EFFECT OF ACOUSTIC CONDITIONS ON ALGORITHMS TO DETECT PARKINSON'S DISEASE FROM SPEECH

J. C. Vásquez-Correa[1,2,3], *J. Serrà*[1], *J. R. Orozco-Arroyave*[2,3], *J. F. Vargas-Bonilla*[2], *E. Nöth*[3]

[1]Telefónica Research, Barcelona, Spain.
[2] Faculty of Engineering, University of Antioquia UdeA, Calle 70 No. 52-21, Medellín, Colombia.
[3]Pattern Recognition Lab, Friedrich-Alexander-Universität Erlangen-Nürnberg, Germany.

jcamilo.vasquez@udea.edu.co

ABSTRACT

Automatic detection of Parkinson's disease (PD) from speech is a basic step towards computer-aided tools supporting the diagnosis and monitoring of the disease. Although several methods have been proposed, their applicability to real-world situations is still unclear. In particular, the effect of acoustic conditions is not well understood. In this paper, the effects on the accuracy of five different methods to detect PD from speech are evaluated. Among the considered conditions, background noise produces the worst effect, while dynamic compression or some speech codecs can even have a marginal positive impact. We also consider, for the first time in this context, the problem of mismatches, i.e., when train/test acoustic conditions are different, and observe a high negative impact on all considered methods. Overall, this study is a step forward in performing a continuous monitoring of the neurological state of the patients in non-controlled acoustic conditions.

Index Terms— Parkinson's disease, Background noise, Telephone channel, Train/Test mismatch.

1. INTRODUCTION

Parkinson's disease (PD) is a neurological disorder that affects the function of the basal ganglia in the midbrain, producing serious motor and non-motor impairments [1]. Speech disorders are among the most prevalent, and an early sign of further motor impairments [2]. Therefore, computational approaches to detect PD from speech represent a major opportunity to support and improve not only the diagnosis of the disease, but also its monitoring [3].

Different techniques have been proposed for the automatic discrimination between PD patients and healthy controls (HC) from speech. In [4], a set of phonation features computed on sustained vowels is considered, including shimmer, jitter, noise measures, Mel frequency cepstral coefficients (MFCCs), and non-linear dynamics features. In [5], different articulation features are considered, including formant frequencies, vowel articulation index, and vowel space area. An alternative set of articulation features is proposed in [6] to model the difficulties of PD patients to start/stop the vocal fold vibration, analyzing the transitions between voiced and unvoiced segments.

Typically, in existing studies, the acoustic conditions of the recordings are carefully controlled, and algorithms deal with clean signals, low-noise levels, and uncorrupted utterances. The recordings obtained from realistic situations can be of much lower quality, potentially including a combination of noises and transformations that can seriously affect the audio content. This fact questions the reliability of existing algorithms to detect PD in real-world applications. Moreover, comparative evaluations of different algorithms under the same data are also scarce.

In this paper, we analyze the impact of several acoustic conditions on the performance of different methods to classify PD patients vs. HC speakers. The main aim in the near future is to develop a system that can track the neurological state of the patients in different acoustic conditions. The acoustic conditions considered here are: saturation, dynamic compression, additive white Gaussian noise (AWGN), different kinds of environmental background noise, codecs used in communication systems, and real telephone channels. Five different methods are considered: (1) phonation modeling based on the analysis of sustained vowels, (2) voiced and unvoiced modeling [7], (3) voiced/unvoiced transitions modeling [6], (4) openSMILE [8] features, and (5) acoustic modeling using super-vectors (SV) [9]. We also consider, for the first time in this context, the effect of mismatched conditions.

2. STUDIES ON ACOUSTIC CONDITIONS

This section describes the few recent contributions to the analysis of speech of PD patients in different acoustic conditions. In [3], speech recordings of 52 PD patients are transmitted over a simulated mobile telephone network. The authors aim to predict the unified Parkinson's disease rating scale [10] score of the patients by means of modeling recordings of the sustained vowel /A/. In [11], 28 subjects (14 PD and 14 HC)

Table 1. Acoustic Conditions tested in this study. Noise DB1 is obtained from [14] and noise DB2 is obtained from [15].

