Comparison of Heterogeneous Feature Sets for Intonation Verification

Mariana Julião[1,2(✉)], Alberto Abad[1,2], and Helena Moniz[1,3,4]

[1] INESC-ID, Lisbon, Portugal
{mariana.juliao,alberto.abad}@l2f.inesc-id.pt, helena.moniz@inesc-id.pt
[2] Instituto Superior Técnico, Universidade de Lisboa, Lisbon, Portugal
[3] FLUL - Faculdade de Letras da Universidade de Lisboa, Lisbon, Portugal
[4] CLUL - Centro de Linguística da Universidade de Lisboa, Lisbon, Portugal

Abstract. The assessment of intonation, to which intonation verification belongs, has many applications, such as health-impaired people training – from individuals with Parkinson's disease to children with Autism Spectrum Disorders – and second language learning. Most of the approaches that are found in the literature are based on intensive preprocessing of the audio signal and hand-crafted feature extraction methods, and most of those works do not tackle the particularities of the Portuguese language. In this paper, we present our work on intonation assessment, developed from a database of binarily-labelled Portuguese intonation imitation. This has been done using the set of Low Level Descriptors (LLDs) and eGeMAPS, both extracted with the openSMILE toolkit, and Problem-Agnostic Speech Encoder (PASE) features. We have taken the most informative feature subsets for prosody out of these. Distances between stimulus and imitation – the so-called similarity intonation scores – have been computed applying Dynamic Time Warping (DTW) for different feature subsets, and have afterwards been used as input features of a binary classifier. Performance achieves up to 66.9% of accuracy in the test data set when considering only one feature set, and it increases up to 77.5% for a set of seven features.

Keywords: Intonation assessment · Prosody imitation verification · Multiple feature fusion

1 Introduction

Although often disregarded among other aspects of language, prosody plays a crucial role in human communication, as it provides a wealth of relevant information. It encompasses pitch, rhythm, energy, and voice quality. Intonation corresponds to the variations of pitch, i.e., the melodic contours. Whereas rhythm helps segment the speech stream into smaller units as words and phrases, intonation provides information on the sentence type. For instance, in languages as Portuguese, where yes-no questions are not lexically distinguishable from affirmations, intonation is what allows to decide between these two, or even between

P. Quaresma et al. (Eds.): PROPOR 2020, LNAI 12037, pp. 1–10, 2020.
https://doi.org/10.1007/978-3-030-41505-1_2

an affirmation and an exclamation. It carries information on intent, focus, attitudes and emotions, not to mention all the information it conveys about the speaker.

Prosody is fluidly acquired in typically developing children. Infants have the ability to mimic the melodic aspects of speech long before they are able to articulate words [5,10]. On the other hand, it reveals itself quite hard to master by second language learners [6], as well as children with Autism Spectrum Disorders (ASD). The prosody produced by the latter, according to [15], is reported to include monotonic intonational patterns, misplaced stress patterns, deficits in pitch and intensity control, and differences in voice quality. Furthermore, according to the same authors, observations suggest that these aspects "tend to persist over time, even when other aspects of language improve".

Intonation assessment has been addressed in different ways. These can broadly be classified as assessment (how good the intonation of a segment is), as in [1,4,12,20], and classification of segments (according to their labels, as H or L, for instance) as in [7,11].

This work presents the preliminary experiments in intonation verification in a data set. Firstly recorded as an imitation task, the augmentation of the data set has made it more similar to a verification task, for considering unrelated samples as bad imitations. This is explained in detail in Subsect. 4.1. Unlike what is commonly seen in the literature, instead of using hand-crafted features alongside preprocessing, we have resorted to simple classification algorithms, and to the available general toolkits for feature extraction, with minimal preprocessing.

The relevance of our work is threefold. First, the task of intonation imitation/verification is a fundamental task in prosody assessment, which has a plethora of applications. Then, this problem has not been solved yet, as there is still no reliable way to compare intonation. Finally, to our best knowledge, no similar work has been done for Portuguese. As the number of L2 speakers of Portuguese keeps increasing, a growth in the demand for learning applications is to be expected, and these should not leave prosody assessment aside.

2 Related Work

The seminal work for goodness of imitation by Arias, Yoma, and Vivanco [1] is still a reference these days. The work considers data from second language learners (L2) of English with Spanish as their native language (L1). MFCCs were aligned using a DTW algorithm, which was afterwards used to compare the F0 curves on a frame-by-frame basis. Their data set consisted short sentences, uttered with different intonation patterns by 16 speakers. It achieved an averaged subjective-objective score correlation of 0.88.

