Robust phoneme recognition for a speech therapy environment

Conference Paper · January 2016

CITATIONS
3

READS
134

4 authors, including:

Isabel Cristina Ramos Peixoto Guimarães
Escola Superior de Saude do Alcoitão
115 PUBLICATIONS **762** CITATIONS
SEE PROFILE

João Magalhães
Universidade NOVA de Lisboa
125 PUBLICATIONS **716** CITATIONS
SEE PROFILE

Sofia Cavaco
Universidade NOVA de Lisboa
55 PUBLICATIONS **206** CITATIONS
SEE PROFILE

Robust phoneme recognition for a speech therapy environment

André Grossinho*, Isabel Guimarães†, João Magalhães* and Sofia Cavaco*
*NOVA LINCS, Department of Computer Science
Faculty of Science and Technology, Universidade Nova de Lisboa
2829-516 Caparica, Portugal
†Escola Superior de Saúde do Alcoitão
Rua Conde Barão, Alcoitão, 2649-506 Alcabideche, Portugal
Email: a.grossinho@campus.fct.unl.pt, iguimaraes@essa.pt, {jm.magalhaes, scavaco}@fct.unl.pt

***Abstract*—Traditional speech therapy approaches for speech sound disorders have a lot of advantages to gain from computer-based therapy systems. With speech recognition techniques the motivation elements of these systems can be automated in order to get an interactive environment that motivates the therapy attendee towards better performances. Here we propose a robust phoneme recognition solution for an interactive environment for speech therapy. We compare the results of hierarchical and flat classification, with naive Bayes, support vector machines and kernel density estimation on linear predictive coding coefficients and Mel-frequency cepstral coefficients.**

***Index Terms*—Speech Therapy, Phoneme Detection, Kernel Density Estimation, Naive Bayes, Support Vector Machines.**

I. Introduction

Speech sound disorders (SSD) of many different types and severities are very common during childhood. As reported by Guimarães et. al [1] for data on European Portuguese (EP), 8.8% of preschool-aged children show SSD. Children with these problems can have difficulty to correctly express themselves, which may affect their related quality of life, and in more severe cases, it can affect the natural development of social skills.

Speech therapy can (and should) be used to address these disorders. Traditional speech therapy approaches for speech sound disorders have many advantages to gain from computer-based therapy systems. Some computer tools for speech therapy exist and many such as SpeechViewer [2] and Box of Tricks [3] are described as already well established [4].

Existing speech therapy tools provide different approaches in aiding speech therapy. Many complex systems, such as TERAPERS (for Romanian) and Ortho-Logo-Paedia (OLP), focus on providing high-quality speech therapy aids and therapy exercises in different forms [5], [6].

Other systems take a less comprehensive approach and target specific aspects of the therapy by providing exercises in a more fun and entertaining way, like ARticulation TUtoR (ARTUR) or the Comunica project [4], [7]. ARTUR uses a virtual tutor, which when needed provides the person with vocal tract animations. The virtual tutor approach makes human computer interaction more natural, as described in [8]. The Comunica [7] framework uses automatic speech recognition to analyse the children's vocalizations (in Spanish) and provide feedback.

For EP there are also some computer aids to speech therapy like a Game for Vowel Training [9], or the Lisling 3D [10] and VITHEA [11]. The Interactive Game for Vowel Training is a simple car racing game where the car actions are controlled by uttering 5 vowels. In Lisling 3D tasks, such as writing or selecting words, are given to patients throughout a virtual 3D environment. The VITHEA system is an online platform where people with aphasia can do exercises from a browser. Exercises are guided with a virtual tutor. An automatic word naming recognition module evaluates the patients responses and provides feedback.

VisualSpeech is an interactive environment for speech therapy for children with SSD [12]. The main novelty of this work is the integration of visual-feedback with gamification components. By combining visual-feedback with adapted traditional speech sound exercises, it is possible to create an environment with motivation focused elements that can improve children's performance and engagement in speech therapy sessions.

VisualSpeech addresses the first stage of speech therapy: phoneme productions. Since languages are different and have a different number and combinations of sounds, computer tools are language bounded. VisualSpeech was designed for European Portuguese (EP).

