Automatic Evaluation of Speech Intelligibility Based on i-vectors in the Context of Head and Neck Cancers

Imed Laaridh, Corinne Fredouille, Alain Ghio, Muriel Lalain, Virginie Woisard

Interspeech 2018
2-6 September 2018, Hyderabad

Automatic Evaluation of Speech Intelligibility Based on i-vectors in the Context of Head and Neck Cancers

Imed Laaridh[1], *Corinne Fredouille*[1], *Alain Ghio*[2], *Muriel Lalain*[2], *Virginie Woisard*[3]

[1]University of Avignon, LIA, France
[2]Aix-Marseille Univ, CNRS, LPL UMR 7309, Aix-en-Provence, France
[3]Service ORL, CHU Larrey, URI Octogone-Lordat, Toulouse, France

imed.laaridh@alumni.univ-avignon.fr, corinne.fredouille@univ-avignon.fr,
{alain.ghio, muriel.lalain}@univ-amu.fr, woisard.v@chu-toulouse.fr

Abstract

In disordered speech context, and despite its well-known subjectivity, perceptual evaluation is still the most commonly used method in clinical practice to evaluate the intelligibility level of patients' speech productions. However, and thanks to increasing computing power, automatic speech processing systems have witnessed a democratization in terms of users and application areas including the medical practice.
In this paper, we evaluate an automatic approach for the prediction of cancer patients' speech intelligibility based on the representation of the speech acoustics in the total variability subspace based on the i-vector paradigm. Experimental evaluations of the proposed predictive approach have shown a very high correlation rate with perceptual intelligibility when applied on the French speech corpora C2SI (r=0.84). They have also demonstrated the robustness of the approach when using a limited amount of disordered speech per patient, which may lead to the redesign and alleviation of the test protocols usually used in disordered speech evaluation context.
Index Terms: automatic speech processing, speech disorders, head and neck cancers, i-vectors, intelligibility

1. Introduction

Speech disorders may affect, depending on their underlying causes, different components of speech production (respiration, phonation, articulation, prosody, etc.). Different measures have been studied in the literature to assess speech quality such as intelligibility, comprehensibility and the severity of the speech alteration. Hence, many evaluation protocols have been designed to gather such measures. These protocols can be related to specific voice and speech disorders such as dysarthria [1], dysphonia [2] or applied to multiple speech disorders [3]. They aim at helping clinicians in their knowledge of the speech impairment and its clinical evaluation, crucial for following the condition progression of patients in the case of treatment or/and of speech rehabilitation. In this context, perceptual evaluation is still the most used method for disordered speech evaluation in clinical practice despite its documented limitations such as irreproducibility and subjectivity[4].

Intelligibility loss is one of the most frequent complaint encountered among patients suffering from speech disorders. To cope with the limitations reported above, automatic approaches have been seen, very early, as potential solutions by providing objective tools for intelligibility assessment. In the literature, we can distinguish two main kinds of approaches: those directly based on automatic speech transcription systems resulting in a word transcription error rate as an intelligibility score [5], and those which are using automatic speech processing technologies so that relevant information can be extracted from speech and used within an automatic prediction system of the degree of intelligibility [6, 7]. In parallel, other automatic approaches have also demonstrated indirect abilities to provide intelligibility scores, as for instance in [8], in which the approach is focused on a sharper analysis of dysarthric speech and is dedicated to the detection of anomalies.

The i-vector paradigm is a state-of-the-art approach successfully applied in speaker recognition applications [9]. It is proven to represent well speaker characteristics [10]. This paradigm has been adapted to several other contexts such as language recognition and even speech quality evaluation. In [11], this representation, combined with a large set of acoustic, syllable-level, and phonotactic features, was used for the automatic prediction of UPDRS ratings of Parkinson's Disease patients, in the specific context of Interspeech 2015 ComParE challenge. In [12], the i-vector paradigm was used as a speaker normalization and involved in a more complex classification approach, combining acoustic and articulatory features for the automatic detection of Amyotrophic Lateral Sclerosis (ALS). In [13], i-vectors were used for the representation of word segments produced by 15 dysarthric speakers resulting in some important correlations between automatically predicted and reference intelligibility measures. Finally, in [14], the authors proposed an approach based on a cosine distance between the i-vector representation of a speech production (test) and two reference i-vectors representing each normal and dysarthric speech.