Cheng considered recordings of Pearson Test Academic English, rated by at least two human experts [4]. Their main contribution is the use of k-means clustering method to build canonical contour models at the word level for F0 and energy, which provided strong predictors of prosody ratings. For the comparison of sequences, they had extracted word boundary information, and then scaled

every word interval to a standard length, after which they applied Euclidean distance, correlation coefficient, and DTW. The authors concluded that, being F0 and energy contours strong predictors of prosody ratings, duration information was the best predictive feature. The linear regression model scores highly correlated with human ratings, $r = 0.80$.

Ma et al. [12] used 48 features based on the normalised F0 measurements, to assess nativeness in a sentence repeat task. Their data set comprised 9k responses from non-native speakers of English, from China and India, and 4k responses from native speakers of US English. The evaluation of the results was done using a weighted F1 score, achieving 0.783 in a text-independent context and 0.741 in a text-dependent one. According to their results, no substantial drop in nativeness detection has been noticed between models trained on polynomial functions on first seven degrees of polynomials and the models trained only on the first three. This means that the "basic descriptive statistics and lower-degree polynomial parameters can capture the primary characteristics".

Also for nativeness assessment, this time focusing on the classification of prosodic contours, the study of Escudero-Mancebo [7] presents a system that computes distances between sequences of prosodic labels. It considers prosodic contours by using a set of metrics based on joint entropy. The data set consisted of fifteen read sentences from the news paragraphs of a bigger corpus, read by a group of Japanese speakers of Spanish. This system performed a pairwise classification, which combined evidence for three complementary types of classifiers (artificial neural networks, decision trees, and support vector machines), which have been combined using a comprehensive fuzzy technique. It allowed for the analysis of the most frequent potential misuses of the Sp_ToBI tones as a cue for possible mistakes that appear in the prosodic productions of non-native speakers.

Automatic intonation classification was done applying multi-distribution DNNs to data from Chinese L2 learners of English [11]. The authors labelled intonation as rising, upper, lower or falling, which are then shrinked into two categories, rising and falling. Ultimately, by considering the last 80 ms of an intonational phrase, they determine the intonation of L2 English speech utterances as either rising or falling, with an accuracy of 93.0%.

3 Imitation Classification Approach

Figure 1 shows the pipeline of the approach followed in this work. First, for both stimulus and imitation speech signal, word segmentation is applied to find start and end of word boundaries (Subsect. 3.1), followed by feature extraction (Subsect. 3.2). Next, the distance between stimulus and imitation is computed using DTW, for each specific feature set (Subsect. 3.3). We refer to the distances obtained with the DTW as *similarity intonation scores*. Then, these intonation scores are fed to a classifier, which classifies them as corresponding to a *good* imitation – stimulus and imitation match – or a *bad* imitation – stimulus and imitation differ (Subsect. 3.4). We try both feeding the classifier one type of intonation score at a time – corresponding to one feature or feature set alone – and

several types of intonation scores (working as different features for the classifier), obtained from different feature sets. We have made two sets of experiences: in train set and in test set. All the experiments in the train set are done in cross-validation with leave one speaker out, and will, henceforth, be referred to as cross-validation.

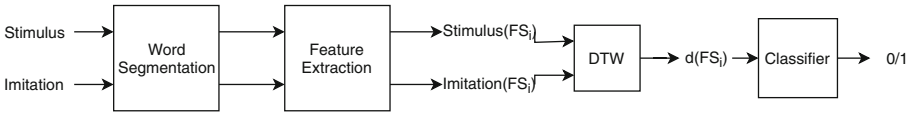

Fig. 1. Intonation comparison approach.

3.1 Word Segmentation

In Portuguese prosodic words, stresses may fall in the last three syllables of a word, and this happens mostly in the antepenultimate syllable (in our data set, all words correspond to this case). As Portuguese has a very strong vowel reduction, prosodic information is by no means evenly spread throughout the word. On the contrary, most of the information comes in the stressed syllable and in what comes afterwards. For this reason, we thought of attending specifically to the words from the beginning of the stressed syllabled until the end. However, as 3 out of 5 words in our set have the stressed syllable as its first one, this would not lead to a significant difference. Therefore, we chose to consider the whole words.