This environment includes motivational elements that aim at keeping the child motivated and focused in the therapy exercises. In particular, the environment has a performance bar which indicates how well the child is performing in the exercise (ice cream in figure 1). These elements also provide useful feedback to the child, who can see changes in the environment that depend on his performance. With a bar that increases or decreases according to the child's performance, the child can be encouraged to outperform his last speech productions.

Since controlling the performance bar is a new task that requires the speech and language therapists' (SLT) care and may divert their attention, this process should have the option of being automatic. For that reason, while we do not want to substitute the SLTs, and the final decision should

Fig. 1. VisualSpeech motivational elements. Progress bar (ice cream) and reward (virtual button).

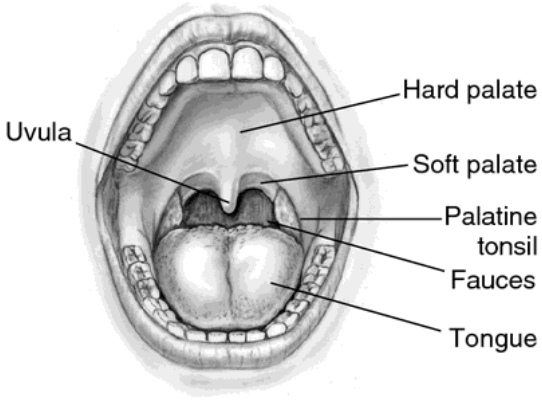


Fig. 2. Regions of the mouth.

be theirs, we are investigating ways of making the progress bar advance automatically. Speech recognition techniques can provide valuable feedback to the SLTs, suggesting when to advance the progress bar. To achieve this, we explored robust speech recognition techniques with EP phonemes (sections II).

In this paper, we propose a robust phoneme recognition solution for analyzing the child's performance in VisualSpeech. We have focused our attention in exercises that allow a robust interaction without false negatives that could be the cause of frustration of the child. While it is our intention to expand this work so as to cover all EP phonemes, we started by giving special attention to the *a, e, i, o, u* vowels and the EP rhotic consonant sound.

II. Scoring Speech Exercises

VisualSpeech focuses on phoneme productions during speech exercises. In order to have the scores computed automatically, the environment needs to perform phoneme recognition. Below we discuss our approach to robust phoneme recognition: The first step consist of extracting audio features from the speech productions, (section II-B), while the second step consists of using those features in a classification algorithm (section II-C).

A. Speech exercises

As mentioned above, the proposed environment addresses phoneme productions. Phonemes do not necessarily represent sounds of letters, since not only a letter can have different sounds in different contexts (words), but also in some cases a sole letter does not represent anything. To accurately represent phonemes in written language, the International Phonetic Alphabet (IPA) is used [13].

While it is our intention to expand this work so as to cover all EP phonemes, we started by giving special attention to the *a, e, i, o, u* vowels. In EP these correspond to the phonemes /a/, /ɛ/, /i/, /ɔ/, /u/.

Following the suggestion of SLTs, we also addressed the EP rhotic consonant sound, that is, the sound of *R* at the beginning of a word, like in *rato* (mouse) or double *R* in the middle of words, like in *carro* (car). This sound is of particular interest because of the accent variations and because often it is one of the last sounds to be mastered in childhood [14].

Rennicke and Martins report that this consonant can be a voiced uvular fricative (/ʁ/), or, less commonly, a voiceless uvular fricative (/χ/). The uvular fricatives are done with the back of the tongue against the uvula (figure 2[1]). They can be voiced if there is vibration of the vocal cords, and voiceless otherwise. This consonant can also be an alveolar trill (/ɾ/), which is made with vibrations of the tip of the tong against the upper alveolar ridge (between the teeth and the hard palate) for longer than two or three periods. When produced in this way, the consonant sounds like the double *R* in the Spanish word *perro* (dog). Another variation is done by a vibration of the palatine uvula, in which case it is known as uvular trill (/ʀ/) and it can be stronger or weaker depending on the vibration. Finally, another not very common variation is the voiceless velar fricative (/x/), which is done with the back of the tongue against the soft palate. Since the voiced uvular fricative is one of the most common pronunciations, we will use the symbol /ʁ/ when denoting the EP rhotic consonant in general.

