How am I reading? Using a chatbot to evaluate prosodic cues in Italian L2

Valentina De Iacovo[1], *Marco Palena*[2], *Antonio Romano*[1]

[1]Università degli Studi di Torino
[2]Politecnico di Torino

valentina.deiacovo@unito.it, marco.palena@polito.it, antonio.romano@unito.it

ABSTRACT

In addition to pronunciation, various prosodic components such as intonation, duration, rhythm and pauses further refine speech. In the evaluation of a foreign language, it is often difficult to untangle these parts and the feedback given is rather generic but thanks to automatic segmentation systems and speech analysis tools it is now possible to compute meaningful parameters that characterise different prosodic components. In this paper we describe an experiment carried out starting from the use of a chatbot to evaluate some prosodic features in the reading of sentences by learners of Italian. After briefly describing the functioning of the chatbot, we focus on fluency and articulatory rate, and their relation to pauses, taking into account the perceptual evaluation performed both by human agents and a machine learning model.

Key words: prosodic evaluation, CALL, Italian L2, educational chatbot, human-computer interaction

1. Introduction

When we speak, we make use of different prosodic features which reveal important communicative information about the modality, the style, the attitude or the regional or social connotations (see, among others, [1], [2], [3]).
These acoustic nuances are essential because they convey multiple pieces of communicative information that, while they may go unnoticed by a native speaker, deserve more attention in a teaching context because they make the learner more aware in mastering the studied language. In fact, this approach reinforces in the learner certain speech recognition and reproduction mechanisms ([4], [5]) that can be trained through various digital tools that exploit user-machine interaction for the assessment of language skills. A tool which seems to represent a good compromise in the educational context is the chatbot for several reasons: it exploits a type of interaction that is already widely used by the public (for example Whatsapp or Telegram), it can be used in an asynchronous mode and is therefore free from the user and structured, thus minimising possible communicative ambiguities.
The so-called educational chatbots ([6], [7]) often assess grammatical or lexical skills ([8], [9]) but also pronunciation has been taken into account ([10]) and recent studies show a growing interest in the use of ASR systems ([11], [12]). Indeed, teaching (and assessing) components in the spoken medium has been a challenge for many years now ([13], [14]), among other things, also for the very description of the skills achieved (see [15], [16], [17]). In fact, in addition to pronunciation, oral skills concern precisely those suprasegmental aspects (such as rhythm, intonation, speech rate) that represent a challenge in evaluation, even more so if by an automatic system.
If we look at Italian, prosodic variation does indeed play an important role because, in addition to the aspects mentioned, it reflects a regional richness ([18], [19]), offering an abundant range of prosodic patterns in oral learning ([20], [21]). It is therefore legitimate to ask whether there is a single prosodic model to be adopted in the teaching of Italian L2 (and which one, if any) or whether it is better to ensure that the student is exposed to several regional varieties and that these represent a starting point for prosodic reflection with respect to spoken Italian.

Starting from these considerations, the aim of this study is to provide a first assessment of the oral production of FL Italian learners based on some acoustic cues. In particular, we analysed the correlation between the acoustic values extracted from the collected data (number of syllables, fluency rate, articulatory rate and number of pauses produced) and the perceptual evaluations (on intonation and fluency rate) to see which of the values have the greatest impact on the perception of a more or less spontaneous speech. After compiling a corpus of sentences read by Italian speakers, we employed them as a basis of comparison to evaluate the same sentences read by students of Italian. We then scored the intonation of both groups on a perceptual basis. Subsequently, a perceptual score was assigned on the basis of speech rate to half of the group of learners in order to train a predictive model capable of predicting the speech rate of the remaining half of the group ([22]). Having set out the structuring of the chatbot and detailed the acoustic parameters used in the

training of the predictive model, we then comment on the results achieved and our observations.

2. Data and Methods

In this section, we describe the source corpus (sentences read by Italian speakers) and how the chatbot used for data collection (sentences read by Italian learners) is structured.

