transform (DFT) in the interval of 3500 to 4750 Hz and application of a discrete cosine transform (DCT) to the DFT results to decorrelate the signal and extract the first 38 coefficients (containing most of the signal energy). Based on the number of available observations and considering the limited vocabulary, classification based on DTW distance to training examples was also used. Following this procedure, the best results yielded a WER of 27.8% for a vocabulary of 10 digits on an isolated word recognition problem. These results compared reasonably well with recognition WER of 67% reported for a continuous speech recognition task of 10 English digits (Srinivasan et al., 2010), although European Portuguese results are for an isolated digit recognition task and the test conditions are not the same.

3.2.3 Visual speech recognition

For VSR, Abreu (2014) used Kinect One to extract geometric and articulatory features from the lips. The green channel of RGB was used to extract the external

Table 1 Words that compose the PT-EMG-A and PT-EMG-B corpora and their respective phonetic transcription

Corpus	Words/word pairs	Phonetic transcription
PT-EMG-A	cato	[katu]
	peta	[petɐ]
	mato	[matu]
	Tito	[titu]
	um (1), dois (2), três (3), quatro (4)	[ũ], [doiʃ], [treʃ], [kwatɾu]
PT-EMG-B	cato/canto	[katu] / [kɐ̃tu]
	peta/penta mato/manto	[petɐ] / [pẽtɐ]/ [matu] / [mɐ̃tu]
	Tito/tinto	[titu]/[tĩtu]

points of the lips and Cr channel from the YCbCr color space[5] was used to obtain the internal points of the lips. After the featureswere extracted, length and distance (from subject to the Kinect) normalizations were applied to the feature vectors. The selected vocabulary consisted of 24 European Portuguese words (10 digits from zero to nine plus 14 common words taken from context free grammars of an ambient assisted living (AAL) application that supports speech input) (Teixeira et al., 2012c). Images of the face of 8 European Portuguese native speakers (7 male) were recorded for 9 repetitions of the selected words and silence state, providing a total of 225 recordings per speaker and a total of 1800 recordings for the 8 speakers (Abreu, 2014). This recordings were processed to extract geometric (the width, height and depth of the lips) and articulatory features (opening, rounding and protrusion of the lips) and used to train and test Support Vector Machines (SVM) classifiers. Best results were obtained using geometric features, with an WER of 32% for the complete set of words and 24% when considering only the AAL subset.

3.2.4 SSI modalities and nasality

Due to the relevance of nasality[6] for European Portuguese and nasality being one of the challenges for SSI, the assessment of the effect of nasality in the various methods explored for European Portuguese SSI was a constant. For example, Freitas et al. (2012a) reported that "for an SSI based on sEMG, performance degradation will be verified due to difficulties of the technique in distinguishing nasal phones from oral ones".

The most interesting results regarding nasality and SSI were presented in Freitas et al. (2015), exploring the existence of useful information about the velum

[5] YCbCr is a family of color spaces used as a part of the color image pipeline in video and digital photography systems. Y, which is luminance and Cb and Cr are the blue-difference and red-difference chroma components.

[6] In European Portuguese there are five nasal vowels, three nasal consonants, several nasal diphthongs and triphthongs.

Table 2 Results obtained with sEMG sensors and DTW in SSI for EP by Freitas et al. (2012a)

Corpus	Best average WER		Best WER	
PT-EMG-A	22.50%	(K = 8)	9.38%	(K = 8)
PT-EMG-B	42.29%	(K = 10)	31.25%	(K = 11)

K stands for the number of repetitions considered to establish the DTW reference, for each word

movement in sEMG signals by time aligning EMG and RT-MRI. Results provided "evidence that it is possible to detect velum movement using sensors positioned below the ear, between mastoid process and the mandible, in the upper neck region".

3.3 Multimodal SSI

With the interesting results obtained with unimodal systems, the obvious next step was exploring multimodal solutions. Multimodal SSI, an advance in SSI state-of-the art, was proposed with a concrete instantiation considering video, depth, surface EMG, and ultrasonic doppler (Freitas et al., 2013). The devices employed were: (1) a Microsoft Kinect, to acquire visual and depth information; (2) an surface EMG acquisition system to capture myoelectric signals from the facial muscles; (3) a custom built Ultrasound Doppler system (UDS). This initial multimodal approach also adopted an example based classification method, using Dynamic Time Warping (DTW) distances and k-Nearest Neighbor (k-NN) classifiers. Results showed that a significant difference in recognition rates can be found between unimodal and multimodal approaches in favor of the latter, and that benefits can be obtained by aligning several modalities, especially when registering Video, Depth and UDS, or Video and Depth. Later, in 2014, feature selection was explored with interesting results (Freitas et al., 2014b). As a secondary result of this work, a multimodal corpora for SSI was created (Freitas et al., 2014c).

3.4 Applications

A context where the speech signal cannot be used or speech recognition performance is highly affected due to ambient noise is AAL. Trying to address these difficulties and exploring the previous advances in SSI for European Portuguese, an application of SSI to AAL was proposed in 2017 by Teixeira et al. (2017). These efforts led to the first working application for European Portuguese of VSR targeting older adults, enabling real-time control of a media player.

Visual information was obtained by a Kinect One camera. Following the classic approach in pattern recognition integrating feature extraction and classifiers, the system integrated: Activity Detection, Feature Extraction, and Classification. 8 words/word sequences considered natural for the target scenario (media player control) were selected (e.g. *Parar* [stop], *Mais lento* [slower]). Lips width and height plus chin position (x and y coordinates) were selected as features. The lips were chosen because it has proven to give good results in previous works (e.g., Abreu, 2014; Rao & Mersereau, 1994); the chin was added due to the role of the lower jaw in the human speech production process. A rich set of classifiers was tried (SVM; Random Forest; Sequential Minimal Optimization—SMO; AdaBoost and Naïve Bayes) and 3 were selected based on their compliance with real time requirements: Random Forest, SMO and Naïve Bayes. A “winner-take-all” approach was adopted to combine the decision of these 3 classifiers. The evaluation of the prototype was made by three users and “consisted in classifying a word in real time for VLC [a well-known media player] controlling purposes” Teixeira et al. (2017). Speaker dependency of the system was also tested, training the system with a database recorded for one speaker and testing with another. To train the classifiers for the evaluation, five different databases were created: 3 databases for Speaker 1 (each recorded at a different distance—0.6 m, 1 m and 2 m away from the Kinect Camera); 1 database for Speaker 2; and 1 database for Speaker 3. Speakers 2 and 3 recorded at 1 m from the Kinect camera. The first tests were performed in matching conditions of test and train regarding speaker and distance. After, the effect of distance was assessed followed by some speaker dependency tests, assessing if the developed system could perform well when trained with data from other speakers. The best result in offline evaluation was achieved for a speaker 2 ms away from the Kinect, with 30% of commands missed. The system also revealed interesting performance in real time control of the media player, with an error rate of 18.7% and 1.3 s to perform a classification.