Condition	Levels/Codecs/Channels
Saturation	Gain $\in \{5, 10, 25, 40\}$ dB
Compression	Ratio $\in \{10{:}5, 20{:}10, 40{:}10, 60{:}20\}$ dB
AWGN	SNR $\in \{15, 10, 5, 0\}$ dB
Street (DB1)	SNR $\in \{10, 6, 0, -6, -10\}$ dB
Street (DB2)	
Cafeteria (DB1)	
Cafeteria (DB2)	
Clinic (DB1)	
Home living (DB2)	
Reverb. room (DB2)	
Car (DB2)	
Codecs	{GSM-FR, G.722, A-law, Silk, Opus}
Real Channels	{Landline, Mobile, Hangouts, Skype}

are classified using speech recordings captured with a low-cost platform and in non-controlled acoustic conditions. A speech enhancement technique is applied to improve the quality of the signals and the classification accuracy. In [12], utterances from 50 patients and 50 HC are compressed by different speech codecs for assessing the impact of such transformation. Regarding the empirical comparison of different methods, in [9] the authors compare the performance of four techniques for PD/HC and PD level classifications, using a database of 88 patients and 88 HC.

3. METHODOLOGY

3.1. Data

We consider recordings from the PC-GITA database (50 PD and 50 HC) [13]. The data-set is balanced in age and gender, and it was recorded in a sound-proof booth, with a dynamic omnidirectional microphone and a professional audio card with a sampling frequency of 44.1 kHz and 16-bit resolution. Among other tasks [13], the participants were asked to pronounce the five Spanish vowels in a sustained manner, to read a text with 36 words (RT), and to produce a monologue about their daily activities (M). All patients were recorded in ON state, i.e., no more than three hours after their morning medication, and were labeled by a neurologist expert.

3.2. Acoustic Conditions

For this study, the PC-GITA recordings were corrupted in several ways: saturation, dynamic compression, AWGN, and different kinds of environmental noise. Additionally, telephony codecs such as the full-rate GSM (GSM-FR), the G.722, the A-law, Silk, and Opus were considered. Finally, the original signals were transmitted and re-captured using 4 real communication systems: Skype, Google Hangouts, landline phone, and mobile phone. For saturation, compression, and noise, different levels of gain, compression ratios, and noise levels, were considered, respectively (Table 1).

3.3. Methods

Phonation Model on Sustained Vowels – The phonation model (PM) is typically based on features related to perturbation measures of the F_0 and amplitude. The feature vector includes the F_0 derivative, log-energy, jitter, and shimmer. Additionally, the first two formant frequencies are considered due to their capability to model several positions of the tongue [16]. Four functionals are calculated upon each descriptor (mean, standard deviation, skewness, and kurtosis). The amplitude perturbation quotient and the pitch perturbation quotient are also computed over the contour of the sustained voiced region, yielding a total of 26 features per vowel.

Voiced/Unvoiced Modeling – The separate analysis of voiced and unvoiced segments for PD detection was introduced in [7]. Voiced modeling (VM) includes a total of 18 descriptors: F_0 derivative, log-energy, jitter, shimmer, the first and second formant frequencies, and 12 MFCCs. Unvoiced modeling (UM) includes 37 descriptors: 25 Bark band energies (BBEs) and 12 MFCCs. The same functionals as with PM are calculated, forming a 72-dimensional feature vector for VM and a 148-dimensional feature vector for UM.

Voiced/Unvoiced Transitions Modeling – The transitions from voiced to unvoiced (offset) and from unvoiced to voiced (onset) were introduced in [6] to model the difficulties observed in PD patients to stop/start vocal fold vibration. The border between the voiced and unvoiced segments is detected, and frames are taken to each side of the border. Both onset and offset are modeled with 25 BBEs and 12 MFCCs. Four functionals are computed, forming a 148-dimensional feature vector both for onset (OnM) and offset (OffM) modeling.

OpenSMILE – This toolkit was considered as the baseline of the INTERSPEECH 2015 computational paralinguistic challenge [17], and it can also be used to extract features to discriminate between PD and HC. The feature vector consists of 6373 static measures formed by several descriptors and functionals calculated using the OpenSmile toolkit v2.1 (OS) [8].

Acoustic Modeling using Super-vectors – This approach was introduced in [9] to detect PD. The method consists of first extracting 13 MFCCs along with their first and second derivatives. Then, a universal background model is trained with the information of all the population from the training set using the expectation maximization algorithm, and specific Gaussian mixture models (GMMs) are adapted for each speaker using the maximum a posteriori rule. Experiments with the number of Gaussians ranging from 4 to 128 were performed, and the best results were obtained with 16 Gaussians. After adaptation, the mean vectors of the GMM are merged together to form a (39×16)-dimensional feature super-vector per speaker.

