Speech Prosody 2022
23-26 May 2022, Lisbon, Portugal

Investigating the usefulness of i-vectors for automatic language characterization

Maureen de Seyssel[1,2], *Guillaume Wisniewski*[2], *Emmanuel Dupoux* [1], *Bogdan Ludusan*[3]

[1]Cognitive Machine Learning (ENS–CNRS–EHESS–INRIA–PSL Research University), France
[2]Université de Paris, CNRS, Laboratoire de linguistique formelle, F-75013 Paris, France
[3]Phonetics Workgroup, Faculty of Linguistics and Literary Studies & CITEC, Bielefeld University, Germany

maureen.deseyssel@gmail.com, bogdan.ludusan@uni-bielefeld.de

Abstract

Work done in recent years has shown the usefulness of using automatic methods for the study of linguistic typology. However, the majority of proposed approaches come from natural language processing and require expert knowledge to predict typological information for new languages. An alternative would be to use speech-based methods that do not need extensive linguistic annotations, but considerably less work has been done in this direction. The current study aims to reduce this gap, by investigating a promising speech representation, i-vectors, which by capturing suprasegmental features of language, can be used for the automatic characterization of languages. Employing data from 24 languages, covering several linguistic families, we computed the i-vectors corresponding to each sentence and we represented the languages by their centroid i-vector. Analyzing the distance between the language centroids and phonological, inventory and syntactic distances between the same languages, we observed a significant correlation between the i-vector distance and the syntactic distance. Then, we explored in more detailed a number of syntactic features and we proposed a method for predicting the value of the most promising feature, based on the i-vector information. The obtained results, an 87% classification accuracy, are encouraging and we envision to extend this method further.

Index Terms: i-vector, language typology, suprasegmental information, prosody, syntax

1. Introduction

Languages differ on a variety of levels, and studying these variations is fundamental in understanding how language is structured [1, 2]. A lot of effort has been put in defining features to classify languages at multiple levels: from phonology [3, 4], morphology and syntax [5] up to semantics [6]. These characterizations of languages are done by expert linguists and have been collected in several typological databases such as WALS [7], PHOIBLE [8] or SSWL [9]. However, this documentation is not complete, as the majority of languages spoken in the world today still lack description in terms of numerous typological features, thus making a general classification of languages difficult to achieve.

Automatic methods of language typology may help with this problem by characterising specific aspects of languages either based on annotated linguistic features [10] or directly from a corpus [11, 12]. Such methods can, in turn, learn to predict missing features [13], or can be used in downstream language-processing models [14]. Although the bulk of the methods that have been proposed in the literature are coming from natural language processing (NLP; see [15] for an overview), there is also some work done towards speech-based language characterization. Those studies include analyses which focus on prosody, either by performing comparative analyses of dialects [16] and languages [17] using suprasegmental information, by employing long-term information for the syntactic description of languages [18], or even by attempting automatic, signal-based, prosodic typology [19].

The results of these studies provide evidence that signal-based approaches, especially those based on prosodic information, may be developed to help automatic typology in different linguistic areas. We investigate here a promising speech representation which captures long-term information, i-vectors, with the goal of aiding the automatic characterization of languages. The i-vectors, features which are able to represent entire utterances of speech into fixed-dimension representations, have been shown to capture speaker-specific characteristics, being initially successfully used in speaker identification applications [20]. Subsequent studies established that these features may capture also language specific characteristics, when language labels are explicitly given, for the task of language identification [21, 22]. Moreover, when the input features used for computing the i-vectors contain prosodic information such as pitch or intensity, this latter type of information is reflected in the composition of the i-vectors [23]. More recent work employing i-vectors for a comparative analysis of dialects [24], has shown that the i-vector distances between the four investigated Latvian dialects correlated with their geographic position (and presumably with the inter-dialect distances, although no objective evaluation was performed to attest this).

We propose to investigate here whether acoustic distance between language, based on i-vectors, can be used to predict various typological distances between languages (Section 3). For this we employ a large set of languages belonging to several linguistic families and we evaluate the method by means of objective distances based on expert linguistic features. In the second part of the study, we explore an approach based on pairwise language distances to directly predict specific features of the given languages (Section 4).