3.5 Recent developments

To face limitations of previously explored SSI methods regarding everyday use (e.g., the need to place equipment in contact with the speaker), environment conditions (e.g lighting for video) and privacy, Ferreira (2021) explored, for the first time, radar for a small vocabulary (a novelty not only for European Portuguese). The prototype uses frequency-modulated continuous wave (FMCW) radar technology (Texas Instruments AWR1642 FMCW sensor) to acquire user’ facial velocity dispersion patterns while they produce speech at an approximate distance of 15 cm from it. During data acquisition, the raw data obtained is assembled into several tag-length-value (TLV) packets. To process these packets, a parser was developed in Matlab to extract all the detected objects’ relevant information (i.e., their Cartesian coordinates (x, y, z) and relative velocities expressed in the radar frame of reference). These data include the entire point cloud detected in the radar field-of-view (FoV), over the time acquisition windows, from which distinct subsets of points are clustered and associated, in real-time, by the the radar board, to represent different objects, or parts of a body, with dissimilar velocity measures. The dispersion of velocity data associated with the users’ facial motions is extracted and used as input for a classification module. A corpus of 13 words was acquired, for 3 speakers, and different classifiers were used (Random Forests, Linear Discriminant Analysis, linear regression (LR), SVM, and bagging (BAG)). The best results were obtained using the BAG classifier with an average error rate of 17.4%, encouraging further investigation of this technology for SSI. Developments of this initial prototype and an improved evaluation were later presented by Ferreira et al. (2022). This more recent evaluation involved 4 European Portuguese native speakers. For each of the 13 words (the same of the initial study), 60 validated repetitions were considered per participant. Regarding the speaker-dependent models, trained and tested with data from each speaker while using 5-fold cross-validation, average WER of 15.5% and 12.0% were respectively obtained from BAG and LR classifiers.

3.6 Comparison with state-of-the-art in SSI

Work in SSI for Portuguese already covered an extended set of modalities (from EMG to Radar) and state-of-the-art results were achieved for most of the models developed at the time of publication. As a result, Portuguese integrated the small set of languages with SSI systems (Silent Speech prototypes have been mainly developed and designed for English. Other languages with SSI systems include French, Japanese and Arabic.). The multimodal SSI system proposed in 2013 was a notable advance to the general SSI state-of-the-art given the number and variety of modalities considered and this kind of approach has, since then, greatly evolved in the literature.

Mainly due to the costs involved in the acquisition of big datasets, impossible to obtain, in general, for research targeting Portuguese, the work continues to focus on the exploration of new modalities and multimodal setups with "classic" Machine Learning classification approaches for small vocabularies while the SSI field evolved for continuous recognition with deep learning approaches (e.g., Tóth & Shandiz, 2020). Also, for Portuguese no work has been published on synthesis having as input the information from SSI sensors, an area quite active for other languages, such as Basque and Spanish (Hernáez Rioja et al., 2021; Gonzalez et al., 2016).

The current state-of-the-art in SSI continues to explore several of the information sources used by the authors, such as video, ultrasound, and radar. A recent special issue edited by Denby et al. (2022) includes some relevant recent research on SSI, being a good source for interested readers. One of the current trends is the evolution in the Machine Learning methods from classic ones, such as SVM, to deep learning (Mohd Shariff et al., 2022). Two examples of SSI implemented using a deep learning approach are SottoVoce (Kimura et al., 2019), based in ultrasound; and EarCommand (Jin et al., 2022), exploring "relationship between the deformation of the ear canal and the movements of the articulator". Another trend is the increase in vocabulary size to obtain large vocabulary systems (e.g., Yu et al. 2022). Recent work on radar-based SSI focused on the use of FMCW radar (Xu et al., 2019; Ferreira et al., 2022), and ultra-wideband radar (UWB) (Lee & Seo, 2019). More recently, Mohd Shariff et al. (2022) proposed the use of continuous-wave (CW) radar images as input for a convolutional neural network (CNN).

4 Computational models applied to European Portuguese production studies

Anthropomorphic synthesizers, the subject of Sect. 2, demand large amounts of detailed anatomic-physiological information, if possible in 3D, and their variation in time. As stated in Martins et al. (2008), "knowledge of the speech production mechanism is essential for the development of speech production models and theories", and, thus, gathering anatomical and physiological data is paramount to evolve the work presented in the previous sections. In this context, it is necessary to obtain information about the configuration of the vocal tract during the production of the various sounds not only for static snapshots, at a fixed time, representative of the main configuration characteristics of that sound, but also how the tract configuration varies, over time, in the various transitions between sounds. This work, started with EMA in 2000 (Teixeira & Vaz, 2001), and continued with MRI (2D, 3D and realtime). Our work contemplated not only image acquisition but also image processing to enable exploration of the large databases in a reasonable time. This section presents, in chronological order, the main developments and the computational models/methods involved.

4.1 2D and 3D MRI processing

The initial work of our group in this line addressed mainly processing of 2D and 3D MRI acquisitions.

Several image processing methods were tried to obtain contours in 2D images (Region Growing, Level Sets, Reaction Diffusion), being the best results obtained with Seeded Region Growing (Martins et al., 2007; Carbone, 2008), a robust method for image segmentation that starts with the placement of a set of seeds (a single pixel or a set of connected pixels) in the image to be segmented. Then, it grows these seeds into regions by successively adding neighboring pixels to them (Fan & Lee, (2015) for more information).

Volumes (3D information) were initially obtained by Martins et al. (2007), Carbone et al. (2007) and Carbone (2008), using direct 3D segmentation and a $2\frac{1}{2}$D method consisting of: determination of the centerline of the vocal tract in a saggittal slice; re-slicing of the volume in normal directions to this centerline; and segmentation of each slice (Martins et al., 2007; Carbone, 2008).

Later, MevisLab and ITK-Snap toolkit were used to implement more complex image processing methods aiming at extracting the tongue and detailed vocal tract areas (3D data) for laterals (Martins et al., 2010).

4.2 Modeling based in 2D and 3D MRI data

Segmentation of tongue images made possible obtaining 3D meshes that were used to study vowels and laterals (Martins et al., 2011, 2012b). As laterals cannot be well characterized by only a sagittal 2D contour, comparison of tongue 3D mesh models was particularly useful to advance knowledge and modeling of European Portuguese laterals (Martins et al., 2010, 2011, 2012b). Results showed that: /l/ has linguo-alveolar contact (laminal or apical), inward-lateral compression and convex shape of the posterior region of the tongue, and lateral channels alongside the tongue; the European Portuguese palatal lateral is articulated at the alveolo-palatal region and not exclusively at the palatal area.

Later, in Teixeira et al. (2012b) a more complete characterization of the palatal lateral was made, including length and average area of the different cavities (back, front, channels and supralingual) and complete area functions. This novel quantitative characterization of the European Portuguese palatal lateral - also novel for palatal lateral in world languages - made possible simulations of the acoustic response with fairly accurate prediction of the formant structure.