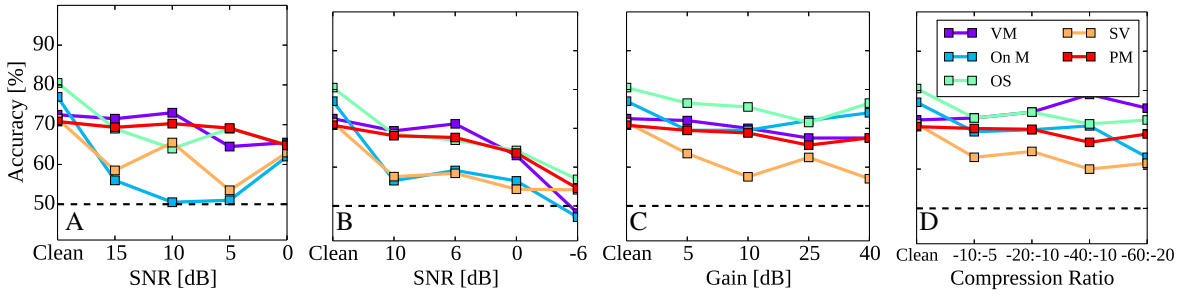

Fig. 1. Accuracies (%) for matched analysis: (A) AWGN, (B) average of all environmental noises, (C) saturation, and (D) dynamic compression. Average accuracies are shown. The complete results with exact numbers can be found online.

Table 2. Accuracies (%) with clean signals for different speech tasks and feature sets: PM of vowel A, VM, OnM, OS, and SV. The results for the missing vowels, OffM, and UM can be found online.

PM	**VM**		**OnM**		**OS**		**SV**	
A	RT	M	RT	M	RT	M	RT	M
71	74	80	82	72	80	81	72	71

3.4. Classification & Evaluation

Following the state-of-the-art, we use a support vector machine (SVM) [4, 6]. For the PM, VM, OnM, and OffM, a Gaussian kernel is used. For the case of SV and OS, a linear kernel with LASSO regularization is considered due to the fact that both are formed with high-dimensional feature vectors. The models are tested following a leave-one-speaker-out cross-validation strategy (LOSOCV), and the meta-parameters C and γ are optimized in a grid search, with selection criterion based on the accuracy obtained in the train set ($C \in \{10^{-5}, 10^{-4}, \dots 10^{4}\}$ and $\gamma \in \{10^{-6}, 10^{-5}, \dots 10^{2}\}$).

4. RESULTS

As mentioned, three scenarios are considered: the clean signals (Sec. 4.1), the matched conditions (Sec. 4.2), and the mismatched conditions (Sec. 4.3). Due to space and clarity constraints, we here report only the most relevant results. The complete results obtained with all the feature sets in all acoustic conditions can be found online[1].

4.1. Clean Signals

A summary of the results for clean signals is reported in Table 2. The highest accuracy is obtained with OnM (82%) for the read text, followed by the results obtained with OS (80–81%). Note that the results obtained with OnM and VM differ from those reported in [6, 7]. This fact is explained for several reasons. Firstly, a re-implementation of both algorithms in a different programming language was done, with the consequent possible changes in basic functions outputs. Secondly, the results reported in the present study correspond to LOSOCV, while the results reported in [6, 7] correspond to a 10-fold CV forcing age and gender balance. Finally, the accuracy reported in the previous papers was computed by optimizing the parameters of the classifier with the test set, whereas here are optimized on the training set. Overall, we believe that the methodology settings in the current work represent a higher standard and a more realistic evaluation of the algorithms according to the available data.

Additional experiments were performed considering the fusion of speech tasks, with the aim of improve the results. For instance, we concatenated the features from the five vowels prior to classification, obtaining an accuracy of 79% (compared to the 71% obtained only with /A/). We also concatenated the features from RT and M for VM, OnM, OS, and SV, obtaining accuracies of 80% (VM), 77% (OnM), 83% (OS), and 83% (SV). In general, combining utterances is beneficial, specially for the algorithms with initially lower accuracies. However, such strategy results in modest improvements for the best performing methods.

4.2. Matched Conditions

Fig. 1A shows the results when we degrade the quality of the signals with AWGN. The performance of PM is not seriously affected, which makes this approach the most robust against AWGN. OnM is the most affected method, with a performance reduction of up to 30%, which might be due to errors that appear in the detection of the transition between the U/V segments. Fig. 1B contains the average results across all background noises. The most affected algorithms are OnM and SV. PM and OS are the less affected. In general, it seems like the V/U detection is being corrupted by the different kinds of noise, producing the high reduction in the accuracy of OnM. The performance for individual environmental noises is reported online. In summary, cafeteria and reverberated room noises are the most critical for almost all algorithms, while street and car noises have the lowest impact. Fig. 1C illustrates the results for the distortion produced by saturating the recording devices. We observe that the effect is less critical than the one produced by the background