The /ʁ/ sounds included in our study are the following: /ʁa/, /ʁɐ/, /ʁə/, /ʁo/ and /ʁu/. More details about these sounds are given in section III-A.

B. Audio features

We extracted two types of features from the speech productions: linear predictive coding (LPC) coefficients and Mel-frequency cepstral coefficients (MFCC).

LPC uses a linear predictive model to estimate the spectral envelope of speech signals. This method assumes that speech sound results from the vocal tract as an all-pole filter, that is applied to the larynx vibrations. This approach tries to predict the current window of a sample as a linear combination of the past windows while minimizing the error. The goal of LPC is to get p coefficients of the p linear equations that minimize the prediction error. Using these coefficients, the formants can be

[1]From http://medical-dictionary.thefreedictionary.com

estimated. These are representations of the acoustic resonance of the human vocal tract. The number of poles used affects the number of formants that can be estimated with LPC. The ideal number of poles varies according to the speakers gender, age, and sampling rate of the audio file.

Although the human ear can hear a wide range of frequencies (20Hz to 20 kHz) our auditory system filters the spectrum, giving more importance to some frequency regions than others. These filters are not uniformly spaced, and our ears have more filters in the lower pitch region of the spectrum and less on the higher pitch region.

The Mel frequency cepstrum (MFC) is a short-time representation of the sound's spectrum that uses the Mel scale: a nonlinear frequency scale of triangular filters for the frequencies in order to approximate the human hear. The coefficients, that is the MFCCs, are time-varying functions. When given a windowed input signal, a filter-bank of n triangular filters is applied and the average spectrum around the center frequency computed. The resulting features, that is the n MFCCs, are cepstral arrays. Since the MFCCs are time-varying, we used the mean of each MFCC to train the learning algorithms described in section II-C. In other words, our feature vectors are vectors of n mean values.

C. Phoneme recognition

In order to score the phoneme productions we compared the performance of three algorithms: the Naive Bayes (NB), Support Vector Machines (SVM) and Kernel Density Estimation (KDE) [15].

The Naive Bayes is a generative classifier that applies the Bayes probability theorem. The NB classifier is well suited when the data dimensionality is high and there is a strong independence among the dimensions. NB estimates the probability of a phoneme label by modeling each dimension independently of the others given the class label (this is the conditional independence assumption). The phoneme sample is then classified with the label that maximizes the sample likelihood.

The SVM is a popular technique, that classifies a sample into one of two classes. A discriminating hyperplane is learned from the training data by selecting samples as support vectors. These support vectors define the hyperplane that maximizes the margin between the two classes. When a new sample is up for classification, this technique projects the sample onto the hyperplane and decides the class of the sample.

The KDE is an approach to estimate the true probability density distribution from the training data. This method uses the entire training set to compute a smoothed estimate of the true probability density distribution. It applies a Kernel function to every point of the training set to compute the contribution of every training sample. This Kernel function is usually a standard probability distribution function. For a matter of convenience, we will use a Gaussian Kernel, $K(z) = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}z^2}$.

To estimate the density function on a given test point x, the aggregated contributions of all training samples correspond to

$$\hat{f}(x) = \frac{1}{n} \sum_{i=1}^{n} K(\frac{x - x_i}{h}).$$

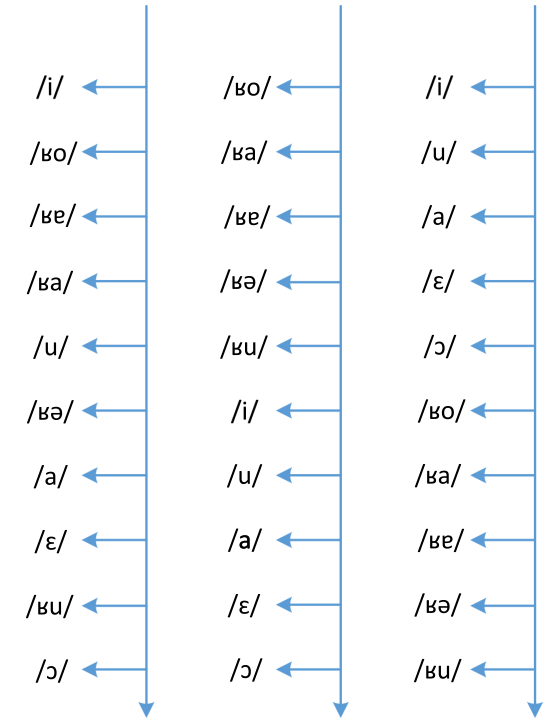
In practice the x variable correspond to the MFCCs and LPCs feature vector. The h parameter is the kernel bandwidth is estimated from the data by some method. We followed the Silverman's rule of thumb and used the data variance σ to set the kernel bandwidth to $h = (\frac{4\sigma^5}{3n})^{1/5}$.