2. General methods

2.1. Materials

We used languages from CommonVoice 6.1 to generate our dataset. This collaborative corpus, an initiative supported by the Mozilla Foundation[1], consists of recordings of people reading

[1]http://voice.mozilla.org/

10.21437/SpeechProsody.2022-94

prompts in various languages and environments. 60 languages were available in the original dataset. We selected utterances from 24 languages[2] to create a balanced dataset, with a total duration of one hour per language, equally split between 60 speakers. A high number of speakers was chosen in order to have a high within-language variability. Preliminary experiments with larger training sizes show that one hour was sufficient for our purposes, while allowing us to employ more languages. The average number of utterances per language set was 761, with an average utterance duration of 4.62 seconds. No significant variance in the number of utterances was observed between languages.

2.2. Training pipeline

We first extracted Mel frequency cepstral coefficient features (MFCCs) [25] for all utterances in the 24 languages, with 13 coefficients including energy (related to intensity), along with double-delta coefficients and pitch information. We then used these features to train a standard i-vector system on all languages, using the Kaldi toolkit [26], with 2,046 Gaussians and i-vectors of dimension 400. In order to maximise the distance between languages, a transformation matrix based on a Linear Discriminant Analysis (LDA) was also computed, and applied to the i-vectors.

Next, we generated i-vector representations for all utterances from our dataset and we calculated the mean i-vector for each language, averaging over all i-vector representations of the language. We call these vectors "centroids".

Finally, we determined the distance between a pair of languages by computing the Euclidean distance between the centroids of those two languages (previous work suggesting that Euclidean distance works best with language i-vectors [24, 27, 28]). Distances were computed for all possible 276 pairs yielded by the languages in our dataset.

3. Experiment 1 : Estimating language distances using i-vectors

In this first experiment, we are comparing the i-vector distances to expert-annotated linguistic distances.

3.1. Linguistic distances

We retrieved the inventory, phonological and syntactic language vectors from the URIEL database [10], having a size of 28, 158 and 103, respectively. These vectors contain various featural information belonging to these three linguistic areas (inventory, phonology and syntax), and were gathered from different typological databases such as WALS and PHOIBLE. To avoid using sparse vectors, missing features were predicted following the method proposed in [13]. We also concatenated, for each language, the vectors corresponding to the three different categories of information into one, which we refer to as the "general" linguistic vector. Distances between languages were then derived for each of the 276 language pairs using the cosine distance, following [10] and [13], for each of the four vectors (three linguistic areas, one general vector).

[2]Arabic, Catalan, Czech, Dutch, Welsh, German, English, Spanish, Basque, Persian, French, Frisian (Netherlands), Italian, Kabyle, Polish, Portuguese, Russian, Kinyarwanda, Swedish, Tamil, Turkish, Tatar, Ukrainian and Mandarin (Mainland China)

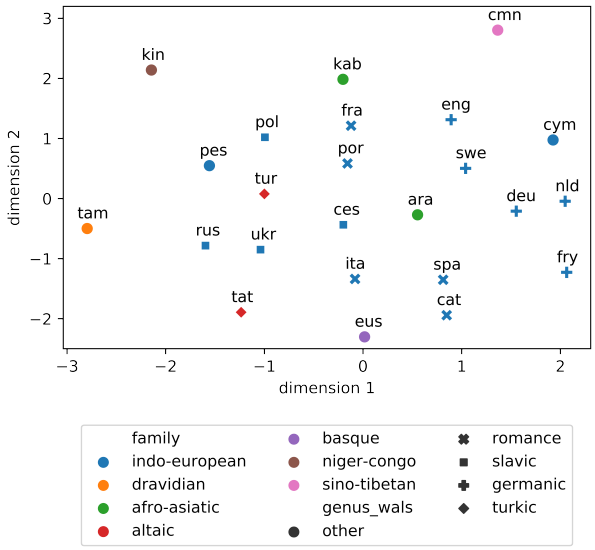

Figure 1: *Visualisation of the centroid language vectors using multidimensional scaling. Colors indicate the language family and shape the language genus as documented in the WALS. Genuses related to only one language are grouped into the "other" category.*

3.2. Results

We applied multidimensional scaling to the centroid i-vectors to visualise how the different languages were scattered around the acoustic space. As shown in Figure 1, the most distinctive languages of our set such as Mandarin, Tamil and Kinyarwanda are separate from the other languages. Similarly, languages sharing common roots seem to located in the same places, as is the case for Russian, Ukrainian, Czech and Polish, or for English, Swedish, German and Dutch. This qualitative analysis seems to corroborate the fact that the information present in the i-vectors may capture some sort of language distance.