4.3 Realtime MRI processing

RT-MRI provides dynamic information of the entire vocal tract (see Fig. 4 for examples) with reasonable spatial and temporal resolution in a non-invasive manner (Teixeira et al., 2012a).

Initial approaches to segmentation of RT-MRI (2D images) were the same as those applied to static 2D imaging (Region Growing, Level Sets, Reaction Diffusion)(Carbone, 2008). But the lower image quality and the dynamic nature motivated other approaches.

Working on the first European Portuguese database of RT-MRI with synchronized audio, Teixeira et al. (2012a) applied a rapid automatic extraction method to determine constriction location targets and to estimate articulatory trajectories using pixel intensity in selected regions of the vocal tract. To study European Portuguese nasal vowels, Oliveira et al. (2012) used MevisLab to develop a segmentation method that applied Region Growing to a region of interest (ROI) roughly encompassing the vocal tract. Due to spatial coherence between the different RT-MRI images collected for each speaker, a single seed could serve for all images and allowed the segmentation of the vocal tract for all image frames. As experiments revealed that, due to different intensities, anatomical variations and artefacts resulting from proximity between structures, the segmentation results could be improved if several regions were considered. In this mindset, Silva et al. (2013) proposed an improvement by dividing the vocal tract in multiple regions, applying different region growing parameters to each one, and blending the outcomes to obtain the contour of the whole vocal tract.

While the previous segmentation approaches already enabled extracting data for initial characterizations of speech dynamics, namely regarding nasality and the role played by the velum, to be able to profit from the shear amount of data provided by RT-MRI, enhancements were required. One important requirement was that the extracted data already provided an annotation of key landmarks and articulators of the vocal tract (e.g., lips, tongue tip, velum) so that (semi-)unsupervised analysis methods could be developed to explore the data.

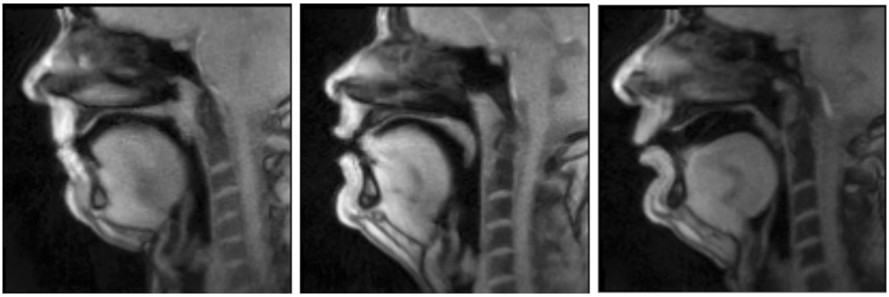

Fig. 4 Examples of images of the vocal tract extracted from RT-MRI sequences for three different speakers and showing, from left to right, the tract configuration to produce a /p/, an /n/, and the nasal vowel /ũ/

In Silva and Teixeira (2013, 2015), a novel segmentation method is presented considering active shape and active appearance models. The images in a particular RT-MRI sequence do not present a very large anatomical difference between neighbouring images. Taking this into consideration, one of the notable features of this method is that it initializes the segmentation of each image with the shape of the vocal tract segmented for the previous image. This means that the adjustments required for segmentation are smaller than what would be required if the method was initialized with the mean shape model, as usual. One of the consequences of this option is that the training of the shape model only requires a small number of images (annotated by hand), which entails less work for the speech scientist.

Finally, the resulting data and evolutions on systematic methods to explore it resulted in the presentation of a framework (Silva & Teixeira, 2014, 2016) encompassing the proposal of a set of measures and visualizations of vocal tract data comparison (an example can be found in Cunha (2019). The main purpose was to elicit discussion and provide overall grounds for the community to move towards forms of analysis and reporting of the results for speech studies that might foster, for instance, increased objectivity and a greater comparability across studies and, even, languages.

The work in this area for Portuguese made several contributions to the advancement of the state-of-the-art, namely: Segmentation and analysis of vocal tract from RT-MRI Image Sequences; quantitative models for production studies; articulatory characterization and modeling of laterals; detailed characterization of European nasals.

The current state-of-the-art methods to gather knowledge regarding speech production continue to actively explore RT-MRI. There is also an increased interest for ultrasound methods (e.g., Albuquerque et al., 2022). For RT-MRI, one of the main current trends for extracting production data is the application of deep-learning, motivated by the successes in other medical image analysis tasks. For example, Ruthven et al. (2023) developed the "first deep-learning-based framework specifically for registering dynamic two-dimensional MR images of the vocal tract during speech". This new method compared favorably with traditional methods and the methods that are not segmentation informed.

The proposed framework includes a deep-learning-based method to estimate segmentations of six anatomical features: the head, soft palate, jaw, tongue, vocal tract and tooth space (Ruthven et al., 2021). Other trends are: the increase in number of subjects integrating the databases (e.g., RT-MRI data from 75 American-English speakers (Lim et al., 2021)); extension of these studies to more languages (e.g., Kannada, a phonetically under-studied Dravidian language, Kochetov et al. 2020).

5 Discussion

From the previous sections, it is clear that research in the three main areas of application of models inspired on how humans produce speech can be characterized by:

Continuous improvement- As made clear in the graphical summary of the work reviewed, Fig. 5, relevant new results were obtained in a continued manner, each one leading to new challenges. The methods and computational models explored were in general aligned with the state-of-the-art, and several times leading it (e.g. RT-MRI automatic extraction of contours (Silva & Teixeira, 2015); framework for quantitative analysis (Silva & Teixeira, 2016); multimodal SSI (Freitas et al., 2016), and radar SSI (Ferreira, 2021)).
Synergies- Research in the three areas considered is not independent, with several examples of synergies among them. The main example being the use of the information from speech production studies to the development of articulatory-based models. For example, MRI information was crucial for the development of the Gestural Model of the complete TTS system and RT-MRI processing models to provide basis for future refinements to this model, by making possible automatic identification of critical articulators. Another example worth mention is the use of velum trajectories obtained from MRI to investigate the potential of EMG based SSI to handle nasal sounds.
Diversifyed use of computation models- Besides the developed models for articulatory synthesis, a diversified set of models was used, for instance, as grounds for the extraction of contours and volumetric information from MRI images (2D, 3D and realtime).
Several uses of Machine Learning- Despite the difficulties in obtaining sufficient data—that, in general, is not available for Portuguese and needs major investments—and the methodological bias toward knowledge-based approaches, Machine Learning approaches were explored in many problems, from grapheme-to-phoneme conversion to the recent experiments in radar-based SSI. The methods used were diverse, including, for example, kNN, MBL, TBL, Random Forests, Discriminant Analysis and SVMs. Deep learning approaches seem to be the next step, already being explored in the evolution of AVS system.
Focus on nasality- One of the distinctive characteristics of Portuguese language, vowels nasality, has been a favored research topic. Advances were made to articula-