2.1. Corpus of Italian speakers

In order to be able to assess reading, we needed a reference model that allows an intonational comparison between the FL Italian learner and the native speaker. A corpus of 10 small sentences containing dates, numbers and various intonation structures (assertive, interrogative, continuative) was therefore compiled. A total of 400 people were involved in reading them in order to have a consistent intonational variation for each sentence. Recording people from different parts of Italy is crucial to ensure the diatopic variation that, as mentioned earlier, affects speech production in prosodic terms. Although diatopic variation was not explicitly observed in this study, participants came from different parts of Italy and included 280 females and 120 males aged between 15 and 70 years for a total of 4000 utterances recorded. Each audio was then resampled to 16kHz and manually labelled. The intonational fluency, i.e. the ability to read the sentence more or less spontaneously in terms of intonation, was evaluated perceptually by two expert phoneticians with a score of 1 (not spontaneous intonation), 2 (acceptable intonation) or 3 (spontaneous intonation). It is important to emphasise that the sentences have been read out, so the highest level of perceptual judgement (3) corresponds to a type of intonation attributable to spontaneous speech. However, it is also useful to remember that in the perceptual assessment of intonation, other factors may influence the judgement: think of pauses or speech rate. In the evaluation made by the two experts, therefore, an attempt was made to set these parameters aside by assessing intonation performance alone.

2.2. Chatbot structure and functioning

The chatbot has been implemented within the instant messaging application Telegram (further information is available in [23]) and involves interaction with the user through questions and answers based on the assessment of technical knowledge (literature, history, mathematics). The 10 sentences mentioned represent the answers: through a series of closedended questions (quizzes), the student must identify the correct answer and send a statement of it by voice message. Each question is presented in written and oral form while he can choose the answer among four options (if the answer is wrong, they can try again). Once the correct answer is chosen, the chatbot suggests to record the answer through a voice message. The elicited utterance is then automatically processed by the bot in order to obtain an evaluation of the intonation level of the user. The audio is first converted to single channel wav format, resampled (if necessary) to 48 kHz and cleaned of background noise. The average amplitude value and signal-to-noise ratio of the resulting audio are then estimated.

If the estimated values are below certain predefined thresholds, the chatbot prompts the learner to record a new utterance in a less noisy environment and/or by speaking in a higher tone of voice. The speech signal is then subjected to segmentation using the WebMAUS Basic web service ([24]) which takes as input the speech signal and the orthographic transcription of the utterance and returns a segmentation into words and phonemes ([25]). The phonetic segmentation of the utterance, provided in TextGrid format to facilitate subsequent processing using Praat software ([26]), is then processed by labelling individual phonemes as vowels or consonants. A Praat script is then invoked to extract the f_0 values of the previously identified vowel phonemes, thus obtaining the intonation curve of the utterance ([27]). The analysis of intonation is carried out by comparing the intonational curve of the user's utterance with the f_0 traces of the corresponding utterances of native speakers, previously collected and evaluated using the same automatic procedure. The comparison is made by calculating a correlation measure ([28]) that compares for each sentence three points (initial, central, final) of f_0 of each vowel segment identified by Maus. The result is a percentage value expressing the intonational proximity to the closest Italian native speaker. In order to provide a feedback that is easily understandable by the user, the resulting percentage is converted into one of three intonational classes, based on predetermined thresholds: close to a native speaker (green), moderately proximate to a native speaker (yellow) and distant from a native speaker (red). At the end of each given answer, the computed intonational class is returned to the user. Once the task is completed, a summary score obtained from the average of the percentage values for each answer is calculated. As the vocalic segments detected for users' utterances and those detected for the corresponding utterances of native speakers might differ, before calculating the correlation, the segments of the two speakers are aligned on the basis of both the phonetic information they contain and their temporal position. The correlation measure is then restricted to all and only those vowel segments to ensure phonetic-segmental homogeneity between the two speakers. Of course, any kind of phenomenon related to the good scansion of the phrase, the completeness of the recording or the lowering of the voice influences the final evaluation.