To achieve a proper segmentation of the words, we ran forced alignment with Kaldi [16] trained with BD-Publico corpus [13]. This provided us the phone boundaries, with which we have segmented the segments of the recordings to consider (words). BD-Publico has been constructed to allow for a broad speech recognition system of continuous speech. Participants from a school of engineering have been recorded reading texts extracted from Público newspaper.

3.2 Feature Sets

In this work, rather than using task-oriented features, we exploit both the large feature sets obtained with the openSMILE toolkit [8] (LLDs and eGeMAPS) and the new bottleneck features from PASE [14]. From the openSMILE based features, we chose the ones which could potentially provide the best prosodic information: those related to pitch, energy, duration, and voice quality. For all cases, we normalised each word feature-wise with zero mean and unit variance before the DTW stage.

LLDs. The set of 130 low-level descriptors corresponds to the ComPaRE 2016 [18] data set. Features were extracted for each utterance with the openSMILE toolkit [8], for windows of 0.060 s with a step of 0.010 s. From this data set, we considered separately F0, MFCCs, and energy.

eGeMAPs. The Geneva Minimalistic Acoustic Parameter Set [9] is a set of functionals computed on top of low-level descriptors, designed specifically to provide a common ground for research in various areas of automatic voice analysis, namely paralinguistics. It comprises both the minimalistic set of 62 features (GeMAPS) and the extended set of 88 features (eGeMAPS), the former appending spectral and frequency-related parameters to the minimalist set.

As these are utterance-level features, and as we need to preserve the sequential aspect of the utterances, for the sake of comparison, we have extracted them for sliding windows of 0.200 s and steps of 0.010 s. From the eGeMAPS set, we considered separately the F0, MFCCs, loudness (to which we call "energy", hitherto), and slope.

Bottleneck Features. Recently, a Problem Agnostic Speech Encoder, PASE, has been presented [14]. It can be used to pre-train a network for speech classification tasks or simply as a speech feature extractor. This is a fully-convolutional speech encoder, followed by seven multilayer perceptron workers, which cooperate to solve different self-supervised tasks. After the convolutional blocks, an additional layer projects 512 features to embeddings of dimension 100. It emulates an overlapping sliding window using a set of convolutions, such that the input signal is decimated by a factor of 160. For sampling rates of 16 kHz, as in our case, this is equivalent to a 10 ms stride, which makes these features as usable in our task as the aforementioned ones.

3.3 Dynamic Time Warping

Dynamic Time Warping (DTW) is an algorithm for comparing time sequences [2], allowing for the comparison of time sequences of different length. It starts by building a distance matrix D, $n \times m$, where n and m are the lengths of the sequences to be compared. Each entry $D_{i,j}$, is the distance between entries i and j of the first and the second sequence, respectively, $D_{i,j} = d(x_i, y_j)$, and d is an adequate distance function.

We consider Euclidean and Cosine distances, defined as $|\boldsymbol{x}^2 - \boldsymbol{y}^2|$, and $1 - \frac{\boldsymbol{x} \cdot \boldsymbol{y}}{\|\boldsymbol{x}\| \|\boldsymbol{y}\|}$, respectively. After computing the distance matrix, the algorithm computes a warping path w, such that $d_{DTW}(x, y) = min \sum_{i=1}^{k} D(w_i)$.

The length normalised distance $d_{DTW}(x, y)$ of the two sequences along the warping path corresponds to the similarity intonation scores. In our problem, the sequences correspond to the whole segmented words. Features from the sequences are normalised to zero mean and unit variance before the distances are computed. Intonation scores are also normalised to zero mean and unit variance before classification.

3.4 Classification

For final imitation verification, we used Support Vector Machines (SVM) binary classifiers trained on the similarity intonation scores provided by the DTW stage.

Whereas any other type of classifier might have been used, SVMs were selected for their particular adequacy for tasks with a small training data size, such in our case. Among other machine learning models, SVMs stand out by being powerful discriminators that are able to perform well in a wide variety of tasks, including those where data is scarce, like in our task.

4 Experimental Setup

4.1 Data Set

Our intonation data set has 20 original stimuli recorded by a native female speaker of Standard European Portuguese and reproductions of it by 17 different speakers: 7 female, 10 male, all of them native speakers of Standard European Portuguese and with no known health impairments. The 20 original stimuli correspond to the possible combinations of five words: *banana*, *bolo*, *gelado*, *leite*, and *ovo*; and four intonations: *affirmation*, *question*, *pleasure*, and *displeasure*. Each utterance corresponds, then, to one word uttered with one particular intonation. This data set collection was initially part of a work for prosodic exercises for children with Autism Spectrum Disorders [19].