In previous work [15], we proposed an approach based on the i-vector paradigm for the automatic prediction of several dysarthric speech evaluation metrics like intelligibility, severity, and articulation impairment. The proposed approach was applied on 129 dysarthric and healthy speakers and high correlation measures (between 0.8 and 0.9) were reached between the different automatic predictions and reference perceptual speech evaluation metrics.

In this paper, the same automatic i-vector-based approach is used for the automatic prediction of speech intelligibility measure in the context of Head and Neck Cancer (HNC) patients. Here, automatic prediction scores are compared to intelligibility measures issued from an original perceptual protocol based on pseudo-word production by patients and healthy speakers. In addition to the evaluation of the i-vector-based approach (in terms of performance), this protocol permits to investigate the impact of the recording length as well as the one of the speech material on the precision of the automatic prediction score of speech intelligibility.

10.21437/Interspeech.2018-1266

This work contributes to a larger research project dedicated to the automatic evaluation of the Quality of Life (QoL) after cancer on the basis of speech quality measures. The rest of this paper is organized as follows. In section 2, the corpus and its perceptual evaluation used in this study is described. The automatic approach architecture is described in section 3. In section 4, the behavior of the proposed approach is investigated whereas section 5 provides a conclusion and directions for future work.

2. Corpus description

2.1. Population

The current study is based on the French HNC speech corpus C2SI [16]. This corpus includes patients suffering from oral cavity or oropharyngeal cancer and healthy speakers, who were asked to record different speech production tasks like sustained /a/, read speech, picture description, spontaneous speech, and isolated pseudo-words. All the patients have undergone a cancer treatment consisting in surgery and/or radiotherapy and/or chemotherapy.
In this study, only the isolated pseudo-words, referred to as DAP (*Décodage Acoustico-Phonétique*, i.e.: Acoustic-Phonetic Decoding) later in the rest of paper, are involved. During this task, all speakers have pronounced 52 pseudo-words. Each pseudo-word had the following phonotactic structure: $C(C)_1V_1C(C)_2V_2$ where $C(C)_i$ was an isolated consonant C or a consonant group CC. These elements were selected from a list of 18 consonants and 16 consonant group of French; 8 different vowels could be selected.
Using this combinatory method, 95000 pseudo-words were automatically generated. This set was then reduced to 90000 elements after the removal of semantically correct French words. For each speaker, a list of 52 items was then randomly built from the 90000 pseudo-words database. All the lists contained the same number of consonants and vowels but with different combinations, which makes the lists equivalent but different. These variations are necessary to avoid learning effects for the listeners who then transcribed the productions of the speakers. In total, 85 patients and 41 healthy speakers were recorded for this task. Some patients did not complete the reading task due to an extreme fatiguability and produced less than the 52 required pseudo-words.

2.2. Perceptual intelligibility evaluation

All of the pseudo-words pronounced by the speakers have been transcribed by 40 listeners. To alleviate the transcription time, each pseudo-word was evaluated by at least 3 of them. The listeners, naives but required to have good spelling level, were confronted with a task that resembles acoustic-phonetic decoding (hence the name of the speech production task) followed by a written transcription. In total, 18360 orthographic transcriptions of the pseudo-words were gathered. The annotation was performed using the Lancelot-Perceval platform [1]. Each listener could listen to each pseudo-word up to 3 times but has then to give his/her transcription.

For each pseudo-word, the average Levenshtein distance, considering acoustic distinctive features, between the expected item and the transcribed responses produced by listeners was estimated. The average distance was then computed, for each speaker, and is considered as an (un)intelligibility measure (high values correspond to the greater distance between the expected pseudo-word and the transcribed response and so characterize the least intelligible speakers).

[1] www.lpl-aix.fr/~lpldev/perceval

3. Methodology

The proposed approach studied here relies on two steps. The first one consists in the parameterization and the representation of each speech utterance in the total variability subspace. Thus, each recording associated with a healthy speaker or a patient is represented with an i-vector [9].
The second step is a regression from the i-vector subspace to the intelligibility space (1 dimension). Support Vector Regression (SVR) is used, considering the limited amount of annotated data available for the study. These phases are reported in figure 1. Despite the large number of patients and healthy speakers available considering the pathological speech context, the amount of data remains limited compared to other "standard" automatic speech processing applications.