2.3. Additional assessed features

Currently, the chatbot returns a value solely on the basis of the intonation correlation. The evaluation of reading should consider other acoustic correlates though (such as duration, rhythm, fluency rate, pauses). As a result, for this study, we focus on fluency rate (the number of syllables uttered by the speaker out of the total duration of the sentence) and articulatory rate (the number of syllables uttered by the speaker out of the duration of each speech-chain in the sentence) in relation to the number of pauses made.

2.4. Training the model

In order to provide the users with a meaningful feedback about their fluency rate, a preliminary experimental evaluation has been conducted to assess the feasibility of training a machine learning

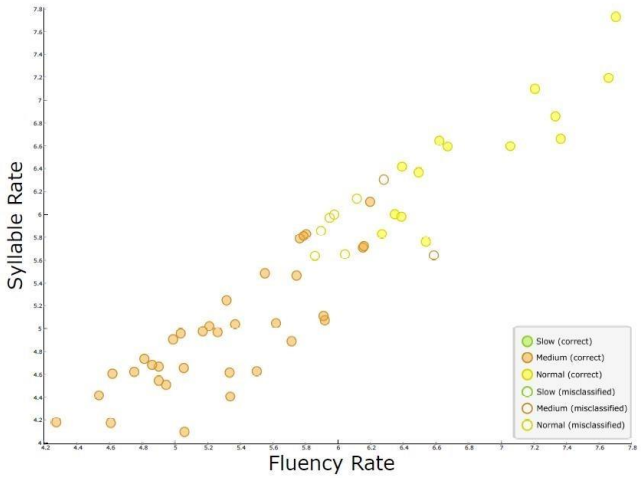
model based on the parameters mentioned above extracted from the speech signals.

First, syllabification of the speech signal is performed using the automatic procedure provided by the WebMAUS Basic web service. Training data has been collected from 63 users whose fluency rate was perceptively assessed according to three levels: slow/not natural (1), medium/read (2), fast/spontaneous (3). Using these perceptive evaluations as targets and the extracted prosodic parameters as features, three classifiers for each sentence were trained to compare the performance of three state-of-the-art classification algorithms: k-Nearest Neighbour ([29]), Logistic Regression ([30]) and Random Forests ([31]). All models were trained using the Orange data mining toolkit ([32]). The models were validated using a 3-fold cross-validation approach and compared in terms of classification accuracy. Results are reported in Table 1. As shown, the model based on logistic regression emerges as a clear winner (in bold), being able to reach an accuracy of more than 80% on most sentences, even considering the relatively small size of the dataset used for training.

Table 1: *Comparison of the classification accuracy of the models learned for each sentence. The most accurate classifier for each sentence reported in bold.*

Algorithm	R01	R02	R03	R04	R05	R06	R07	R08	R09	R10
kNN	79.7%	80.3%	80.4%	82.3%	56.9%	**78.6%**	64.5%	79.0%	79.0%	52.2%
Random Forests	64.4%	82.0%	83.9%	80.6%	48.3%	67.9%	**72.6%**	77.4%	74.2%	39.1%
Logistic Regression	74.6%	**83.6%**	**85.7%**	**82.3%**	**62.1%**	**78.6%**	64.5%	**80.6%**	**80.6%**	**56.7%**

Figure 1: *Comparison of the classification accuracy of the models learned for each sentence. The most accurate classifier for each sentence reported in bold.*



3. Results

The task has been submitted to 63 students of Italian. Based on the sociolinguistic information given at the beginning of the registration, the languages spoken by the users are quite heterogeneous (mostly French, Spanish and English though), half of them have lived in Italy and their Italian level is B2-C1 according to the CEFR level. From the total number of sentences (630), 120 were excluded because the user did not read them or because he/she read them incorrectly.