Formally, we have one function $\hat{f_{l_j}}(x)$ for each label l_j of our problem, where $j = 1, ..., L$. In our case, we will have a function $\hat{f_{l_j}}(x)$ for each training phoneme sample. Thus, each training sample contributes exclusively to the density function of its own label.

It is now straightforward to address the multi-class nature of phoneme detection. Using Bayes' theorem we can merge all individual density estimates $\hat{f_{l_j}}(x)$ with $j = 1, ..., L$:

$$p(l = l_j | X = x_0) = \frac{\pi_{l_j} \hat{f_{l_j}}(x_0)}{\sum_{i=1}^{L} \pi_i \hat{f_{l_i}}(x_0)},$$

where π_i corresponds to the label i prior. This definition allow us to compute the probability of one speech sample x_0 corresponding to a certain phoneme label l_j. To classify test phonemes, we only need to find the label l_j that maximizes the above expression.



(a) Hierachy 1. (b) Hierachy 2. (c) Hierachy 3.

Fig. 3. Hierarchical classifiers.

	NB	SVM	KDE	KDE Silverman
MFCC 9	82.51	87.93	91.73	89.35
MFCC 10	83.65	88.12	91.73	89.54
MFCC 11	82.89	88.02	91.73	88.78
MFCC 12	81.94	88.02	**91.83**	89.26
MFCC 13	80.52	87.26	91.54	88.59
…	…	…	…	…
MFCC 21	70.91	85.36	91.25	84.03
MFCC 22	69.87	85.08	91.44	83.65
MFCC 23	67.97	84.79	91.35	83.84
LPC 22	57.13	48.19	0.481939	N/A
LPC 24	58.27	48.29	0.476236	N/A

TABLE I
ACCURACY RESULTS OF HIERARCHY 1

Feature	NB	SVM	KDE	KDE Silverman
MFCC 9	82.60	87.83	92.02	90.40
MFCC 10	83.84	88.02	**92.11**	90.21
MFCC 11	83.08	87.83	92.02	89.64
MFCC 12	82.03	87.93	92.02	89.83
MFCC 13	81.46	87.17	91.73	89.26
…	…	…	…	…
MFCC 21	74.62	85.27	91.54	84.79
MFCC 22	73.67	84.98	91.73	84.32
MFCC 23	73.48	84.70	91.64	84.51
LPC 22	59.98	48.19	53.42	N/A
LPC 24	60.17	48.29	52.95	N/A

TABLE II
ACCURACY RESULTS OF HIERARCHY 2

Feature	NB	SVM	KDE	KDE Silverman
MFCC 9	82.60	87.83	92.02	90.40
MFCC 10	83.84	88.02	**92.11**	90.21
MFCC 11	83.08	87.93	92.02	89.64
MFCC 12	82.04	87.93	92.02	89.83
MFCC 13	81.46	87.17	91.73	89.26
…	…	…	…	…
MFCC 21	74.62	85.36	91.54	84.79
MFCC 22	73.67	85.08	91.73	84.32
MFCC 23	73.48	84.79	91.64	84.51
LPC 22	52.85	48.19	22.81	N/A
LPC 24	53.80	48.29	22.15	N/A

TABLE III
ACCURACY RESULTS OF HIERARCHY 3

Feature	KDE	KDE Silverman
MFCC 9	92.59	93.35
MFCC 10	92.68	94.68
MFCC 11	93.06	94.39
MFCC 12	93.35	**95.06**
MFCC 13	93.25	94.68
…	…	…
MFCC 21	93.82	92.59
MFCC 22	94.11	92.21
MFCC 23	94.39	91.54
LPC 22	44.20	N/A
LPC 23	41.92	N/A
LPC 24	41.73	N/A

TABLE IV
ACCURACY RESULTS WITH THE FLAT APPROACH.