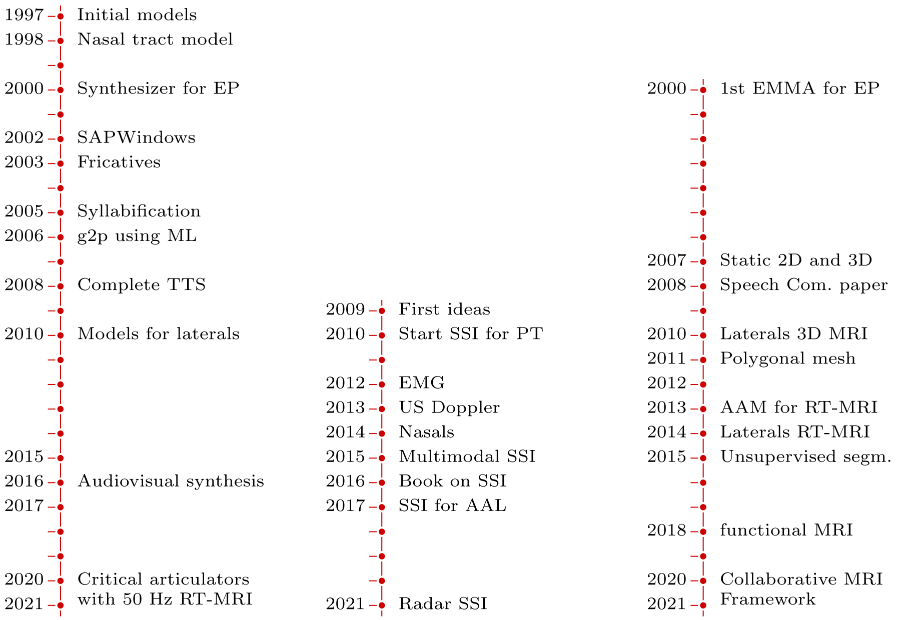

Fig. 5 Main events of the research for European Portuguese (EP) in Human-inspired models and supporting speech production studies. From the left, articulatory-based synthesis, Silent Speech Interfaces and Speech Production studies

tory models for nasals (vowels and consonants), the capability of SSI methods to handle nasal sounds investigated (e.g. Freitas et al. 2012a, 2014a), and results from MRI processing explored in a variety of phonetic studies (e.g. Oliveira & Teixeira, 2007; Martins et al., 2012a; Oliveira et al., 2012).

The amount of data and methods proposed throughout the years, along with a strong interest in further increasing the synergies among the different areas of research has also compelled us to move towards the creation of a framework to support a stronger and more efficient articulation among them. While at its infancy (Almeida et al., 2020), currently, research such as the one performed for the determination of critical articulators by Silva et al. (2019) is already performed with this platform providing an API for selecting and retrieving vocal tract contour data, for instance, from a Matlab script (Fig. 6).

Furthermore, important steps such as the revision of the segmented vocal tract data, to ensure data quality, can also be performed collaboratively, from anywhere. And, finally, by creating such a framework, to support the research in a field that is inherently multidisciplinary, we also intend that, in the long-term, researchers that may not have the technical skills to tackle the methods for data segmentation or analysis, can more strongly contribute with their expertise. This can be made possible by providing them with these resources of-the-shelf, in an interactive and usable framework for them to build upon.

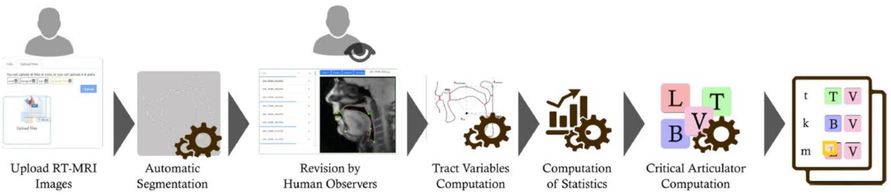
Upload RT-MRI Images
Automatic Segmentation
Revision by Human Observers
Tract Variables Computation
Computation of Statistics
L
T
V
B
Critical Articulator Computation
t
k
m
T
V
B
V
V

Fig. 6 Illustrative pipeline supported by ongoing work on a framework for supporting speech studies, aiming to integrate a set of resources resulting from our research, through the years, potentially improving how the different areas can be articulated

6 Conclusion

This paper presents a survey of research in speech technologies adopting in some degree Human-inspired models, covering three broad areas (synthesis, silent speech interfaces, and speech production studies).

Overall, it shows the continuous evolution of the models adopted to face increasingly complex challenges, summarized in Fig. 5, which contributes to both the existence of state-of-the-art models and systems not only for Portuguese, but also for the state-of-the-art of the area. The many studies made possible contribute to advances in knowledge about speech production, particularly of nasals and laterals. From these lines of research results a rich set of publications, with several of the developments published in reference journals or books (Freitas et al., 2016).

The adoption, as part of the main objectives, of Portuguese as target language, with scarce availability of needed resources, particularly corpora/datasets, was one of the major challenges for the concretization of the research reviewed. In many cases the only option was to acquire new datasets. This added to the duration and costs of the research projects to, among others: acquire the needed knowledge and comprehension of the state-of-the-art in the topic; establish the team with the needed expertise—in general multidisciplinary, combining fields such as Phonetics, Linguistics, Medical Imaging, Image Processing, and Computer Science; obtain access (e.g., for RT-MRI acquisition) or create the experimental setup (e.g., for SSI); and define the corpora. To reduce these problems—cost and duration—the following approach was adopted: combining careful study of state-of-the-art, to augment efficiency of resources creation; collaboration with research teams with the needed experience and facilities to make possible acquisitions (e.g. first EMA data was acquired at LMU in Munich supported by Phil Hoole, one of the specialists in this technique); and development of automatic processing methods (e.g. MRI processing). Nevertheless, while enabling the acquisition of small and medium sized datasets, covering a wide range of state-of-the-art technologies, the development of large scale data resources was not possible, a requirement to explore recent processing methods, particularly those involving deep learning. The need to create new resources (data and tools) did not bring only problems and challenges. On the positive side, it allowed designing them to directly fit our research needs, to contribute to expansion of the set of languages with resources for speech production studies, the development of Human-inspired methods, and, in several cases, to the state-of-the-art in this type of resources (e.g., multimodal dataset for SSI).

Evolution has not stopped, in last years research started to address new challenges, methods for obtaining information and computational models, for example: articulatory-based speech synthesis (Silva et al., 2016); radar based SSI using Machine Learning (Ferreira, 2021; Ferreira et al., 2022); ultrasound imaging to complement MRI (Albuquerque et al., 2021); and the extension of speech production studies to the brain, initiated with functional MRI of inner speech (Ferreira et al., 2018; Ferreira, 2020).

Acknowledgements Work reviewed was partially funded by FCT Portuguese Research Agency (mainly projects HERON, HERON II and MEMNON). Authors thank all the informants for their help and tolerance during acquisition sessions; all co-authors; first author mentors (Francisco Vaz, José Carlos Princípe and Donald Childers); PhD students (Paula Martins, Catarina Oliveira, João Freitas, Roberto Martinez, Carlos Ferreira, David Ferreira, Fábio Barros); Masters' students (Inês Carbone, Luís Silva); University of Aveiro colleagues (Augusto Silva, Francisco Vaz, Rosa Lídia Coimbra, Lurdes Moutinho, Luís Jesus). The authors also thank the Radiology Department of the Coimbra University Hospital (HUC) and IBILI, their Directors Professors Filipe Caseiro Alves and Miguel Castelo Branco and technical staff.