3.1. The total variability subspace

The total variability paradigm was first introduced in the context of automatic speaker recognition. It consist on the conversion of an acoustic vector (recording) into a single low-dimensional vector representing the whole speech utterance. The speaker- and session-dependent super-vector s of concatenated Gaussian Mixture Model (GMM) means is assumed to obey a linear model of the form :

$$s = m + Tw \tag{1}$$

where m is the mean super-vector of the Universal Background Model (UBM), T is the low-rank projection matrix trained using a large dataset by MAP estimate (it represents the "total variability" subspace) and w is a latent variable, called "i-vector", having a standard normal distribution $\mathcal{N}(0, I)$. For more details, the algorithms for the estimate of T and the extraction of i-vectors are described in [17].

3.2. I-vector extraction

In this work, the parametrization used consisted on 19 LFCC, their 19 first (Δ) and 11 second ($\Delta\Delta$) derivatives. Also, a mean and variance normalization (MVN) was applied to the LFCC features computed on the speech portions of each recording (silences and pauses were eliminated), detected using an automatic text-constrained phone alignment. Then, a gender-dependent 512 component UBM and a total variability matrix T of low rank 400 estimated using French Ester 1&2, REPERE and ETAPE speech corpora (7690 sessions from 2906 speakers) [18] are used to extract one i-vector per speech recording. The LIA_SpkDet package of the ALIZE open source toolkit [19, 20] is used for the estimate of the total variability matrix and the i-vector extraction.

3.3. Support Vector Regression

The basic idea of Support Vector Regression is to find a function that maps between a representation and a prediction subspaces. In ϵ-SVR, this consist in finding a function that has, at most, ϵ deviation from target reference values for all the training data. When such a task is not feasible, trade-off and slack variables are introduced to cope with the optimization problem [21].
For each test vector, and given the training vectors $x_i \in R^{400}$,

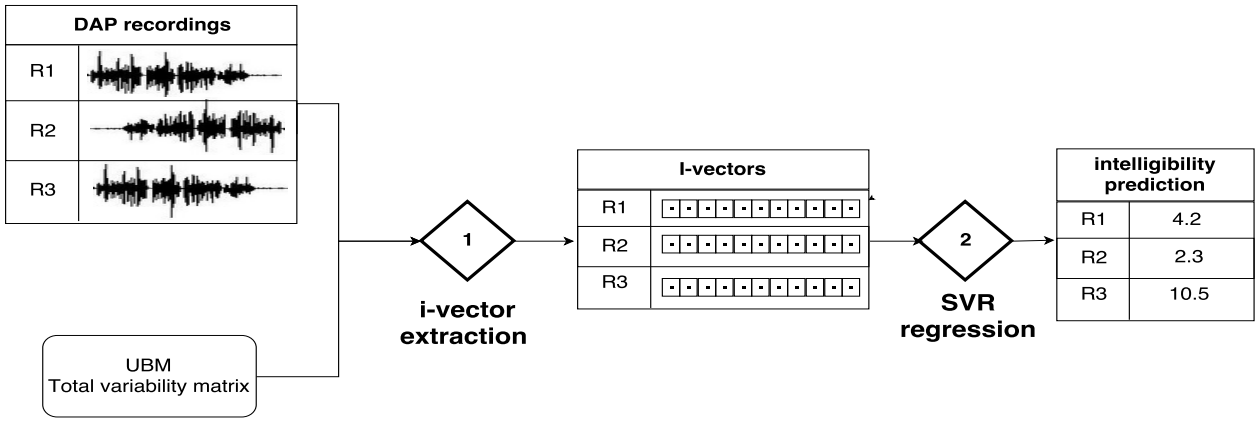

Figure 1: *Representation of the different phases of the proposed methodology.*

$i = 1, ..., n$, the decision function is:

$$f(x) = \sum_{i=1}^{n} (\alpha_i - \alpha_i^*) K(x_i, x) + b \qquad (2)$$

where α_i and α_i^* are Lagrange multipliers, K is the kernel function and b is the bias. In this work, RBF kernels were used. Also, and considering the high performance achieved by the i-vector representation in speaker recognition context, a 10 fold cross-validation process was implemented to avoid bias related to same speakers being present on both training and testing phases.

4. Results and discussions

To evaluate the performance of the automatic approach on the C2SI corpus, Pearson Correlation (r) and Root Mean Square Error (RMSE) measures were computed between the automatically predicted and perceptually evaluated intelligibility scores (described in section 2.2).