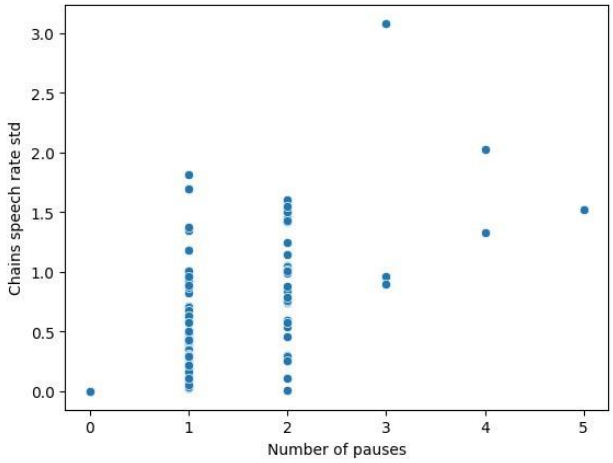
3.1. The intonational score

Regarding the perceptive evaluation, out of the 4000 sentences read by Italian speakers, 91% received a high score (3), 8% medium (2) and 1% low (1). The sentences read by students received similar percentages: 70% had a max score, 26% medium and 4% low. We can therefore say that in both groups intonation is perceptually associated with a quite spontaneous way of speaking. If we observe the relation of the perceptual score between the native speakers and the students, the association is constant (the student receives the same score as the native speaker) in 69.3% of the cases, improving in 26.3% (the student receives a lower score than the native speaker) or worsening in 4.4% (the student receives a higher score than the native speaker). Where the association is constant, the majority of answers given (66%) received a maximum score (3) with an average intonational proximity of 73%. The second highest percentage (23%) concerns the association between students with a score of 2 and native speakers with a score of 3: in this case the average score is 70% which means that, although students were assessed perceptually with a medium score (2), the automatic assessment still associated them with a native speaker with a higher score. This is encouraging on the one hand because it seems to bring out an autonomous evaluation of intonation with respect to other potential disfluencies in speech (hesitations, repetitions) but leads to a widening of the range of parameters to be evaluated to enrich the final feedback to give to the student.

3.2. The trained model

Figure 1 provides an example of how the 63 instances of the dataset are classified in terms of fluency rate for sentence 3 ("*Ciao Salvatore, allora ci andiamo a fare una partita a calcio uno di questi giorni?*") according to the best performing classifier learned (logistic regression). Students' mean fluency rate for this utterance is 5.4 while articulatory rate is 5.8, compared to the Italian corpus where fluency rate is 6.1 and articulatory rate is 6.3 (see Table 2 afterwards). Utterances are correctly classified with a score of 2 (medium/read) when their fluency rate is less than 5.8 and 3 (fast/spontaneous) when is more than 6.2. Between these two values (5.8 and 6.2) utterances are misclassified.

Figure 2: *Relation between fluency rate and number of pauses for sentence 3. The blue line represents the FR mean for Italian speakers.*



3.3. Fluency rate vs pauses

Fluency rate is generally lower for students who also tend to take more pauses within the sentence. Figure 2 shows the relation between fluency rate and number of pauses for sentence 3 produced by the students: most of the sentences contain a pause and were rated perceptually as fast (blue dots) if close to or above the fluency rate threshold of 6.0 (the blue line indicates the average fluency rate for Italian speakers); it can also be observed that those with a lower fluency rate also tend to produce more pauses.

3.4. Chains fluency rate variance vs pauses

Table 2 shows the comparison between Italian speakers and students based on the total duration of sentence 3, the number of syllables, fluency and articulatory rate and the duration of pauses.

Table 2: *Comparison between Italian speakers (ItS) and students (St) for sentence 3.*

	Tot dur. (s)	N. of syllables	Fluency rate	Artic. rate	Dur. of pauses (s)
ItS	4.9	26.3	6.1	6.3	0.2
St	6.2	27.7	5.4	5.8	0.3

Since fluency rate and articulation rate give an overall reference of the utterance, we wanted to investigate if there is a link between the varying rate of each speech chain (variance) and the number of pauses. Indeed, comparing the results obtained for sentence 3 by the Italians speakers (Figure 3) and the students (Figure 4), we notice that most Italians take 1 or 2 pauses and have a fluency rate that varies between 0 and about 2 syllables/s between each chain; on the contrary, the distribution of pauses is more heterogeneous in for students who present a compact variance mainly between 0 and 1 instead.