We computed the Pearson correlation coefficient between the i-vector distance scores and the general linguistic vectors distance scores for all language pairs, as well as its 95% confidence interval using bootstrapping with 9,999 samples. We then compared it to the correlation obtained when randomly pairing i-vector distances and linguistic distances (for 9,999 times, by resampling with replacement). The correlation is significant when the boundaries of the two confidence intervals do not overlap. As reported in Table 1, the i-vector distances and the general linguistic distances were positively correlated, further supporting the fact that i-vectors encode language information.

We continued our analysis by calculating the correlation between i-vector distances and each of the three categorical linguistic distances, in order to gain more knowledge regarding which type of information is captured by the i-vectors. As presented in Table 1, there was no significant correlation neither between the phonology distances and the i-vector distances nor between the inventory distances and the i-vector distances. The syntactic distances however, yielded a significant positive correlation with the i-vector distance scores. Because the datapoints in our correlations correspond to language pairs and are therefore not totally independent from each other, in addition to computing the random permutations, we also re-ran the analysis on the syntax and i-vector distances correlation removing each

Table 1: *Correlation scores between the i-vector distances and each of the general, phonology, inventory and syntax distances. The median Pearson R value is reported over the bootstrapped alternative hypothesis along with its 95% confidence interval (* indicates significance). CI for the random permutation is also reported.*

	Pearson R	95% CI	Random perm. 95% CI
general	0.52*	[0.44, 0.59]	[-0.29, 0.34]
phonology	0.34	[0.23, 0.44]	[-0.29, 0.36]
inventory	0.22	[0.12, 0.32]	[-0.28, 0.35]
syntax	0.55*	[0.47, 0.62]	[-0.28, 0.33]

time one of the languages (so 23 language pairs). A significant correlation was obtained every time, suggesting that the initial results are robust.

4. Experiment 2: Predicting syntactic features from speech representations

We have seen in Experiment 1 that the distances between the i-vectors centroids correlate best (among the distances investigated here) with the syntactic distance between language pairs. In this experiment we would like to determine which are the syntactic features most correlated with the i-vector distances. Moreover, we conduct a preliminary investigation into using the information given by the i-vector distances to predict values for languages which have not been yet described.

Based on evidence from prosodic phonology [29], showing that prosodic information (the placement of prosodic prominence within phonological phrases) is correlated with the relative order of heads and complements in a language, and from speech processing revealing that long-term information (the shape of the amplitude modulation spectrum) may discriminate between head-complement and complement-head languages [18], we focused our analysis on word order features.

4.1. Methods

We chose those word order features from the WALS for which our languages were represented only by two classes and an optional third class, for "mixed" or "no dominant order". One of the two classes was then coded with the value 1, while the other class with the value 0. In case the optional mixed/no dominant order class existed, it was coded with the value 0.5. We then kept only those features which had at least three instances of languages for each of the two classes (0 or 1). Languages for which their feature value was not recorded in the WALS, were marked with a question sign (see Table 2) and were not used to compute the correlation. The following six features were employed in this experiment:

- 83A: Order of Object and Verb
- 85A: Order of Adposition and Noun Phrase
- 86A: Order of Genitive and Noun
- 87A: Order of Adjective and Noun
- 90A: Order of Relative Clause and Noun
- 93A: Pos. of Interrogative Phrases in Content Questions

Table 2: *The WALS syntactic features employed in this experiment. We used features for which the considered languages had only two distinct values (coded by 0/1) and, optionally, a mixed/no dominant order (coded by 0.5). Features missing a value are coded by a question mark in the table and not used in the correlation computation. The last column shows the prediction of the feature 90A using the proposed approach.*

Lang.	**WALS features**						**Pred.**
	83A	**85A**	**86A**	**87A**	**90A**	**93A**	**90A**
ara	0	0	0	0	1	0.5	1
cat	0	0	0	0	1	?	1
ces	0	0	0.5	1	1	0	1
cmn	0	0.5	1	1	0	0	0
cym	0	0	0	0	1	1	1
deu	0.5	0	0	1	1	1	1
eng	0	0	0.5	1	1	1	1
eus	1	1	1	0	0	0	0.5
fas	1	0	0	0	1	0	1
fra	0	0	0	0	1	1	1
fry	0.5	0	0.5	1	1	1	1
ita	0	0	0	0	1	?	1
kab	0	0	0	0	1	?	1
kin	0	0	?	0	?	0	0
nld	0.5	0	0	1	1	?	1
pol	0	0	0	1	1	1	1
por	0	0	0	0	1	?	1
rus	0	0	0	1	1	1	1
spa	0	0	0	0	1	1	1
swe	0	0	1	1	1	1	1
tam	1	1	1	1	0	0	0
tat	1	1	1	1	0	?	0.5
tur	1	1	1	1	0	0	1
ukr	0	0	?	1	1	?	1