4.1. Intelligibility prediction at the speaker level

The first experiment we performed consisted in the automatic prediction of the intelligibility of each speaker using his/her entire speech recording, in which he/she pronounced the 52 pseudo-words. This meant that we had a large amount of data to estimate the i-vector representing each speaker's acoustic production. Figure 2 depicts the automatically predicted intelligibility measure compared to the reference perceptual evaluation. We observe that the automatic approach is capable of performing a good separation between the patients and healthy speaker groups. The regression slope confirms the system ability to detect and represent the loss of intelligibility measured perceptually by the listeners. Indeed, r and RMSE measures reach 0.84 and 2.339 respectively. This correlation rate is consistent with previous results observed over read speech produced by dysarthric patients [15]. Also, the resulting RMSE measure is quite low considering that the interval of the reference measure is characterized by a wide range ([0,22] for this corpus) and an extreme sensibility. Indeed, the perceptual evaluation is measured on a discrete interval, where the smallest difference for any phoneme, in terms of acoustic distinctive features, between the reference and the manual pseudo-word transcription is, at least, a 1 point distance).

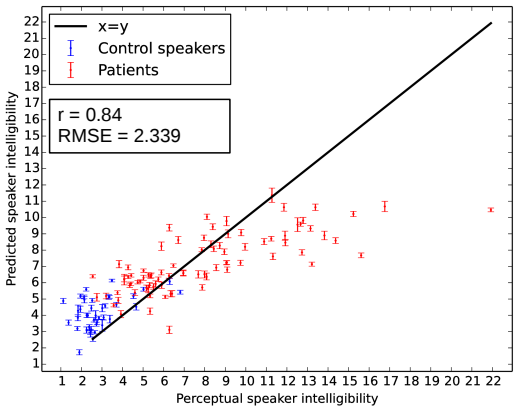

Figure 2: *Automatically predicted intelligibility computed at the recording level (52 pseudo-words) according to perceptual evaluation and a slope 1 line (black).*

4.2. Intelligibility prediction at the sublist level

Since the approach capacity of predicting the intelligibility measure on pseudo-words is confirmed, we propose here to study the amount of speech required to perform a reliable prediction. This is important considering that the heaviness of the task performed by patients during the test protocol is crucial in disordered speech evaluation context. As reported in section 2, fatiguability may prevent patients from completing the reading task of 52 pseudo-words.

Considering that almost each speaker produced a list of 52 pseudo-words, we divided each speaker's list into 5 sublists of around 10 words each. Each sublist represented around 7 seconds only of speech. Apart from their size, the pseudo-word distribution among the sublists was totally random and did not consider their structure or composing phonemes. Each sublist was assigned a perceptual intelligibility measure by averaging distances measured on their pseudo-words (section 2.2). In total, 623 sublists were extracted representing the 126 speakers. A 10 fold cross-validation was implemented on the speaker level to avoid bias resulting from the use of sublists produced by the same speaker in both train and test phases.

Figure 3 depicts the automatically predicted intelligibility measure compared to the reference perceptual evaluation on each sublist. We observe that the discriminative attribute of the approach between patients and healthy speakers is maintained and that the 10 pseudo-word sublists used were sufficient for the automatic approach to detect the intelligibility loss for the patients.

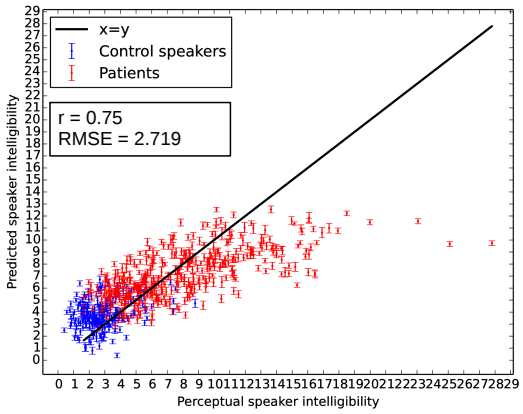

Figure 3: *Automatically predicted intelligibility computed at the subset level (∼10 pseudo-words) according to perceptual evaluation and a slope 1 line (black).*

Considering both the r and RMSE measures, we note that even though a high correlation measure between the automatic predicted and perceptual intelligibility scores is reached (0.75), a 0.09 absolute loss is observed compared with the prediction score reached with the totality of the 52 pseudo-words. Nonetheless, we support that the performance loss is not extreme and does not challenge the automatic prediction reliability since the RMSE still remains quite low with a value of 2.719 against 2.339 considering the 52 pseudo-words.