Figure 3: *Relation between chains fluency rate variance and number of pauses for sentence 3 by Italian speakers.*

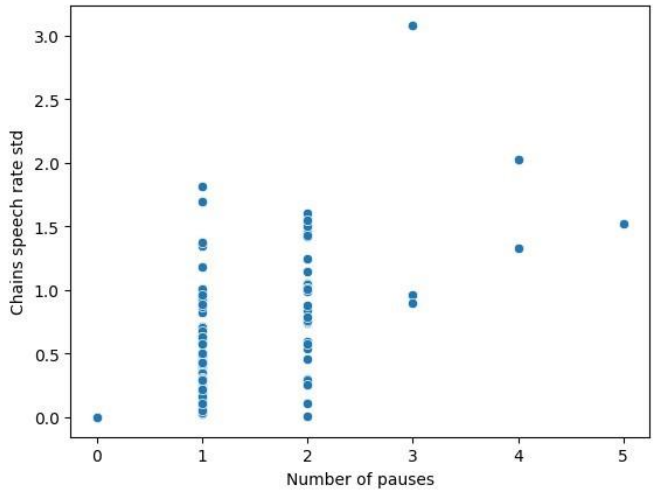
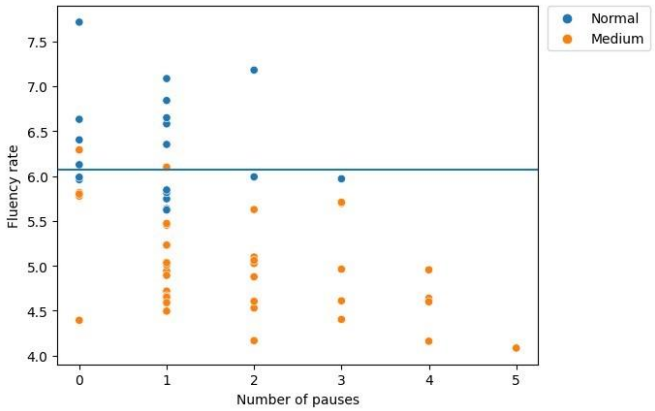


Figure 4: *Relation between chains fluency rate variance and number of pauses for sentence 3 by students.*



4. Discussion

We discuss here the results starting from utterance 3 which obtained the best accuracy level from logistic regression. Although the corpus of Italian speakers and students presents limitations due to the homogeneity of the data (average spontaneous reading, fairly young speakers), the approach using the machine learning model seems promising in terms of using the chatbot to assess the oral competence of a learner of Italian. The results show that a sentence is considered perceptually spontaneous if its fluency rate is above 6 syl/s. Below this threshold, it is considered less spontaneous instead. At the same time, the trained model has difficulty classifying the same sentence if it has a rate between 5.8 and 6.2 syl/s. If we relate the fluency rate to the number of pauses (Figure 2), we note that the utterances produced by students that are considered spontaneous perceptually almost always have a fluency rate equal to or greater than 6 syl/s. These also have a number of pauses equal to 0 or 1, while those with pauses greater than 1 also have a lower fluency rate. Finally, we found it useful to analyse the relationship between the articulatory rate in individual speech chains related to the number of pauses produced. In fact, comparing tables 3 and 4, we deduce that Italian speakers tend to make one or two pauses while students tend to make more pauses within the same utterance.

5. Conclusion

Assessing the acoustic correlates of prosody can be very useful for an FL learner because it allows him/her to notice nuances specific to the language being studied. In this paper, we have presented a part of a project we are currently conducting which studies how the use of a chatbot can represent a didactic tool in the automatic evaluation of read speech by language learners. The results presented seem encouraging but point to a complexity of prosodic features which combine perfectly in speech but which must be considered individually in order to be evaluated. In speech, in fact, there are several components related to the number of syllables elicited, the rhythm and speed of speech used, the phonetic, semantic and syntactic complexity of the sentence, as well as the level reached by the Italian speaker as L2. The comparison with the perceptual scores shows a certain correspondence between the perception of fluency rate