We then determined, for each feature, the distance (*feat_dist*) between all pairs of languages which had values for that particular features, by computing the absolute difference between the value of the two classes. For example, if one class had the value 0 and the other one value 1, the absolute difference between them was equal to 1. Thus, languages belonging to a class always had a difference equal to 0 to the other languages from the same class, a distance of 1 to the instances of the other class and a distance of 0.5 to the mixed/no dominant class elements. The pairwise *feat_dist* for all language pairs was subsequently correlated to the distance between the centroid of the i-vectors of the same pairs of languages. The R software [30] was used to compute the Pearson r correlation coefficient and to test the significance of the correlation.

Finally, we employed the most promising feature (the one having the highest correlation to the i-vector distance) to predict the values of languages, both of those that are described and of those for which no value is given in the WALS. For the languages which had values, we proceeded as follows: we replaced the original value of the language by either 0, 1 or 0.5 and we recomputed *feat_dist* and its correlation to the i-vectors. We then considered as the predicted class the one which gave the highest correlation among the three. Also for the languages without values in the WALS, an identical procedure was applied (the only difference is that we actually consider all the pairs which contain that particular languages, as they were initially not included due to not having a value for that feature).

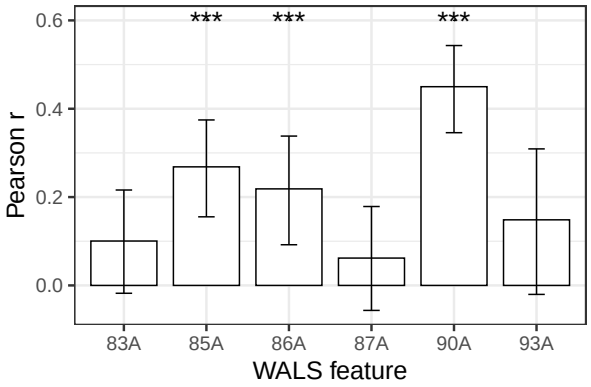

Figure 2: *Obtained correlations between the i-vector distances and the feat_dist corresponding to that particular WALS feature. The error bars represent confidence intervals. The asterisks on top illustrate the significance level (***, $p < 0.001$).*

4.2. Results

The correlation results between the pairwise distances *feat_dist* and the i-vector ones are illustrated in Figure 2. We observe positive low to medium correlations for all the investigated features, with a maximum value of 0.45 obtained for the feature 90A. Three of the feature distances, 85A, 86A and 90A reached significant correlations with the i-vector distances.

We then used the best feature, 90A, to predict the values of the languages that had values for this feature in the WALS, as well as for Kinyarwanda, the only language from our set of 24 languages which was missing a value for that feature. These results are presented in the last column of Table 2. Comparing the predictions with the values given in the WALS we see that most languages are correctly predicted. The three languages which were not correct, belonged to the class 0, with two of them being classified as mixed/no dominant order. The proposed approach predicted Kinyarwanda to be a *Relative Clause - Noun* language (class 0), similar to the prediction made by [13].

5. Discussion and conclusions

Using an i-vector model of language identification, and relying on the average representation of each language in our train set, we were able to compute distances between language pairs. We found that these languages correlated with the general distance from [13], based on the concatenation of multiple expert-annotated linguistic features, at different levels. These results extend those of [24], by showing that i-vectors encode relevant information to discriminate also between languages. Moreover, we evaluated our distances against expert-derived observations, thus providing robust evidence for the suitability of using i-vectors and showing their appropriateness for methods for automatic language characterisation.

One of the main advantage of this approach is that only relatively small amount of speech per language is required (here we used one hour, but we could probably reduce it further). However, it is important to have sufficient within-language variability in the training set languages (e.g., by increasing the number of different speakers or recording conditions). An alternative would be to first train a model on a fixed number of languages which have enough data, and use it to compute representation vectors of novel languages with less data. Assuming we have an adequate amount of data and diversity among the languages in the training set, it might be possible to compute distances to new languages with only a few utterances. Finally, whilst not reported here, we also found a significant, although slightly weaker, correlation when no Linear Discriminant Analysis was applied, suggesting that the model can capture language characteristics even in a totally unsupervised fashion.