D. Hierarchical vs Flat classifiers

To achieve phoneme scoring the features and classifiers used in phoneme recognition need to be able to detect all meaningful speech sounds. To this purpose two different approaches were experimented, the hierarchical and the flat. The former aims at classifying in binary in-line fashion, and the latter all possibilities concurrently. Since in our experiments we used ten different speech sounds the number of possible combinations for the hierarchical approach are far too many to try. We selected three different hierarchies based on the accuracies obtained in preliminary tests: hierarchy 1 orders from best to worse accuracies across both vowels and phonemes; hierarchy 2 does the same but distinguishes vowels from phonemes; and hierarchy 3 does the same as hierarchy 2 but in reverse order. All three hierarchies are illustrated in figure 3, they are sorted from best to worst accuracy to minimize misclassification along the binary decisions of each hierarchy level.

III. EVALUATION

To compare the performance of the different approaches discussed in the previous section (the flat and the hierarchical approach with each of the three classification algorithms discussed and the two types of features) we used two data sets of speech productions. The results are discussed below.

A. Phoneme Data

In order to address phoneme recognition, we started with a data set with the EP vowel phonemes /a/, /ɛ/, /i/, /ɔ/, /u/ [16]. These samples were recorded from 44 different speakers: 27 child speakers, 11 female adult speakers and 6 male adult speakers. The data set contains a total of 220 manually segmented samples, that is, 44 samples for each phoneme. (For more details, please see [16].)

Following the suggestion of SLTs, we also created a data set with the uvular sonorant /ʁ/. This data set was created from 67 audio recordings performed at Escola Superior de Saúde do Alcoitão (ESSA). For these recordings the 67 participants read an EP version of a phonetic balanced short tale 'The story of Arthur the Rat' with six words that include the sound /ʁ/ [17].