Also, we computed, for each speaker, an average of his/her sublist-based predicted intelligibility scores and compared it with the reference speaker perceptual evaluation used in the first experiment. The r and RMSE measures reach 0.84 and 2.354 respectively, which is quite similar to values observed in when the prediction was performed at the speaker level.

4.3. Discussion

One key assumption made in the previous experiment was that the different sublists, extracted randomly, were equivalent and carried the same information from an intelligibility point of view. However, we observed that the correlation measure between the predicted and perceptual evaluations computed over the 5 sublists extracted per speaker could vary between 0.72 to 0.80. This variability meant that the sublists, and therefore the pseudo-words associated with, were considered differently by the automatic prediction system, and the i-vector-based approach. Based on this observation, the choice of sublists to use for the prediction task may have a major impact.

In order to highlight this behavior, table 1 presents the best (or worst) r and RMSE measures that could be achieved if we only consider the sublists that minimize (or maximize) the prediction errors. We observe that the correlation measure could reach up to 0.87 and the RMSE only 1.685 if the sublists used for the evaluation are well chosen. In contrast, the r measure drops down to 0.58 if the sublists used hold "less significant" words. Even though, in this case, the performance loss is important, this value could be seen as a threshold in terms of bad predictions for the automatic approach.

Table 1: *Pearson correlation (r) and Root Mean Square Error (RMSE) measures between automatically predicted and perceptual intelligibility scores when using the best and worst sublist selections.*

	r	RMSE
Best sublist selection	0.87	1.685
Worst sublist selection	0.58	4.278

As seen above, the pseudo-word discriminative potential in terms of intelligibility can notably vary. Therefore, in addition to the sophisticated method implemented to design the initial list of 90000 pseudo-words (section 2.1), a more thoughtful way of building the sublists used for speech production by speakers should result in even higher gains in automatic prediction precision (for example: maximizing the presence of consonants in the speech). This observation can be highly useful when implementing new protocols for the automatic/perceptual evaluation of disordered speech. Indeed, most of the battery of assessment tests are substantial and require much effort from both the patient and the evaluating listener. A more relevant linguistic unit content selection can be extremely helpful to drastically alleviate them.

5. Conclusion

This paper investigates an automatic approach for the prediction of speech intelligibility based on the i-vector paradigm and Support Vector Regression-based models. This approach was applied on a dedicated reading protocol of pseudo-words produced by speakers suffering from a head or neck cancers. High correlation (0.84) was achieved between the automatically predicted and perceptual evaluations when using 52 pseudo-words per speaker. Moreover, the approach proved to be stable and robust to the lack of data since r=0.75 was achieved when using only about 20% of each speaker's speech production (∼10 pseudo-words). In addition, the effect of the sublists used for the evaluation, and consequently of the relevancy of pseudo-words associated with in terms of intelligibility prediction, was established. Future work will investigate information carried by each pseudo-word and the impact of its phonetic content on the intelligibility evaluation, from the perceptual and automatic point of view.

6. Acknowledgements

This work has been carried out thanks to the French National Cancer Institute project (INCA - C2SI project). We deeply thank Moez Ajili for his help providing models of the UBM and the total variability matrix realized within the FABIOLE project (ANR-12-BS03-0011).

7. References

[1] P. Enderby, "Frenchay dysarthric assessment," *Pro-Ed, Texas*, 1983.

[2] F. L. Wuyts, M. S. De Bodt, G. Molenberghs, M. Remacle, L. Heylen, B. Millet, K. Van Lierde, J. Raes, and P. H. Van de

Heyning, "The dysphonia severity index: an objective measure of vocal quality based on a multiparameter approach," *Journal of Speech, Language, and Hearing Research*, vol. 43(3), pp. 796–809, 2000.

[3] A. Lowit and R. D. Kent, *Assessment of motor speech disorders*. Plural publishing, 2010, vol. 1.

[4] S. Fex, "Perceptual evaluation," *Journal of voice*, vol. 6, no. 2, pp. 155–158, 1992.

[5] H. Christensen, S. Cunningham, C. Fox, P. Green, and T. Hain, "A comparative study of adaptive, automatic recognition of disordered speech," in *Proceedings of Interspeech'12*, Portland, USA, 2012.