and the acoustic values extracted. There is certainly a need to increase the number of users, extending it especially to the most elementary levels in order to have a better representativeness ([33]). Finally, the biggest challenge is to ensure that a fully automatic system is able to return an evaluation as accurate as possible: for this reason, cases of disfluencies, false starts, hesitations, pauses, etc. must be identified. We therefore want to complete the evaluation of other acoustic parameters and ensure a more diatopically balanced reference corpus. In fact, we started with a corpus of read speech because it is easier to monitor compared to spontaneous speech ([34]) but we still need to verify other aspects related to ASR: for example, the reliability of the segmentation done by WebMaus, which is also related to the potential false starts or hesitations made by the speaker. The next step is to collect additional data to train more accurate logistic regression models that will then be integrated in the bot to provide the users with a more comprehensive evaluation of their prosodic production.

6. Acknowledgments

This study is part of the project CALL-UniTO funded by the Fondazione CRT - Bando Erogazioni Ordinarie 2020. We would like to thank Donatella Bisconti, Francesca Nicora and Liliana Vocale for the coordination and all the participants who took part in the recording and assignment.

7. References

[1] Delattre, P. (1966). Les dix intonations de base du français. *French review*, 1-14.

[2] Hirst, D. (1983). Structures and categories in prosodic representations. In Cutler A. & Ladd R. (eds.) Prosody: Models & Measurement (pp. 93-109). Berlin: Springer.

[3] Cruttenden, A. (1997). *Intonation*. Cambridge University Press.

[4] Cresti, E. (1999). Force illocutoire, articulation topic/comment et contour prosodique en italien parlé. *Faits de langues*, 7(13), 168-181.

[5] Chun, D. M. (2002). *Discourse intonation in L2: From theory and research to practice*. John Benjamins Publishing.

[6] Pereira, J. (2016). Leveraging chatbots to improve self-guided learning through conversational quizzes. *Proc of the 4th International Conference on Technological Ecosystems for Enhancing Multiculturality, Salamanca, Spain,* 911-918.

[7] Fernoagă, V., Stelea, G.-A., Gavrilă, C., & Sandu, F. (2018). Intelligent education assistant powered by chatbots. *eLearning & Software for Education*, 376-383.

[8] Nghi, T. T., Phuc, T. H., & Thang, N. T. (2019). Applying AI chatbot for teaching a foreign language: An empirical research. *International Journal of Scientific and Technology Research*, 8(12), 897-902.

[9] Kim, N.-Y. (2019). A study on the use of artificial intelligence chatbots for improving English grammar skills. *Journal of Digital Convergence*, 17(8), 37-46.

[10] Franco, H. Bratt, H. Rossier, R. Rao Gadde, V. Shriberg, E. Abrash, V & Precoda, K. (2010). Eduspeak®: A speech recognition and pronunciation scoring toolkit for computer aided language learning applications. *Language Testing*, 27(3), 401-418.

[11] Colace, F., De Santo, M., Lombardi, M., Pascale, F., Pietrosanto, A., & Lemma, S. (2018). Chatbot for e-learning: A case of study. *International Journal of Mechanical Engineering and Robotics Research*, 7(5), 528-533.

[12] Cheng, V. C.-W., Lau, V. K.-T., Lam, R.W.-K., Zhan, T.-J. & Chan, P.K. (2020). Improving English phoneme pronunciation with automatic speech recognition using voice chatbot. *Proc of the 5th International Conference on Technology in Education*, Macau, China, 88-99.

[13] Chun, C. (1998). Signal analysis software for teaching discourse intonation. *Language Learning & Technology*, 2(1), 74-93.

[14] Cazade, A. (1999). De l'usage des courbes sonores et autres supports graphiques pour aider l'apprenant en langues. *Apprentissage des Langues et Systèmes d'Information et de Communication*, 2(2), 3-32.

[15] James, E. (1976). The acquisition of prosodic features of speech using a speech visualizer. *International Review of Applied Linguistics*, 143, 227-243.