Funding Open access funding provided by FCT|FCCN (b-on).

Data availability Some of the developments reviewed can be made available. The authors should be contacted by interested researchers.

Declarations

Conflict of interest The authors declare that they have no conflict of interest.

References

Abreu, H. (2014). Visual speech recognition for European Portuguese. Master thesis, Universidade do Minho. Retrieved from https://hdl.handle.net/1822/37465. Accessed 23 Mar 2023.

Albuquerque, L., Valente, A.R., Barros, F., Teixeira, A., Silva, S., Martins, P., & Oliveira, C. (2021). The age effects on EP vowel production: An ultrasound pilot study. In Proc. IberSpeech.

Albuquerque, L., Valente, A. R., Barros, F., Teixeira, A., Silva, S., Martins, P., & Oliveira, C. (2022). Exploring the age effects on European Portuguese vowel production: An Ultrasound study. *Applied Sciences*. https://doi.org/10.3390/app12031396

Almeida, N., Silva, S., Teixeira, A., & Cunha, C. (2020). Collaborative quantitative analysis of RT-MRI. In: Proceedings of the 12th International Seminar on Speech Production (ISSP).

Bastos, R., & Dias, M. S. (2009). *FIRST-fast invariant to rotation and scale transform: Invariant image features for augmented reality and computer vision*. VDM Verlag.

Bay, H., Tuytelaars, T., & Gool, L.V. (2006). Surf: Speeded up robust features. In: *European conference on computer vision* (pp. 404–417). Springer.

Birkholz, P. (2013). Modeling consonant-vowel coarticulation for articulatory speech synthesis. *PLoS ONE, 8*(4), e60603.

Birkholz, P. (2022). Synthesis examples. Retrieved from https://www.vocaltractlab.de/index.php?page=vocaltractlab-examples. Accessed 23 Mar 2023.

Blandin, R., Arnela, M., Félix, S., Doc, J. B., & Birkholz, P. (2022). Efficient 3d acoustic simulation of the vocal tract by combining the multimodal method and finite elements. *IEEE Access, 10*, 69922–69938. https://doi.org/10.1109/ACCESS.2022.3187424

Bouma, G. (2003). Finite state methods for hyphenation. *Natural Language Engineering, 9*(1), 5.

Branco, A., Teixeira, A., Tomé, A., & Vaz, F. (1997). An articulatory speech synthesizer. In *Portuguese Conference on Pattern Recognition (RecPad), Univ. Coimbra, Dep. Engenharia Electrotécnica, FCTUC* (vol. 9, pp. 205–208).

Calliess, J. P., & Schultz, T. (2006). *Further investigations on unspoken speech*. Karlsruhe: Institut für Theoretische Informatik Universität Karlsruhe (TH).

Carbone, I. (2008). Segmentação do tracto vocal a partir de estudos imagiológicos de ressonância magnética. Masters dissertation, Dep Electrónica Telecomunciações e Informática, Universidade de Aveiro

Carbone, I., Martins, P., Silva, A., & Teixeira, A. (2007). Volumetric MRI acquisition and processing. *Journal of the Acoustical Society of America, 122*(5), 3030–3030.

Cooper, F. S. (1962). Speech synthesizers. In *Proceedings of the 4th International Congress of Phonetic Sciences (ICPhS'61)* (pp. 3–13).

Cunha, C., Silva, S., Teixeira, A., Oliveira, C., Martins, P., Joseph, A., & Frahm, J. (2019). On the role of oral configurations in european portuguese nasal vowels. In *Interspeech, Graz, Austria* (pp. 3332–3336). https://doi.org/10.21437/Interspeech.2019-2232

Denby, B., Csapó, T. G., & Wand, M. (2022). Future speech interfaces with sensors and machine intelligence. Retrieved from https://www.mdpi.com/journal/sensors/special_issues/FSI-SMI. Accessed 23 Mar 2023.

Denby, B., Schultz, T., Honda, K., Hueber, T., Gilbert, J. M., & Brumberg, J. S. (2010). Silent speech interfaces. *Speech Communication, 52*(4), 270–287.

Duffy, B. R. (2003). Anthropomorphism and the social robot. *Robotics and Autonomous Systems, 42*(3), 177–190.

Fan, M., & Lee, T. (2015). Variants of seeded region growing. *IET Image Processing*. https://doi.org/10.1049/iet-ipr.2014.0490

Fant, G., Liljencrants, J., & Qg, Lin. (1985). A four-parameter model of glottal flow. *STL-QPSR, 4*(1985), 1–13.

Ferreira, C. D. (2020). Functional mapping of the inner speech brain related areas. Phd thesis, Universidade de Aveiro

Ferreira, C., Direito, B., Sayal, A., Simões, M., Cadório, I., Martins, P., Lousada, M., Figueiredo, D., Castelo-Branco, M., Teixeira, A. (2018). Functional mapping of inner speech areas: A preliminary study with Portuguese speakers. In *SPECOM*.

Ferreira, D., Silva, S., Curado, F., & Teixeira, A. (2021). RaSSpeR: Radar-based Silent Speech Recognition. In *Proceedings of the Interspeech 2021*.

Ferreira, D., Silva, S., Curado, F., & Teixeira, A. (2022). Exploring silent speech interfaces based on frequency-modulated continuous-wave radar. *Sensors, 22*(2), 649.

Freitas, J., Dias, M.S., & Teixeira, A. (2012a). Towards a silent speech interface for Portuguese: Surface electromyography and the nasality challenge. In *International Conference on Bio-inspired Systems and Signal Processing (BIOSIGNALS 2012), Vilamoura, Portugal*.

Freitas, J., Dias, M. S., Teixeira, A. (2014a). Can ultrasonic doppler help detecting nasality for silent speech interfaces? An exploratory analysis based on alignement of the doppler signal with velum aperture information from Real-Time MRI. In *Proceedings of PhyCS*.

Freitas, J., Ferreira, A., Figueiredo, M., Teixeira, A., Dias, M. S. (2014b). Enhancing multimodal silent speech interfaces with feature selection. In *Proceedings of the InterSpeech*.

Freitas, J., Teixeira, A., & Dias, M. S. (2013). Multimodal silent speech interface based on video, depth, surface electromyography and ultrasonic doppler: Data collection and first recognition results. In *Workshop on Speech Production in Automatic Speech Recognition, Lyon*.

Freitas, J., Teixeira, A., Dias, M. S., & Bastos, C. A. C. (2011). Towards a multimodal silent speech interface for European Portuguese. In *Speech Technologies, INTECH*.

