USE OF ACOUSTIC LANDMARKS AND GMM-UBM BLEND IN THE AUTOMATIC DETECTION OF PARKINSON'S DISEASE

L. Moro-Velazquez[1*], J. I. Godino-Llorente [1, 3], J. A. Gómez-García[1], J. Villalba[2], S. Shattuck-Hufnagel[3], N. Dehak[2]

[1]Centro de Tecnología Biomédica, Universidad Politécnica de Madrid, Madrid, España
[2]Center for Language and Speech Processing, Johns Hopkins University, Baltimore, USA
[3]Speech Communication Group, Research Laboratory of Electronics, MIT, Boston, USA.
laureano.moro@upm.es, igodino@ics.upm.es, jorge.gomez.garcia@upm.es, jvillal7@jhu.edu, sshuf@mit.edu, ndehak3@jhu.edu

Abstract: **New tools based on speech analysis can improve and accelerate diagnosis of Parkinson's Disease. In this work, the use some specific segments of speech, around the so called Acoustic Landmarks, are used with different families of features such as acoustic cues or Rasta-PLP and GMM-UBM-Blend classification methods to detect Parkinson's Disease. Results of 87% of accuracy are obtained.**

Burst segments provide the most relevant information when detecting Parkinson's Disease while GMM-UBM-Blend is revealed as a promising technique when using small databases and segmented speech.

Keywords: **Parkinson's Disease, GMM-UBM, Acoustic Landmarks, Rasta-PLP.**

I. INTRODUCTION

Diagnosis of Parkinson's Disease (PD) is a challenging task which might require several years, depending on the patient. New tools based on motor analysis such as speech analysis can provide the means to do a more rapid and robust diagnosis.

Literature reports multiple efforts to detect and assess PD using voice and speech. These works can be classified as phonatory, articulatory, prosodic and linguistic, depending on the type of material employed and the analyzed speech/voice features.

In the present work, which can be framed into the articulatory group, the detection of PD is performed employing some specific points of speech, called Acoustic Landmarks [1], which are detected along several speech tasks. Some acoustic measurements associated to these landmarks (acoustic cues), or Rasta-Perceptual Linear Predictive (Rasta-PLP) features calculated over several time windows around the landmarks, are used to detect the presence of PD, employing GMM and GMM-UBM Blend classification techniques in two different databases.

Acoustic Landmarks were first defined by Stevens in [1] as "*a discrete representation of the speech stream in terms of a sequence of segments, each of which is described by a set (or bundle) of binary distinctive features*". These landmarks can be determined following the procedure described in [2], in which their detection is mainly based upon the analysis of the energy changes in six frequency bands. In this study, three types of landmarks are considered: b-Lmk, which are related to bursts during articulation; g-Lmk, coinciding with the beginning or ending of vocal fold vibration; and s-Lmk which mark the transitions between vowels and sonorant consonants or vice-versa. Fig. 1 shows the spectrogram and acoustic landmarks of a normal voice during a diadochokinetik (DDK) test.

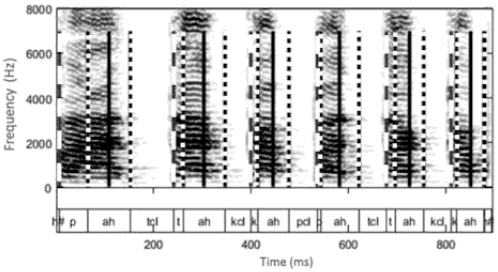

Fig. 1. *Landmarks on a DDK speech task (pa-ta-ka). Dashed lines represent b-lmk, dotted lines are related to g-lmk (pointing the beginning and end of vocal fold vibration). Additionally, black continuous lines mark the Vowel landmark, in the middle of the two g-lmk.*

The acoustic landmarks have already been employed in [3] for PD detection where these are used as a mean to characterize prosody. Other works like [4], [5] do not specifically utilize acoustic landmarks but employ particular segments of speech to characterize and detect parkinsonian speakers.

In this work, three different approaches for the automatic detection of PD are assessed. On each one, a different family of features characterizing speech and classification scheme are considered.

II. METHODS

Overview: The main objective of this study is to automatically detect PD using some articulatory-related features which are introduced in a classification scheme. The families of features can be acoustic cues, probability of a candidate (PoC) or Rasta-PLP coefficients. The classification schemes can be GMM + Logistic regression, GMM-Blend and GMM-UBM-Blend. Some combinations of families of features and classification schemes are performed aiming to obtain the highest accuracies.