One non-expert annotator labelled the utterances binarily as *good* or *bad* imitations. After discarding non-suitable recordings, we were left with 335 imitations. The data set was split in train and test data subsets: 13 speakers for train (5 female, 8 male) and 4 speakers for test (2 female, 2 male). As expected from native speakers of the same variant of a language, most of the imitations were labelled as *good*. To increase the number of *bad* imitations in our data set, we have augmented it by adding pairs of stimulus and imitation considering orthogonality between some types of intonation: *Affirmation* orthogonal to *Question*, and *Pleasure* orthogonal to *Displeasure*. In that sense, a good imitation of an affirmation cannot be, at the same time, a good imitation of a question. Therefore, good imitations have been considered bad imitations of their counterpart. With this procedure, the final data sets consist of 494 imitations in the train set (of which 248 are *good* imitations), and 142 imitations in the test set (of which 72 are *good*).

4.2 Cross-Validation Experiments

Discrimination of Similarity Intonation Scores. In this section, we investigate the importance of the similarity intonation scores obtained with each individual feature set, as well as the proper distance metric for each one of them in the DTW stage. We considered the most informative feature sets for our task, according to literature: F0, MFCC, energy, slope – from LLD and eGeMAPS. We have trained SVMs with the intonation scores from each of these feature sets separately (cf. Fig. 1), for both euclidean and cosine distance in the DTW. The results in Table 1 are the average of the accuracies obtained in the SVM with leave-one-speaker-out cross-validation, and the corresponding standard deviations. We have also considered the PASE-BNF features. For each feature set, we

highlight the best result between euclidean and cosine. Clearly, the euclidean distance provides better results than the cosine distance for all feature sets except EG_F0. These best results are the ones to be considered hereafter.

Table 1. Cross-validation accuracy (%) and standard deviation for each feature subset.

	LLD			eGeMAPS				PASE-BNF
	F0	MFCC	Energy	F0	MFCC	Energy	Slope	All
euc	**75.5** ± 4.7	**63.0** ± 6.5	**64.7** ± 10.2	67.2 ± 5.0	**61.8** ± 4.4	**64.1** ± 5.4	**69.8** ± 6.0	**59.9** ± 6.8
cos	68.0 ± 7.2	62.6 ± 8.1	62.6 ± 8.1	**70.8** ± 9.9	58.8 ± 5.8	61.4 ± 8.0	64.4 ± 4.3	56.6 ± 7.1

Fusion and Selection of Intonation Scores. In this section, we investigate different intonation score combinations at a late fusion stage. That is, we considered the intonation scores computed with the distance metric (euclidean or cosine) which has provided a better performance, and use them to train the imitation classifier. Rather than testing all possible combinations, we run an iterative search approach based on the sequential forward selection (SFS) [17], in which a model is trained with an incremental number of features. Starting with no features, at each iteration the accuracy of the model is tested by adding, one at a time, each of the features that were not selected in a previous iteration. The feature that yields the best accuracy is retained for further processing. The same leave-one-speaker-out cross-validation protocol on the training data partition as previously has been applied.

Feature sets have been chosen with the following order: LLD_F0, EG_slope, LLD_MFCC, PASE-BNF, EG_MFCC, EG_F0, EG_energy, LLD_energy. One can consider that the selection order illustrates the relevance of the features with respect to the task at hand. Therefore, as expected, LLD_F0 is the first feature selected, as it is the one which more directly relates to the task. Afterwards, EG_slope, which is, as well, very much related to the imitation of contours, and LLD_MFCC. Then, the selected set is the PASE-BNF. On the one hand, this is surprising, as it is selected before energy, which is expected to complement the existing information. On the other hand, this is indicative of the properties of this feature set, which does not provide much discriminative power when considered alone, but seems to add robustness when added to the other feature sets. Energy, both LLD and EG come at the end of feature selection. This may be due to the fact that although energy is a parameter of prosody, its correlation with intonation may not be that informative. The blue curve in Fig. 2 shows the mean and deviation accuracy achieved in cross-validation with an increasing number of feature sets selected.