[16] De Bot, K. (1983). Visual feedback of intonation: Effectiveness and induced practice behavior. *Language and speech*, 26(4), 331-350.

[17] Martin, P. (2010). Learning the prosodic structure of a foreign language with a pitch visualizer. *Proceedings of the 5th Speech Prosody, Chicago, USA*, 1-4.

[18] Canepari, L. (1983). *Italiano standard e pronuncia regionale*. CLEUP.

[19] Sorianello, P. (2006). *Prosodia: Modelli e ricerca empirica*. Carocci.

[20] De Meo, A. & Pettorino, M. (2013). *Prosodic and rhythmic aspects of L2 acquisition: the case of Italian*. Cambridge Scholars Publishing.

[21] De Marco, A., Sorianello, P. & Mascherpa, E. (2014). L'acquisizione dei profili intonativi in apprendenti di italiano L2 attraverso un'unità di apprendimento in modalità blended learning. In De Meo, A. D'Agostino, M., Iannaccaro G. & Spreafico, L. (eds.), *Varietà dei contesti di apprendimento linguistico*, (pp. 189-211). Collana Studi AItLA.

[22] Papi, S. Trentin, E. Gretter, R. Matassoni, M., & Falavigna, D. (2021). Mixtures of deep neural experts for automated speech scoring. arXiv preprint arXiv:2106.12475.

[23] De Iacovo, V., Palena, M., & Romano, A. (2022). Evaluating prosodic cues in Italian: the use of a Telegram chatbot as a CALL tool for Italian L2 learners. In Bernardasci, C., Dipino, D., Garassino, D., Negrinelli, S., Pellegrino, E. & Schmid, S. (eds.), *Speaker individuality in phonetics and speech sciences: speech technology and forensic applications* (pp. 283-298). OfficineVentuno.

[24] Kisler, T. Reichel, U. & Schiel, F. (2017). Multilingual processing of speech via web services. *Computer Speech & Language*, 45, 326-347.

[25] Schiel, F. (1999). Automatic phonetic transcription of non-prompted speech. *Proc of the 14th International Conference of Phonetics Sciences, San Francisco, USA*, 607-610.

[26] Boersma, P. & Weenink, D. (2022). Praat: doing phonetics by computer [Computer program]. Version 6.3.03, retrieved 2 May 2023 from http://www.praat.org/

[27] Romano, A., Contini, M., & Lai, J. P. (2014). L'Atlas Multimédia Prosodique de l'Espace Roman: uno strumento

per lo studio della variazione geoprosodica. *Proc 20th Jahre digitale Sprachgeographie, Berlin, Germany*, 27-51.
[28] de Castro Moutinho, L., Coimbra, R. L., Rilliard, A., & Romano, A. (2011). Mesure de la variation prosodique diatopique en portugais européen. *Estudios de fonética experimental*, 20, 33-55.
[29] Dasarathy, B. V. (1991). Nearest Neighbor (NN) Norms: NN Pattern Classification Techniques. IEEE Computer Society Press tutorial. IEEE Computer Society Press.
[30] McCullagh P. and Nelder, J. A. (1989). *Generalized Linear Models*. Chapman & Hall / CRC.
[31] Breiman, L. (2001). Random forests. *Machine Learning*, 45(1), 5-32.
[32] Demšar, J., Curk, T., Erjavec, A., Gorup, Č., Hočevar, T., Milutinovič, M., & Zupan, B. (2013). Orange: data mining toolbox in Python. *The Journal of machine Learning research*, 14(1), 2349-2353.
[33] Lacheret-Dujour, A. (2001). Modéliser l'intonation d'une langue. Où commence et où s'arrête l'autonomie du modèle? l'exemple du français parlé. *Colloque international: Journées Prosodie, Grenoble, France.* 57-60.
[34] Nakamura, M., Iwano, K., & Furui, S. (2008). Differences between acoustic characteristics of spontaneous and read speech and their effects on speech recognition performance. *Computer Speech & Language*, 22(2), 171-184.