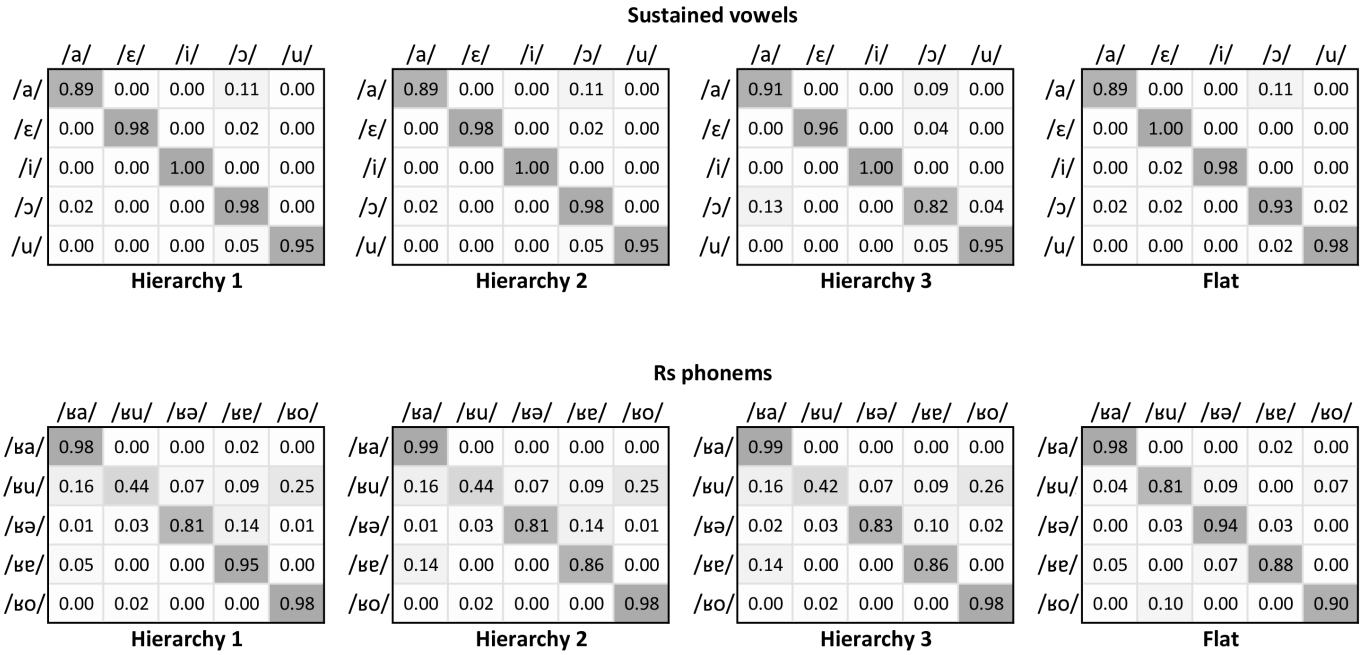


Fig. 4. Phonemes detection confusion matrices.

65 adult female speakers and 2 male adult male speakers participated in the recordings. Most were university students but there were also some voice professionals participating in these recordings. The participants were from different regions of Portugal, which means that the recordings include different accents.

To create the data set, we manually extracted the section containing the uvular /ʁ/ followed by a vowel from the six words with this sound:

- /ʁa/ from *Rato*,
- /ʁɐ/ from *Terra*,
- /ʁə/ from *Respondia* and *Repente*,
- /ʁo/ from *Terror*, and
- /ʁu/ from *Ruínas*.

This data set is composed of 827 samples: 530 /ʁa/ samples, 57 /ʁɐ/ samples, 120 /ʁə/ samples, 63 /ʁo/ samples and 57 /ʁu/ samples. Since in spoken EP it is common to have reduction or deletion of unstressed vowels, some vowels present in the words above are not heard in some of the samples. This is the case for /ə/ in *respondia* and *repente*.

B. Scoring Phonemes

As discussed in section II, in order to score the produced phonemes, we used hierarchical and flat approaches with features extracted from LPC (with 22 and 24 poles) and MFCCs (with 9 to 23 cepstra) on NB, SVM and KDE classifiers. Here we compare the results of the different approaches. For the KDE classifier, we experimented with a $h = 0.2$ kernel bandwidth and we also tried the KDE with the Silverman's method for bandwidth selection. The leave-one-out cross-validation method was used to tune all the parameters.

The results obtained for the hierarchical approach are presented in tables I, II and III. The first column in the tables indicates the features used and the number of cepstra and poles for the MFCCs and LPC, respectively. The remaining columns show the accuracy results for each of the classifiers used. The best accuracies are indicated in bold. As illustrated in the tables, the MFCCs performed much better than LPC. All methods performed well with accuracies above 80.0% with the NB performing worse. The best was the KDE method with constant bandwidth. While hierarchy 1 shows the worst results, there seems to be no much difference between hierarchies 2 and 3.

The flat classification approach uses KDE for comparison with the hierarchical classifiers configuration (with both $h = 0.2$ bandwidth and Silverman's method). The results for all /ʁ/ and vowel phonemes are presented in table IV. In this approach the standard KDE performed as well as the hierarchical approach, but worse than the flat KDE with Silverman's bandwidth, that reached 95% overall accuracy making it the best method.

A visualization of the cross-phonemes' confusion is presented in figure 4. There was no detected confusion between phonemes of vowels and /ʁ/. The key message to extract from these confusion matrices is that the confusion across the different classes is very low, making these methods and speech exercises robust to be used in an interactive environment with children.

IV. Conclusions

Here we discussed robust phoneme recognition of EP vowel phonemes and the EP rhotic consonant sound for a speech

therapy environment. The environment includes motivational elements such as a progress bar. While we do not want to make the bar fully automatic (since SLTs should always have some control of the bar), robust phoneme recognition can give the SLTs valuable feedback, suggesting when to advance the progress bar.

We explored hierarchical and flat classifiers with NB, SVM and KDE that use LPC coefficients and MFCCs. The best results were obtained with the flat KDE with Silverman's bandwidth using MFCCs. A reproducible evaluation of the cross-phoneme confusion showed it to be robust enough to be used in interactive environments.

As future work we plan to explore the recognition of other more complex utterances and speech productions.