For cross-validation, the best average accuracy was 78.1%, attained for four features selected. Nevertheless, we can see in Fig. 2 that the performance in cross-validation is quite stable across different feature sets, although, as we see from error bars, the variability across speaker folds increases, having a minimum when two feature sets are selected. This can indicate that, although other features can

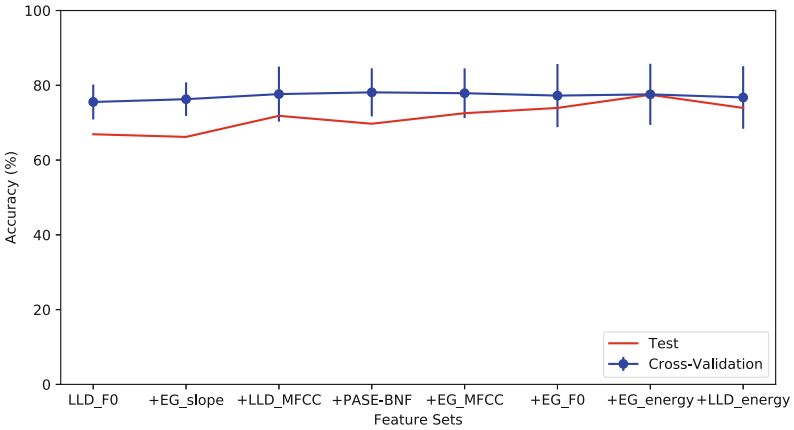

Fig. 2. Accuracy in cross-validation experiments and in the test set for an increasing number of feature sets selected.

Table 2. Accuracy (%) for isolated feature sets in test.

LLD			eGeMAPS				PASE-BNF
F0	MFCC	Energy	F0	MFCC	Energy	Slope	All
66.9	62.0	62.7	49.2	57.7	63.4	60.0	52.8

add relevant information for our task, F0 and EG_slope are the ones achieving the most inter-speaker consistency. Nevertheless, the observations about the relative importance of each feature set and the optimum number of features to select must be taken cautiously due to the considerably high accuracy variance observed.

4.3 Experiments in the Test Set: Results and Discussion

In this subsection, we report the results obtained in the test set when the intonation classifiers are trained using the complete training data set. Table 2 shows results of the intonation imitation classifiers trained using one single intonation score. We notice that, except for the case of EG_F0, there is not a large performance degradation with respect to previous cross-validation experiments. We notice, as well, that in spite of the 8.6% degradation for LLD_F0, this feature set is still the one yielding the best individual results.

When considering the fusion of intonation scores, the order of the selected features follows the order learned in the previous cross-validation experiments. For the best configuration in cross-validation – obtained selecting only four features (78.1%) –, the performance achieved in the test set drops to 69.7%. However, when training the classifiers with seven features, the performance in the test set increases up to 77.5%. This performance difference between the cross-validation and test set results is in line with the degradation already observed from Table 1

to Table 2. The fact that the best result in the test set does not correspond to the set of features which yielded the best result for cross-validation, as pointed out previously, is most likely related with the reduced amount of data and the high variability accuracy observed in cross-validation.

5 Conclusions and Future Work

We investigated the performance of different feature sets for the task of intonation verification. As a single feature, F0 from the LLD set provided the best performance, with 75.5% of well classified pairs in cross-validation and 66.9% in the test set. The fusion of features led to improved results, as 78.1% for the cross-validation experiments and up to 77.5% in test set. F0 and F0 slopes are the first features chosen for the task, as foreseen. Interestingly, the PASE-BNF come as a relevant part in classification, as they improve the classification in train, but they are also the last features to be added when reaching the maximum. This is rather encouraging, in the sense that these bottleneck features have been extracted from a network previously extracted for other languages and tasks, and still add robustness to the results.

As future work, we would like to further investigate the possibilities of fusion, namely using tooklits as BOSARIS [3], which would contribute to better tuning each component. We would also like to re-annotate the data set, so that we had a regressive metric for goodness of imitation and could then improve the comparability of our results. The results we have had with the bottleneck features suggest transfer learning should be further addressed in the future tasks with cross-language information.

Acknowledgements. This work was supported by national funds through Fundação para a Ciência e a Tecnologia (FCT), with reference UIDB/50021/2020, through PhD grant SFRH/BD/139473/2018, and project CMUP-ERI/TIC/0033/2014.