Having found that the i-vector distances correlated with general expert-derived linguistic features, we analysed further whether there were particular levels of linguistics that correlated with this new distance. We looked at three different levels: inventory, phonology and syntax, and found that the syntax-derived distances yielded a significant correlation with the i-vector distances. The fact that the i-vector distances do not correlate with neither inventory nor phonology was surprising but not unexpected, as previous attempts to take into consideration phoneme information using i-vectors were done by modeling phoneme information in a supervised fashion [31, 32]. Further analyses will also be required to determine whether the structure of the employed corpus might have had an effect on these results. Finally, the fact that syntactic distances correlated with i-vector distances can be explained by the links between prosody and syntax (e.g. [29]), the former type of information being likely captured by our model.

In order to better investigate which syntactic distances might relate to those captured by i-vector distances, we tested six word-order features from the WALS. We observed that three features significantly correlated with our i-vector distances, with two of them capturing phrase-level word order characteristics. These results are in line with the prosodic phonology theory [29], stating that prosody information may help determine word order, as well as with the findings of previous speech processing studies (e.g. [18]). Finally, we found that our approach was able to correctly predict the 90A feature for 20 out of the 23 languages for which we had this information, and that the value it predicted for the only language missing this information (Kinyarwanda) was the same as the one predicted by the method in [13]. These preliminary results are promising in that they suggest that i-vectors could potentially be used in prediction of missing linguistic features.

We can see multiple applications to using i-vector models for language characterization. First, their output could be employed in downstream speech processing tasks, in the same way as text-based models are used in downstream NLP tasks, for example to select which languages to pretrain models from, in the case of under-resourced speech recognition. Secondly, the preliminary results we obtained on feature prediction are encouraging in that such models can bring additional knowledge to be used in predicting some features, particularly syntactic. Future work could focus on using these representations along with supervised or unsupervised learning paradigms, rather than with correlations, for determining specific language features.

6. Acknowledgements

This work was performed using HPC resources from GENCI-IDRIS (Grant 20XX-AD011012315). MS work was partly funded by l'Agence de l'Innovation de Défense.

7. References

[1] J. H. Greenberg, "The nature and uses of linguistic typologies," *International Journal of American Linguistics*, vol. 23, no. 2, pp. 68–77, 1957.

[2] B. Comrie, "Linguistic typology," *Annual Review of Anthropology*, vol. 17, no. 1, pp. 145–159, 1988.

[3] R. Jakobson, *Child language, aphasia and phonological universals*. De Gruyter Mouton, 2014.

[4] L. M. Hyman, "Where's phonology in typology?" *UC Berkeley PhonLab Annual Report, 2(2)*, 2007.

[5] B. Comrie, *Language universals and linguistic typology: Syntax and morphology*. University of Chicago press, 1989.

[6] N. Evans *et al.*, *Semantic typology*. Oxford University Press, 2010.

[7] M. S. Dryer and M. Haspelmath, "The world atlas of language structures online (max planck institute for evolutionary anthropology, leipzig)," *Available at wals. info. Accessed October*, vol. 9, p. 2014, 2013.

[8] S. Moran and D. McCloy, Eds., *PHOIBLE 2.0*. Jena: Max Planck Institute for the Science of Human History, 2019. [Online]. Available: https://phoible.org/

[9] C. Collins and R. Kayne, "Syntactic structures of the world's languages," *New York: New York University*, 2009.

[10] P. Littell, D. R. Mortensen, K. Lin, K. Kairis, C. Turner, and L. Levin, "Uriel and lang2vec: Representing languages as typological, geographical, and phylogenetic vectors," in *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*, 2017, pp. 8–14.

[11] B. Snyder and R. Barzilay, "Unsupervised multilingual learning for morphological segmentation," in *Proceedings of acl-08: hlt*, 2008, pp. 737–745.

[12] S. B. Cohen and N. A. Smith, "Shared logistic normal distributions for soft parameter tying in unsupervised grammar induction," in *Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, 2009, pp. 74–82.

[13] C. Malaviya, G. Neubig, and P. Littell, "Learning language representations for typology prediction," *arXiv preprint arXiv:1707.09569*, 2017.

[14] H. O'Horan, Y. Berzak, I. Vulić, R. Reichart, and A. Korhonen, "Survey on the use of typological information in natural language processing," in *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, 2016, pp. 1297–1308.