* orcid.org/0000-0002-3033-7005

Features: On this study three families of features are used. The first one is integrated by the acoustic cues, defined in [2]. Each landmark type (b-Lmk, g-Lmk and s-Lmk) has several associated specific acoustic cues which consist on some measurements over the speech signal around the acoustic landmark. For b-Lmk, acoustic cues are abruptness (i.e., difference of energy level between two points separated by a certain time window) and silence (i.e., energy level on both sides of the landmark). For g-Lmk, acoustic cues are abruptness and vocalic level (again, energy level on both sides of the landmark). For s-Lmk, the acoustic cues used in this work are abruptness and energy statistics (as mean, maximum, minimum and tilt).

The second family of features is the PoC. As not all the landmarks detected by the algorithm are true landmarks, the acoustic cues of each candidate of being a landmark are introduced in a statistical model trained with the TIMIT database as explained in [2]. After this, it is possible to calculate the probability of a candidate of being a true landmark, obtaining the PoC values.

The third family of features is Rasta-PLP. On this work, these last features are extracted along the whole signal or are calculated only in three overlapped time windows located around each specific landmark, as represented in Fig. 2. In these last cases, the rest of the signal is discarded. Thus, these features are called Lmk-based Rasta-PLP and can be related to b-Lmk, g-Lmk or s-Lmk, depending on the landmark around which these features are extracted.

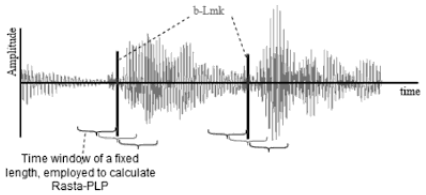

Fig. 2. *Selection of three time windows around b-Lmk to calculate Lmk-based Rasta-PLP.*

Databases: Three databases are used in this study: The first one is the GITA database [6], which includes speech from 50 parkinsonian and 50 control Colombian speakers. A DDK task (/pa-ta-ka/) and two sentences are selected in this work from all the available materials, being the two sentences: Sentence 1: *"Los libros nuevos no caben en la mesa de la oficina"*; and Sentence 2: *"Luisa Rey compra el colchón duro que tanto le gusta".*This database is used to train and test different classification models as it is proposed in the methodology. The second database is the Neurovoz corpus which is employed for validation purposes and contains DDK tasks (/pa-ta-ka/) of 46 and 26 speakers in the parkinsonian and control groups respectively. The third database consists on the first corpus of the Albayzin database [7] which is used to create the UBMs to train the GMM-UBM models.

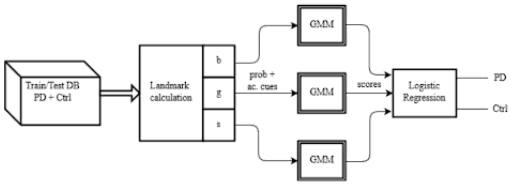

Fig. 3. *First approach for PD detection.*

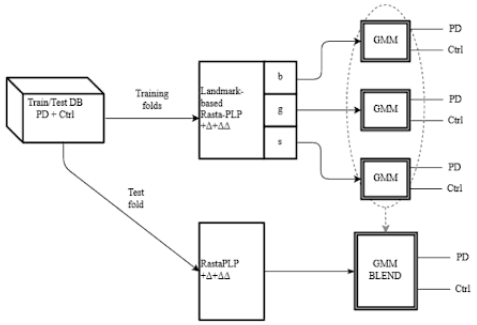

Fig. 4. *Second approach for PD detection.*

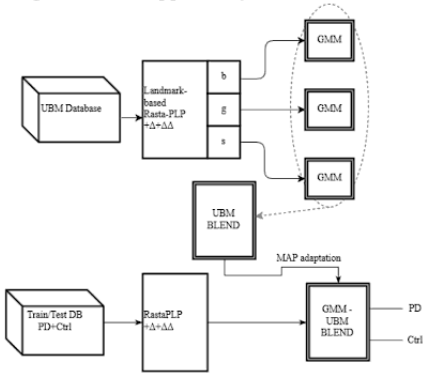

Fig. 5. *Third approach for PD detection.*

Methodology: Firstly, a preliminary analysis of acoustic cues and PoC using DDK utterances from GITA database is performed, to evaluate the statistical behavior of these families of features and their separability on the parkinsonian and control classes. Then, three different approaches are considered depending on the used features and the classification scheme. All the training-testing iterations on each approach follow a k-folds validation scheme with k=7. On the first approach one different GMM classifier is trained and tested for each acoustic landmark type, namely b-Lmk, g-Lmk and s-Lmk, using acoustic cues and PoC joined in a feature vector for each landmark point. Therefore, three global scores are obtained per speaker, one for each type of landmark. These three