Freitas, J., Teixeira, A., Dias, M. S. (2014c). Multimodal corpora for Silent Speech Interaction. In *Proceedings of the LREC, Reykjavik, Iceland*.

Freitas, J., Teixeira, A., Dias, M. S., & Silva, S. (2016). *An introduction to silent speech interfaces*. Springer.

Freitas, J., Teixeira, A., Silva, S., Oliveira, C., & Dias, M. S. (2015). Detecting nasal vowels in speech interfaces based on surface electromyography. *PLoS ONE, 10*, e0127040.

Freitas, J., Teixeira, A., Vaz, F., & Dias, M. S. (2012b). Automatic speech recognition based on ultrasonic doppler sensing for european Portuguese. In *Advances in Speech and Language Technologies for Iberian Languages* (vol. CCIS 328). Springer

Freixes, M., Arnela, M., Socoró, J. C., Alías, F., & Guasch, O. (2019). Glottal source contribution to higher order modes in the finite element synthesis of vowels. *Applied Sciences*. https://doi.org/10.3390/app9214535

Gonzalez, J. A., Cheah, L. A., Gilbert, J. M., Bai, J., Ell, S. R., Green, P. D., & Moore, R. K. (2016). A silent speech system based on permanent magnet articulography and direct synthesis. *Computer Speech & Language, 39*, 67–87.

Hernáez Rioja, I., González López, J. A., Navas, E., Pérez Córdoba, J. L., Saratxaga, I., Olivares, G., Sanchez, J., Galdón, A., García Romillo, V., Gónzalez Atienza, M., Schultz, T., Green, P. D., Wand, M., Marxer, R., & Diener, L. (2021). Voice restoration with silent speech interfaces (ReSSInt). In IberSPEECH, ISCA.

Hueber, T., Chollet, G., Denby, B., Dreyfus, G., & Stone, M. (2008). Phone recognition from ultrasound and optical video sequences for a silent speech interface. In Ninth Annual Conference of the International Speech Communication Association.

Jackson, P. J., & Singampalli, V. D. (2009). Statistical identification of articulation constraints in the production of speech. *Speech Communication, 51*(8), 695–710.

Jin, Y., Gao, Y., Xu, X., Choi, S., Li, J., Liu, F., Li, Z., & Jin, Z. (2022). Earcommand: "Hearing" your silent speech commands in ear. In *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* (vol. 6, issue no. 2). https://doi.org/10.1145/3534613,

Ke, Y., & Sukthankar, R. (2004). Pca-sift: A more distinctive representation for local image descriptors. In *Proceedings of the 2004 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2004. CVPR 2004* (vol. 2, pp. II–II). IEEE.

Kimura, N., Kono, M., & Rekimoto, J. (2019). Sottovoce: An ultrasound imaging-based silent speech interaction using deep neural networks. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems, Association for Computing Machinery, New York, NY, USA, CHI* (vol. '19, pp. 1–11). https://doi.org/10.1145/3290605.3300376

Kochetov, A., Savariaux, C., Lamalle, L., Noûs, C., & Badin, P. (2020) An MRI-based articulatory characterization of Kannada coronal consonant contrasts. Retrieved from https://hal.science/hal-03031319, working paper or preprint. Accessed 23 Mar 2023.

Kröger, B. J., & Birkholz, P. (2009). Articulatory synthesis of speech and singing: State of the art and suggestions for future research. In *Multimodal Signals: Cognitive and Algorithmic Issues* (pp. 306–319).

Krug, P. K., Stone, S., Birkholz, P. (2021) Intelligibility and naturalness of articulatory synthesis with VocalTractLab compared to established speech synthesis technologies. In *Proceedings of the 11th ISCA Speech Synthesis Workshop (SSW 11)* (pp. 102–107). https://doi.org/10.21437/SSW.2021-18

Lee, S., & Seo, J. (2019) Word error rate comparison between single and double radar solutions for silent speech recognition. In *2019 19th International Conference on Control, Automation and Systems (ICCAS)* (pp. 1211–1214). https://doi.org/10.23919/ICCAS47443.2019.8971653

Levelt, W. J. (1993). *Speaking: From intention to articulation*. MIT press.

Lim, Y., Toutios, A., Bliesener, Y., Tian, Y., Lingala, S. G., Vaz, C., Sorensen, T., Oh, M., Harper, S., Chen, W., Lee, Y., Töger, J., Monteserin, M. L., Smith, C., Godinez, B., Goldstein, L., Byrd, D., Nayak, K. S., & Narayanan, S. (2021). A multispeaker dataset of raw and reconstructed speech production real-time MRI video and 3d volumetric images. *Scientific Data, 8*(1), 1–14. https://doi.org/10.1038/s41597-021-00976-x

Linguateca. (2008). CETEMPúblico (Corpus de Extractos de Textos Electrónicos MCT/Público). Retrieved July 18, 2022, from https://www.linguateca.pt/CETEMPublico/

Lowe, D. G. (2004). Distinctive image features from scale-invariant keypoints. *International Journal of Computer Vision, 60*(2), 91–110.

Martins, P., Carbone, I., Silva, A., & Teixeira, A. (2007). An MRI study of European Portuguese nasals. In *Interspeech*.

Martins, P., Carbone, I., Silva, A., & Teixeira, A. (2008). European Portuguese MRI based speech production studies. *Speech Communication, 50*, 925–952.

Martins, P., Oliveira, C., Silva, A., & Teixeira, A. (2010). Articulatory characteristics of European Portuguese laterals: A 2D & 3D MRI study. In *FALA 2010*.

Martins, P., Oliveira, C., Silva, S., & Teixeira, A. (2012a). Velar movement in European Portuguese nasal vowels. In *Proceedings of IberSPEECH 2012—VII Jornadas en Tecnología del Habla and III Iberian SLTech Workshop, Madrid, Spain*.

Martins, P., Silva, S., Oliveira, C., Ferreira, C., Silva, A., & Teixeira, A. (2012). Polygonal mesh comparison applied to the study of European Portuguese sounds. *International Journal of Creative Interfaces and Computer Graphics, 3*, 28.

Martins, P., Silva, S., Oliveira, C., Silva, A., & Teixeira, A. (2011). Investigating the differences between European Portuguese sounds: An approach using polygonal mesh comparison. In *Proceedings of the SIACG, Faro, Portugal*.

Mateus, M. H., & d'Andrade, E. (2000). *The phonology of Portuguese*. OUP Oxford.

Mermelstein, P. (1973). Articulatory model for the study of speech production. *The Journal of the Acoustical Society of America, 53*(4), 1070–1082.

Mohd Shariff, K. K., Nadiah Yusni, A., Md Ali, M. A., Syahirul Amin Megat Ali, M., Megat Tajuddin, M. Z., & Younis, M. A. A. (2022) Cw radar based silent speech interface using CNN. In *2022 IEEE Symposium on Wireless Technology & Applications (ISWTA)* (pp. 76–81). https://doi.org/10.1109/ISWTA55313.2022.9942730

Nam, H., Browman, C., Goldstein, L., Proctor, M., Rubin, P., & Saltzman, E. (2001). Tada: Task dynamic model of inter-articulator speech coordination, version 0.9782. Retrieved July 20, 2022 from, https://haskinslabs.org/about-us/features-and-demos/tada-task-dynamic-model-inter-articulator-speech-coordination

Nam, H., Goldstein, L., Browman, C., Rubin, P., Proctor, M., & Saltzman, E. (2006). *TADA (TAsk Dynamics Application) manual*.