References

1. Arias, J.P., Yoma, N.B., Vivanco, H.: Automatic intonation assessment for computer aided language learning. Speech Commun. **52**(3), 254–267 (2010)
2. Berndt, D.J., Clifford, J.: Using dynamic time warping to find patterns in time series. In: KDD Workshop, Seattle, WA, vol. 10, pp. 359–370 (1994)
3. Brümmer, N., De Villiers, E.: The BOSARIS toolkit user guide: theory, algorithms and code for binary classifier score processing. Documentation of BOSARIS toolkit, p. 24 (2011)
4. Cheng, J.: Automatic assessment of prosody in high-stakes English tests. In: Twelfth Annual Conference of the International Speech Communication Association (2011)
5. Christophe, A., Gout, A., Peperkamp, S., Morgan, J.: Discovering words in the continuous speech stream: the role of prosody. J. Phon. **31**(3–4), 585–598 (2003)
6. Chun, D.M.: Discourse Intonation in L2: From Theory and Research to Practice, vol. 1. John Benjamins Publishing, Amsterdam (2002)

7. Escudero-Mancebo, D., González-Ferreras, C., Aguilar, L., Estebas-Vilaplana, E., Cardeñoso-Payo, V.: Exploratory use of automatic prosodic labels for the evaluation of Japanese speakers of L2 Spanish (2016)
8. Eyben, F., Weninger, F., Gross, F., Schuller, B.: Recent developments in openSMILE, the Munich open-source multimedia feature extractor. In: Proceedings of the 21st ACM International Conference on Multimedia, MM 2013, pp. 835–838. ACM, New York (2013). https://doi.org/10.1145/2502081.2502224, https://doi.org/10.1145/2502081.2502224
9. Eyben, F., et al.: The Geneva Minimalistic Acoustic Parameter Set (GeMAPS) for voice research and affective computing. IEEE Trans. Affect. Comput. **7**(2), 190–202 (2015)
10. Levitt, A.G.: The acquisition of prosody: evidence from French- and English-learning infants. In: de Boysson-Bardies, B., de Schonen, S., Jusczyk, P., McNeilage, P., Morton, J. (eds.) Developmental Neurocognition: Speech and Face Processing in the First Year of Life. NATO ASI Series (Series D: Behavioural and Social Sciences), vol. 69, pp. 385–398. Springer, Dordrecht (1993). https://doi.org/10.1007/978-94-015-8234-6_31
11. Li, K., Wu, X., Meng, H.: Intonation classification for L2 English speech using multi-distribution deep neural networks. Comput. Speech Lang. **43**, 18–33 (2017)
12. Ma, M., Evanini, K., Loukina, A., Wang, X., Zechner, K.: Using F0 contours to assess nativeness in a sentence repeat task. In: Sixteenth Annual Conference of the International Speech Communication Association (2015)
13. Neto, J.P., Martins, C.A., Meinedo, H., Almeida, L.B.: The design of a large vocabulary speech corpus for Portuguese. In: Fifth European Conference on Speech Communication and Technology (1997)
14. Pascual, S., Ravanelli, M., Serrà, J., Bonafonte, A., Bengio, Y.: Learning problem-agnostic speech representations from multiple self-supervised tasks. arXiv preprint arXiv:1904.03416 (2019)
15. Paul, R., Shriberg, L.D., McSweeny, J., Cicchetti, D., Klin, A., Volkmar, F.: Brief report: relations between prosodic performance and communication and socialization ratings in high functioning speakers with autism spectrum disorders. J. Autism Dev. Disord. **35**(6), 861 (2005)
16. Povey, D., et al.: The Kaldi speech recognition toolkit. In: IEEE 2011 Workshop on Automatic Speech Recognition and Understanding. IEEE Signal Processing Society (2011)
17. Pudil, P., Novovičová, J., Kittler, J.: Floating search methods in feature selection. Pattern Recogn. Lett. **15**(11), 1119–1125 (1994)
18. Schuller, B., et al.: The INTERSPEECH 2016 computational paralinguistics challenge: deception, sincerity and native language. In: Proceedings of the 21st ACM International Conference on Multimedia, pp. 2001–2005, September 2016. https://doi.org/10.21437/Interspeech.2016-129
19. da Silva Sousa, M.S.: Prosodic exercises for children with ASD via virtual therapy. Master's thesis, Instituto Superior Técnico, Lisboa, Portugal (2017)
20. Truong, Q.T., Kato, T., Yamamoto, S.: Automatic assessment of L2 English word prosody using weighted distances of F0 and intensity contours. In: Interspeech, pp. 2186–2190 (2018)