global scores are fused following a logistic regression scheme in order to classify the speaker as parkinsonian or control using the equal error rate as threshold. The diagram of this stage is depicted in Fig. 3. In the second approach, features are Lmk-based Rasta-PLP plus derivatives (Δ+ΔΔ) obtained employing windows of 15 ms with 50% overlapping and. On this second approach, three different GMM are trained too, one for each landmark type. Then, the three resulting GMM are blended into a new GMM which is tested with using Rasta-PLP+Δ+ΔΔ features from the testing fold at each iteration of the cross-validation. The GMM blending consists on the creation of a model containing all the Gaussians of the three original models and the weightings of these pondered by a factor. On this study, this factor is always 1/3. Fig. 4 depicts this approach. On the third approach, Lmk-based Rasta-PLP+Δ+ΔΔ extracted from the Albayzin database are employed to obtain three different UBMs, one for each landmark type. Then, these three UBMs are blended into one and this is readjusted into a new GMM using MAP adaptation and Rasta-PLP features from the utterances included the training folds. A diagram of this approach is presented in Fig. 5. This third approach is repeated using Rasta-PLP+Δ+ΔΔ to characterize UBM database (and, therefore, avoiding segmentation and the GMM-UBM-Blend) in order to compare results obtained with and without the use of landmark-based segmentation. The three approaches are achieved using the DDK task from the GITA database. Additionally, the third approach is repeated using the two sentences from this database and the DDK utterances from Neurovoz database. In all three approaches, the number of Gaussians of the GMM models are varied in the range [4, 8, 16, 32, 64, 128] while the number of Rasta-PLP coefficients is 12.

III. Results

Results regarding the preliminary study and the three approaches are included in this section. Only results leading to the highest accuracy are included.

Fig. 6 shows the boxplots of some acoustic cues and PoC associated to the three types of landmarks. For the sake of simplicity, some of the acoustic cues are not referred.

The accuracy, confidence interval (CI), area under the curve (AUC), specificity and sensitivity obtained on the three different approaches are included in tables 1, 2 and 3.

Table 1. Best results on first approach

Lmk	*Accu. (%)*	*CI (%)*	*AUC*	*Spec.*	*Sens.*
b	**80**	**±8**	**0,83**	**0,82**	**0,78**
g	70	±9	0,79	0,70	0,70
s	75	±8	0,81	0,76	0,74
Fusion	77	±8	0,84	0,70	0,84

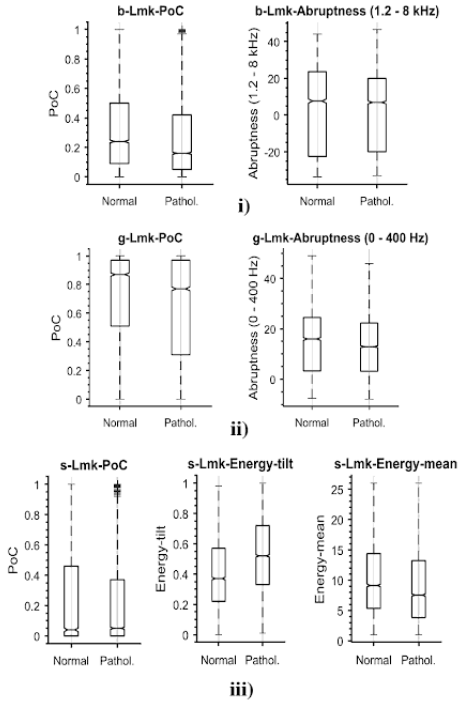

Fig. 6. *Boxplots of PoC and acoustic cues for b (i), g (ii) and s (iii) landmarks. All values, except by PoC are expressed in dB.*

Table 2. Best results on second approach

Lmk	*Accu. (%)*	*CI (%)*	*AUC*	*Spec.*	*Sens.*
none	76	±8	0,8	0,74	0,78
b	74	±9	0,8	0,74	0,74
g	75	±8	0,81	0,74	0,76
s	74	±9	0,82	0,68	0,8
All	**78**	**±8**	**0,85**	**0,78**	**0,78**

Results are referred to DDK task of GITA database in all cases unless otherwise specified. Specifically, other speech tasks and databases are used additionally in the third approach.