Nascimento, F., Marques, L., & Segura, L. (1987). *Português fundamental: métodos e documentos*. Tomo I e II Lisboa: Instituto de Investigação Científica, Centro de Lingüística da Universidade dc Lisboa.

Oliveira, C. (2009). From grapheme to gesture. Linguistic contributions for an articulatory based text-to-speech system. Ph.d. thesis, University of Aveiro

Oliveira, C., & Teixeira, A. (2007) On gestures timing in European Portuguese nasals. In *ICPhS* (pp. 405–408).

Oliveira, C., de Castro Moutinho, L., Teixeira, A. (2005a). On automatic European Portuguese syllabification. In *III Congreso de Fonética Experimental, Universidade de Santiago de Compostela, Espanha*.

Oliveira, C., de Castro Moutinho, L., & Teixeira, A. (2005b). On European Portuguese automatic syllabification. In *InterSpeech, L2F/ISCA, Lisboa, Portugal*.

Oliveira, C., Martins, P., Silva, S., & Teixeira, A. (2012). An MRI study of the oral articulation of European Portuguese nasal vowels. In *13th Annual Conference of the International Speech Communication Association (InterSpeech), Portland, USA*.

Oliveira, C., Paiva, S., de Castro Moutinho, L., & Teixeira, A. (2004). Um novo sistema de conversação grafema-fone para o português europeu baseado em transdutores. In *II Congresso Internacional de Fonética e Fonologia*.

Prado, P. (1991). A target-based articulatory synthesizer. Phd thesis, University of Florida

Rao, R., & Mersereau, R. M. (1994). Lip modeling for visual speech recognition. In *Proceedings of 1994 28th Asilomar Conference on Signals, Systems and Computers* (vol 1, pp. 587–590). IEEE.

Rubin, P., Saltzman, E., Goldstein, L., McGowan, R., Tiede, M., & Browman, C. (1996). Casy and extensions to the task-dynamic model. In *1st ETRW on Speech Production Modeling: From Control Strategies to Acoustics; 4th Speech Production Seminar: Models and Data*.

Ruthven, M., Miquel, M. E., & King, A. P. (2021). Deep-learning-based segmentation of the vocal tract and articulators in real-time magnetic resonance images of speech. *Computer Methods and Programs in Biomedicine, 198*, 105814. https://doi.org/10.1016/j.cmpb.2020.105814

Ruthven, M., Miquel, M. E., & King, A. P. (2023). A segmentation-informed deep learning framework to register dynamic two-dimensional magnetic resonance images of the vocal tract during speech. *Biomedical Signal Processing and Control, 80*, 104290. https://doi.org/10.1016/j.bspc.2022.104290

Saltzman, E. L., & Munhall, K. G. (1989). A dynamical approach to gestural patterning in speech production. *Ecological Psychology, 1*(4), 333–382.

Sampson, R. (1999). *Nasal vowel evolution in Romance*. Oxford linguistics, Oxford University Press.

Schroeder, M. R. (1999). *Computer speech: Recognition, compression, synthesis* (Vol. 35). Springer.

Schultz, T., & Wand, M. (2010). Modeling coarticulation in EMG-based continuous speech recognition. *Speech Communication, 52*(4), 341–353.

Silva, S., & Teixeira, A. (2013) AAM based vocal tract segmentation from Real-Time MRI image sequences. In *Proceedings of the RecPad 2013*.

Silva, S., & Teixeira, A. (2014). A framework for analysis of the upper airway from real-time MRI sequences. In *Proceedings of the Visualization and Data Analysis (VDA 2014)*. SPIE.

Silva, S., & Teixeira, A. (2015). Unsupervised segmentation of the vocal tract from real-time MRI sequences. *Computer Speech and Language*. https://doi.org/10.1016/j.csl.2014.12.003

Silva, S., & Teixeira, A. (2016). Quantitative systematic analysis of vocal tract data. *Computer Speech and Language, 36*, 307–329.

Silva, S., & Teixeira, A. (2017a). An anthropomorphic perspective for audiovisual speech synthesis. In *Proceedings of the BIOSIGNALS*.

Silva, S., & Teixeira, A. (2017b). Critical articulators identification from RT-MRI of the vocal tract. In *Proceedings of the Interspeech 2017, Stocholm, Sweden*.

Silva, S., Teixeira, A., Oliveira, C., & Martins, P. (2013). Segmentation and analysis of vocal tract from midsagittal Real-Time MRI. In *Proceedings of the ICIAR 2013, vol. SPRINGER LNCS 7950* (pp. 459–466).

Silva, S., Teixeira, A., & Orvalho, V. (2016). Articulatory-based audiovisual speech synthesis: Proof of concept for European Portuguese. In *Proceedings of the IberSPEECH, Lisboa*.

Silva, S., Almeida, N., Cunha, C., Joseph, A., Frahm, J., & Teixeira, A. (2020). Data-driven critical tract variable determination for European Portuguese. *Information*. https://doi.org/10.3390/info11100491

Silva, S., Cunha, C., Teixeira, A., Joseph, A., & Frahm, J. (2020b). Towards automatic determination of critical gestures for European Portuguese sounds. In *International Conference on Computational Processing of the Portuguese Language* (pp. 3–12). Springer

Silva, L. N., Teixeira, A., & Santos, B. S. (2002). Visualization of articulatory and acoustic information on an articulatory synthesizer. In *Portuguese Conference on Pattern Recognition (RecPad), IEETA, Universidade de Aveiro*.

Silva, S. S., Teixeira, A. J., Cunha, C., Almeida, N., Joseph, A. A., & Frahm, J. (2019). Exploring critical articulator identification from 50hz RT-MRI data of the vocal tract. In *INTERSPEECH* (pp. 874–878)

Srinivasan, S., Raj, B., & Ezzat, T. (2010). Ultrasonic sensing for robust speech recognition. In *2010 IEEE International Conference on Acoustics, Speech and Signal Processing* (pp. 5102–5105). IEEE.

Stone, S., Azgin, A., Mänz, S., & Birkholz, P. (2020). Prospects of articulatory text-to-speech synthesis. In *International Seminar on Speech Production (ISSP)*.

Story, B. H. (2019). History of speech synthesis. In W. F. Katz & P. F. Assmann (Eds.), *The Routledge Handbook of Phonetics* (pp. 9–33). Routledge.

Teixeira, A. (2000). Síntese articulatória das vogais nasais do Português Europeu. Phd thesis, Universidade de Aveiro.