REFERENCES

[1] I. Guimarães, C. Birrento, C. Figueiredo, and C. Flores, "Teste de articulação verbal," Oficina Didáctica, Lisboa, Portugal, 2014.

[2] Frank R Adams, Hubert Crepy, David Jameson, and J Thatcher, "Ibm products for persons with disabilities," in *Global Telecommunications Conference, 1989, and Exhibition. Communications Technology for the 1990s and Beyond. GLOBECOM'89., IEEE*. IEEE, 1989, pp. 980–984.

[3] K Vicsi, P Roach, A Öster, Z Kacic, P Barczikay, A Tantos, F Csatári, Zs Bakcsi, and A Sfakianaki, "A multimedia, multilingual teaching and training system for children with speech disorders," *International Journal of speech technology*, vol. 3, no. 3-4, pp. 289–300, 2000.

[4] Olov Engwall, Olle Bälter, Anne-Marie Öster, and Hedvig Kjellström, "Designing the user interface of the computer-based speech training system artur based on early user tests," *Behaviour & Information Technology*, vol. 25, no. 4, pp. 353–365, 2006.

[5] Mirela Danubianu, Stefan-Gheorghe Pentiuc, Ovidiu Andrei Schipor, Marian Nestor, Ioan Ungureanu, and Doina Maria Schipor, "Terapers-intelligent solution for personalized therapy of speech disorders," *International Journal On Advances in Life Sciences*, vol. 1, no. 1, pp. 26–35, 2009.

[6] Protopapas A Öster AM, D House, and A Hatzis, "Presentation of a new eu project for speech therapy: Olp (ortho-logo-paedia)," .

[7] Oscar Saz, Shou-Chun Yin, Eduardo Lleida, Richard Rose, Carlos Vaquero, and William R Rodríguez, "Tools and technologies for computer-aided speech and language therapy," *Speech Communication*, vol. 51, no. 10, pp. 948–967, 2009.

[8] Ron Cole, Dominic W Massaro, Jacques de Villiers, Brian Rundle, Khaldoun Shobaki, Johan Wouters, Michael Cohen, Jonas Baskow, Patrick Stone, Pamela Connors, et al., "New tools for interactive speech and language training: using animated conversational agents in the classroom of profoundly deaf children," in *MATISSE-ESCA/SOCRATES Workshop on Method and Tool Innovations for Speech Science Education*, 1999.

[9] M. Carvalho, "Interactive game for the training of portuguese vowels," M.S. thesis, Faculdade de Engenharia da Universidade do Porto, 2008.

[10] Yves Rybarczyk, José Fonseca, and Ricardo Martins, "Lisling 3d: a serious game for the treatment of portuguese aphasic patients," in *Proc. 12th conference of the Association for the Advancement of Assistive Technology in Europe*, 2013.

[11] Alberto Abad, Anna Pompili, Angela Costa, Isabel Trancoso, José Fonseca, Gabriela Leal, Luisa Farrajota, and Isabel P Martins, "Automatic word naming recognition for an on-line aphasia treatment system," *Computer Speech & Language*, 2012.

[12] A. Grossinho, S. Cavaco, and J. Magalhaes, "An interactive toolset for speech therapy," in *Proceedings of Advances in Computer Entertainment Technology Conference (ACE)*, 2014.

[13] Madalena Cruz-Ferreira, "Portuguese (european)," in *Handbook of the International Phonetic Association, A guide to the use of the international phonetic alphabet*. Cambridge, University Press, 1999.

[14] I. Rennicke and P. Martins, "As realizações fonéticas de /R/ em português europeu: análise de um corpus dialetal e implicações no sistema fonológico," in *Encontro Nacional da Associação Portuguesa de Linguística*, 2013, pp. 509–523.

[15] Bernard W Silverman, "Using kernel density estimates to investigate multimodality," *Journal of the Royal Statistical Society. Series B (Methodological)*, pp. 97–99, 1981.

[16] Aníbal JS Ferreira, "Static features in real-time recognition of isolated vowels at high pitch," *The Journal of the Acoustical Society of America*, vol. 122, no. 4, pp. 2389–2404, 2007.

[17] I. Guimarães, "A ciência e a arte da voz humana," Escola Superior de Saúde do Alcoitão (ESSA), Alcoitão, Alcabideche, 2007.