Table 3. Best results on third approach

Database	*Speech task*	*Lmk-based & GMM-Blend*	*Accu. (%)*	*CI (%)*	*AUC*	*Specif.*	*Sensit.*
GITA	DDK	Yes	82	±8	0,87	0,82	0,82
		No	75	±8	0,82	0,72	0,78
GITA	Sentence 1	Yes	82	±8	0,88	0,91	0,71
		No	80	±8	0,88	0,90	0,70
GITA	**Sentence 2**	**Yes**	**87**	**±7**	**0,91**	**0,92**	**0,82**
		No	78	±8	0,88	0,92	0,64
Neurovoz	DDK	Yes	82	±9	0,89	0,77	0,85
		No	78	±10	0,84	0,69	0,83

IV. Discussion

Preliminary results, as shown in Fig. 6, reveal that acoustic cues for the three types of landmarks have a different statistical distribution in the two classes, especially in the case of Energy tilt and Energy mean for s-Lmk and PoC for b-Lmk and g-Lmk. Although the used speech tasks in this case (DDK) do not include s-Landmarks (/pa-ta-ka/ only contains b-Lmk and g-Lmk), these are used for the rest of the study as in many occasions, s-Lmk candidates are detected, especially for parkinsonian speakers. This can be caused by the motor perturbation associated to articulation that many PD patients suffer in which some burst or plosive consonants become sonorant. This sign may be a consequence of the reduction of the articulation ranges. That is the reason why s-Lmk and its acoustic cues provide considerable outcomes, as it can be inferred from tables 1 and 2.

Regarding the first approach, acoustic cues + PoC extracted from b-Landmarks provide the best results, with 80% of accuracy. Fusion of scores of the three types of landmarks does not result in better accuracies as it can be inferred from Table 1. Table 2 shows the results for the second approach, where the use of Lmk-based Rasta-PLP+Δ+ΔΔ provides lightly lower accuracies (75% for g-Lmk) than in the case in which all speech frames are used (76%) while the GMM-Blend, considering the models of the three types of landmarks, provides the best results of this second approach (78%). Finally, Table 3 shows the results of the GMM-UBM-Blend approach using different types of speech materials in the train/test databases. The accuracy employing the DDK task is higher in this third approach than in the rest (reaching 82%) while best results are obtained using Sentence 2 (87%) where a relative improvement of 11% is achieved with respect to the non Lmk-based segmentation and non GMM-UBM-Blend scenario. On this approach, the obtained specificity and sensitivity repeating the methodology with the Neurovoz database are comparable to those obtained with GITA database. Therefore, this approach seems to be appropriate to be used in voice pathology detection schemes in which the databases are relatively small and some parts of the speech are more relevant for detection than others. In future works, new studies based on the use of specific segments such as plosive or fricative consonants should consider the GMM-UBM-Blend technique. It has been observed that the landmark detection techniques detect much more candidates than the true number of landmarks present in a sequence and, therefore, more precise techniques such as forced alignment [8] might be employed in the future to detect specific segments.

V. Conclusion

In this work, several approaches for the detection of PD using speech have been analyzed. From all the proposed schemes, the use of GMM-UBM-Blend provides the best results, 87% of accuracy, employing Lmk-based Rasta-PLP to train the UBM model and Rasta-PLP when performing the MAP adaptation. Results evidence that the use of GMM-UBM-Blend techniques with acoustic landmark segmentation in the UBM database provide better results than just GMM-UBM typical use. The employment of the acoustic landmark segmentation for the training of GMM models along with Rasta-PLP coefficients, provides relative improvements up to 11% respect non-segmentation scenarios and must be considered for future works.

Acknowledgements

This work was supported by the grants EEBB-I-17-12092 and BES-2013-062984 within project TEC2012-38630-C04-01, RR01/2011, PRX15/00385, *XV Ayudas Consejo Social-UPM* and *Ayudas EEBB para PDI-UPM*, with special thanks to the Fulbright Foundation.

References

[1] K. N. Stevens, "Toward a model for lexical access based on acoustic landmarks and distinctive features," *J. Acoust. Soc. Am.*, 2002.

[2] C. Park et al., "Automatically determining acoustic landmark sequences using physiological constraints," in *ICASSP*, 2008.

[3] Huici, H. D. et al., "Speech rate estimation in disordered speech based on spectral landmark detection". *Biomedical Signal Processing and Control*, 2016.

[4] J. R. Orozco-Arroyave et al., "Voiced/unvoiced transitions in speech as a potential bio-marker to detect Parkinson's disease," in *INTERSPEECH*, 2015.

[5] Novotný, M. et al., "Automatic evaluation of articulatory disorders in Parkinson's disease." IEEE/ACM Transactions on Audio, Speech and Language Processing (TASLP), 2014.

[6] J. R. Orozco-Arroyave et al., "New Spanish speech corpus database for the analysis of people suffering from Parkinson's disease," Proc. Ninth Int. Conf. Lang. Resour. Eval., 2014.

[7] A. Moreno, D. Poch et al., "Albayzín speech database: Design of the phonetic corpus," Eurospeech 1993. Proc. 3rd Eur. Conf. Speech Commun. Technol., 1993.

[8] Moreno, P. J. et al., "A factor automaton approach for the forced alignment of long speech recordings." ICASSP 2009.