[15] E. M. Ponti, H. O'horan, Y. Berzak, I. Vulić, R. Reichart, T. Poibeau, E. Shutova, and A. Korhonen, "Modeling language variation and universals: A survey on typological linguistics for natural language processing," *Computational Linguistics*, vol. 45, no. 3, pp. 559–601, 2019.

[16] A. Suni, M. Włodarczak, M. Vainio, and J. Šimko, "Comparative Analysis of Prosodic Characteristics Using WaveNet Embeddings," in *Proc. Interspeech 2019*, 2019, pp. 2538–2542.

[17] J. Šimko, A. Suni, K. Hiovain, and M. Vainio, "Comparing Languages Using Hierarchical Prosodic Analysis," in *Proc. Interspeech 2017*, 2017, pp. 1213–1217.

[18] L. Varnet, M. C. Ortiz-Barajas, R. G. Erra, J. Gervain, and C. Lorenzi, "A cross-linguistic study of speech modulation spectra," *The Journal of the Acoustical Society of America*, vol. 142, no. 4, pp. 1976–1989, 2017.

[19] U. Reichel, K. Mády, and S. Benus, "Acoustic profiles for prosodic headedness and constituency," in *Proc. Speech Prosody 2018*, 2018, pp. 699–703.

[20] N. Dehak, P. J. Kenny, R. Dehak, P. Dumouchel, and P. Ouellet, "Front-end factor analysis for speaker verification," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 19, no. 4, pp. 788–798, 2010.

[21] D. Martinez, L. Burget, L. Ferrer, and N. Scheffer, "ivector-based prosodic system for language identification," in *2012 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2012, pp. 4861–4864.

[22] D. Martinez, O. Plchot, L. Burget, O. Glembek, and P. Matějka, "Language recognition in ivectors space," in *Twelfth annual conference of the international speech communication association*, 2011.

[23] D. Martinez, E. Lleida, A. Ortega, and A. Miguel, "Prosodic features and formant modeling for an ivector-based language recognition system," in *2013 IEEE International Conference on Acoustics, Speech and Signal Processing*. IEEE, 2013, pp. 6847–6851.

[24] A. A. Bērziņš, "Usage of i-vectors for automated determination of a similarity level between languages," *Proceedings of the Institute for System Programming of the RAS (Proceedings of ISP RAS)*, vol. 31, no. 5, pp. 153–164, 2019.

[25] P. Mermelstein, "Distance measures for speech recognition, psychological and instrumental," *Pattern recognition and artificial intelligence*, vol. 116, pp. 374–388, 1976.

[26] D. Povey, A. Ghoshal, G. Boulianne, L. Burget, O. Glembek, N. Goel, M. Hannemann, P. Motlicek, Y. Qian, P. Schwarz *et al.*, "The kaldi speech recognition toolkit," in *IEEE 2011 workshop on automatic speech recognition and understanding*, no. CONF. IEEE Signal Processing Society, 2011.

[27] H. Behravan, T. Kinnunen, and V. Hautamäki, "Out-of-set i-vector selection for open-set language identification." in *Odyssey*, vol. 2016, 2016, pp. 303–310.

[28] E. San Segundo, A. Tsanas, and P. Gómez-Vilda, "Euclidean distances as measures of speaker similarity including identical twin pairs: a forensic investigation using source and filter voice characteristics," *Forensic Science International*, vol. 270, pp. 25–38, 2017.

[29] M. Nespor and I. Vogel, *Prosodic phonology*. De Gruyter Mouton, 2012.

[30] R Core Team, *R: A language and environment for statistical computing*, R Foundation for Statistical Computing, Vienna, Austria, 2020. [Online]. Available: https://www.R-project.org/

[31] J. Franco-Pedroso and J. Gonzalez-Rodriguez, "Linguistically-constrained formant-based i-vectors for automatic speaker recognition," *Speech Communication*, vol. 76, pp. 61–81, 2016.

[32] L. F. D'Haro Enríquez, O. Glembek, O. Plchot, P. Matějka, M. Sou?far, R. de Córdoba Herralde, and J. Černocký, "Phonotactic language recognition using i-vectors and phoneme posteriogram counts," in *InterSpeech 2012 - 13th Annual Conference of the International Speech Communication Association*, 2012, pp. 1–4. [Online]. Available: http://oa.upm.es/20403/