[6] C. Middag, J.-P. Martens, G. Van Nuffelen, and M. De Bodt, "Automated intelligibility assessment of pathological speech using phonological features," *EURASIP Journal on Advances in Signal Processing*, vol. 2009, no. 1, pp. 1–9, 2009.

[7] T. Khan, J. Westin, and M. Dougherty, "Classification of speech intelligibility in parkinson's disease," *Biocybernetics and Biomedical Engineering*, vol. 34(1), pp. 35–45, 2014.

[8] I. Laaridh, C. Fredouille, and C. Meunier, "Automatic detection of phone-based anomalies in dysarthric speech," *ACM Transactions on accessible computing*, vol. 6, no. 3, pp. 9:1–9:24, May 2015.

[9] 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, 2011.

[10] P. Verma and P. K. Das, "i-vectors in speech processing applications: a survey," *International Journal of Speech Technology*, vol. 18, no. 4, pp. 529–546, 2015.

[11] G. An, D. G. Brizan, M. Ma, M. Morales, A. R. Syed, and A. Rosenberg, "Automatic recognition of unified parkinsons disease rating from speech with acoustic, i-vector and phonotactic features," in *Proceedings of Interspeech'15*, Dresden, Allemagne, September 2015.

[12] J. Wang, P. V. Kothalkar, B. Cao, and D. Heitzman, "Towards automatic detection of amyotrophic lateral sclerosis from speech acoustic and articulatory samples," in *Proc. of INTERSPEECH*, 2016.

[13] D. Martínez, E. Lleida, P. Green, H. Christensen, A. Ortega, and A. Miguel, "Intelligibility assessment and speech recognizer word accuracy rate prediction for dysarthric speakers in a factor analysis subspace," *ACM Transactions on Accessible Computing (TACCESS)*, vol. 6, no. 3, p. 10, 2015.

[14] N. Garcia, J. R. Orozco-Arroyave, L. DHaro, N. Dehak, and E. Nöth, "Evaluation of the neurological state of people with parkinsons disease using i-vectors," *Proceedings of the 18th INTERSPEECH (2017, in Press) Google Scholar*, 2017.

[15] I. Laaridh, W. Ben Kheder, C. Fredouille, and C. Meunier, "Automatic prediction of speech evaluation metrics for dysarthric speech," *Proc. Interspeech 2017*, pp. 1834–1838, 2017.

[16] C. Astesano, M. Balaguer, J. Farinas, C. Fredouille, P. Gaillard, A. Ghio, L. Giusti, I. Laaridh, M. Lalain, B. Lepage, J. Mauclair, O. Nocaudie, J. Pinquier, O. Pont, G. Pouchoulin, P. Michele, D. Robert, E. Sicard, and V. Woisard, "Carcinologic Speech Severity Index Project: A Database of Speech Disorders Productions to Assess Quality of Life Related to Speech After Cancer," in *Language Resources and Evaluation Conference (LREC), Miyazak, Japon*. http://www.elra.info: European Language Resources Association (ELRA), mai 2018.

[17] D. Matrouf, N. Scheffer, B. G. Fauve, and J.-F. Bonastre, "A straightforward and efficient implementation of the factor analysis model for speaker verification." in *Interspeech*, 2007, pp. 1242–1245.

[18] M. Ajili, J.-F. Bonastre, W. Ben Kheder, S. Rossato, and J. Kahn, "Phonetic content impact on forensic voice comparison," in *Spoken Language Technology Workshop (SLT), 2016 IEEE*. IEEE, 2016, pp. 210–217.

[19] J.-F. Bonastre, N. Scheffer, D. Matrouf, C. Fredouille, A. Larcher, A. Preti, G. Pouchoulin, N. W. Evans, B. G. Fauve, and J. S. Mason, "Alize/spkdet: a state-of-the-art open source software for speaker recognition." in *Odyssey*, 2008, p. 20.

[20] A. Larcher, J.-F. Bonastre, B. G. Fauve, K.-A. Lee, C. Lévy, H. Li, J. S. Mason, and J.-Y. Parfait, "Alize 3.0-open source toolkit for state-of-the-art speaker recognition." in *Interspeech*, 2013, pp. 2768–2772.

[21] A. J. Smola and B. Schölkopf, "A tutorial on support vector regression," *Statistics and computing*, vol. 14, no. 3, pp. 199–222, 2004.