Teixeira, A., & Vaz, F. (2000a), Articulatory synthesis: The use of biological models in production of high quality speech. In *Congresso Português de Engenharia Biomédica (BioEng'2000), Coimbra* (vol. 5).

Teixeira, A., & Vaz, F. (2000b). Síntese articulatória dos sons nasais do Português. In Anais do V Encontro para o Processamento Computacional da Língua Portuguesa Escrita e Falada (PROPOR, (Ed.), *das Graças Volpe Nunes M* (pp. 183–193). Atibaia.

Teixeira, A., & Vaz, F. (2001). European Portuguese Nasal Vowels: An EMMA study. In *7th European Conference on Speech Communication and Technology, EuroSpeech—Scandinavia, CPK/ISCA, Aalborg, Dinamarca* (vol. 2, pp. 1843–1846).

Teixeira, A., de Lima, V. S., Caldas de Oliveira, L., & Quaresma, P. (Eds.). (2008). *Computational Processing of the Portuguese Language Lecture Notes in Artificial Intelligence, LNAI* (Vol. 5190). Springer.

Teixeira, A., Jesus, L. M. T., & Martinez, R. (2003). Adding fricatives to the Portuguese articulatory synthesiser. In *8th European Conference on Speech Communication and Technology (Eurospeech)* (pp. 2949–2952). IDIAP/ISCA.

Teixeira, A., Martinez, R., Silva, L., Jesus, L. M. T., & Vaz, F. (2004). Articulatory synthesis of Portuguese. In *The International Workshop Dedicated to the Memory of Farkas Kempelen (Wolfgang von Kempelen), Budapeste*.

Teixeira, A., Martinez, R., Silva, L., Jesus, L., Príncipe, J. C., & Vaz, F. (2005). Simulation of human speech production applied to the study and synthesis of European Portuguese. *EURASIP Journal of Applied Signal Processing, Special Issue on Anthropomorphic Proc of Audio and Speech, 9*, 1435–1448.

Teixeira, A., Martins, P., Oliveira, C., Ferreira, C., Silva, A., & Shosted, R. (2012a). Real-time MRI for Portuguese. In *Computational Processing of the Portuguese Language, PROPOR 2012, Lecture Notes in Computer Science/LNAI* (Vol. 7243).

Teixeira, A., Martins, P., Oliveira, C., & Silva, A. (2012b). Production and modeling of the European Portuguese palatal lateral. In *Computational Processing of the Portuguese Language, PROPOR 2012, Lecture Notes in Computer Science/LNAI* (Vol. 7243).

Teixeira, A., Oliveira, C., & Barbosa, P. (2008b). European Portuguese articulatory based text-to-speech: First results. In *Computational Processing of the Portuguese Language, The International Conference on Computational Processing of Portuguese, PROPOR 2008, Lecture Notes in Computer Science/LNAI* (Vol. 5190). Springer.

Teixeira, A., Oliveira, C., & Moutinho, L. (2006). On the use of machine learning and syllable information in european Portuguese grapheme-phone conversion. In Vieira, R., Quaresma, P., das Graças Volpe Nunes, M., Mamede, N. J., Oliveira, C., & Dias, M. C. (Eds.) *Computational Processing of the Portuguese Language, The International Conference on Computational Processing of Portuguese, PROPOR 2006, Lecture Notes in Computer Science/LNAI, Vol. 3960, Springer Verlag, Itatiaia, RJ, Brasil, no. LNAI 3960 in Lecture Notes in Artificial Intelligence* (pp. 212–215).

Teixeira, A., Silva, L., Martinez, R., & Vaz, F. (2002). Sapwindows—towards a versatile modular articulatory synthesizer. In *IEEE-SP Workshop on Speech Synthesis, Santa Mónica, CA, E. U. A.*

Teixeira, A., Vaz, F., & Príncipe, J. C. (1997a). A Software Tool to Study Portuguese Vowels. In *5th European Conference on Speech Communication and Technology (Eurospeech'97), Ródes, Grécia* (vol. 5, pp. 2543–2546).

Teixeira, A., Vaz, F., & Príncipe, J. C. (1998a). A comprehensive nasal model for a frequency domain articulatory synthesis. In Muge, F., Pinto, R. C., & Piedade, M. (Eds.) *Portuguese Conference on Pattern Recognition (RecPad), APRP, IST, Lisboa*, (vol. 10, pp. 333–338).

Teixeira, A., Vaz, F., Príncipe, J. C. (1998b). Some studies of European Portuguese nasal vowels using an articulatory synthesizer. In *5th IEEE International Conference on Electronics, Circuits and Systems (ICECS Lx98), Instituto Superior Técnico, Lisboa, Portugal* (vol .3, pp. 507–510).

Teixeira, A., Vaz, F., Príncipe, J. C., & Childers, D. G. (1997b). Articulatory synthesis of Portuguese vocoids. In *Portuguese Conference on Pattern Recognition (RecPad), Univ. Coimbra, Dep. Engenharia Electrotécnica, FCTUC* (vol. 9, pp. 219–224).

Teixeira, A., Vitor, N., Freitas, J., & Silva, S. (2017). Silent speech interaction for ambient assisted living scenarios. In *Proceedings of the HCI International*.

Teixeira, V., Pires, C., Pinto, F., Freitas, J., Dias, M. S., & Rodrigues, E. M. (2012c). Towards elderly social integration using a multimodal human-computer interface. In *Proceedings of the International Living Usability Lab Workshop on AAL Latest Solutions, Trends and Applications*. AAL.

Toth, A. R., Kalgaonkar, K., Raj, B., & Ezzat, T. (2010). Synthesizing speech from doppler signals. In *2010 IEEE International Conference on Acoustics, Speech and Signal Processing* (pp. 4638–4641). IEEE.

Tóth, L., & Shandiz, A. H. (2020) 3D Convolutional Neural Networks for Ultrasound-Based Silent Speech Interfaces. In *International Conference on Artificial Intelligence and Soft Computing* (pp. 159–169). Springer.

Tran, V. A., Bailly, G., Lœvenbruck, H., & Toda, T. (2010). Improvement to a nam-captured whisper-to-speech system. *Speech Communication, 52*(4), 314–326.

Wang, J., Hou, Q., Liu, N., & Zhang, S. (2015) Model of human visual cortex inspired computational models for visual recognition. In *2015 IEEE International Conference on Multimedia Big Data* (pp. 88–91). https://doi.org/10.1109/BigMM.2015.29

Xu, C., Li, Z., Zhang, H., Rathore, A. S., Li, H., Song, C., Wang, K., & Xu, W. (2019). Waveear: Exploring a mmwave-based noise-resistant speech sensing for voice-user interface. In *Proceedings of the*

17th Annual International Conference on Mobile Systems, Applications, and Services, Association for Computing Machinery, New York, NY, USA, MobiSys (vol. 19, pp. 14–26) https://doi.org/10.1145/3307334.3326073

Yu, W., Zeiler, S., & Kolossa, D. (2022). Reliability-based large-vocabulary audio-visual speech recognition. *Sensors*. https://doi.org/10.3390/s22155501

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.