[1] https://www5.cs.fau.de/en/our-team/orozco-rafael/projects/IS-2016-Camilo

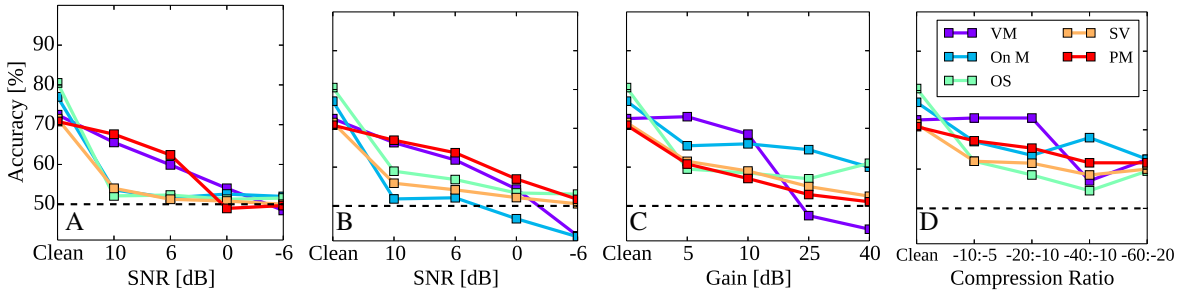

Fig. 2. Accuracies (%) for mismatched analysis: (A) environmental noises for Mismatch 1, (B) environmental noises for Mismatch 2, (C) saturation for Mismatch 1, and (D) dynamic compression for Mismatch 1. Average accuracies are shown. The complete results with exact numbers can be found online.

Table 3. Accuracies (%) for telephone codecs and channels.

	PM	**VM**		**OnM**		**OS**		**SV**	
	A	RT	M	RT	M	RT	M	RT	M
Clean	71	74	80	82	72	80	81	72	71
Codecs									
Opus	70	79	69	86	61	87	81	65	69
Silk	71	75	73	75	73	75	67	61	57
A-law	66	78	76	73	64	62	64	62	62
G.722	69	82	72	87	76	79	63	63	61
GSM-FR	72	70	82	68	70	69	76	60	68
Channels									
Hangouts	71	76	64	79	67	85	77	64	58
Skype	62	66	73	61	55	75	79	63	71
Landline	68	70	75	63	66	66	78	77	63
Mobile	65	66	76	50	65	64	76	58	71

noise (Fig. 1B). Fig. 1D shows the results for the dynamic compression. We see that specific compression ratios can improve the accuracy for certain algorithms, while reducing the performance of others. For the saturation and compression, the V/U detection is not highly affected, implying a low impact on the overall performance of VM and OnM.

Table 3 contains the results when the signals are compressed by telephony codecs and re-captured from real telephone channels. Interestingly, we find that some codecs improve the results, specially Opus and G.722 in OnM and OS. GSM-FR can also improve the results for PM and VM. Such improvements can be explained by the different signal processing stages involved in the coding schemes, which may perform noise filtering, dynamic compression, pitch re-synthesis, and intensify certain zones of the spectrum.

For the case of real telephone channels, in general, there is not much effect on the accuracy of the algorithms (even some channels such as Hangouts can improve accuracy). As mentioned, some of the processes involved in the transmission over such channels may be beneficial for the detection. The highest impact is observed for the mobile channel, and can be explained by its transmission rate (the transmission rate of Google Hangouts ranges from 6 to 52 kbps,the bit rate of Skype ranges from 6 to 40 kbps, the bit rate of landline is 64 kbps and, conversely, the bit-rate for mobile channels is around 12.2 kbps).

4.3. Mismatched Conditions

Mismatches occur when clean data is used for training and noisy recordings are considered for testing (Mismatch 1, Fig. 2A) or viceversa (Mismatch 2, Fig. 2B). In general, we find the impact is clearly higher than in the matched conditions (compare with Fig. 1B). For Mismatch 1, the results are close to random for OS, SV, and OnM. PM and VM are the less affected, but still present a considerable impact. Fig. 2C contains the results for Mismatched 1 with distortion produced by saturation. We now see the effect is apparent, as compared to the matched condition (Fig. 1C). The same applies for dynamic compression (Figs. 2D and 1D).

5. CONCLUSION

This study evaluates the effect of several acoustic conditions on different methods to detect PD from speech. According to the results, background noise has the strongest influence in the classification task. Thus for a continuous monitoring of the neurological condition, it has to be considered with special attention. The effect of saturation only appears in the mismatched condition. Dynamic compression and the codecs can improve the results, the latter benefiting from several signal processing procedures to reduce noise and intensify spectrum zones. The impact of telephone channels is not critical, except for the mobile channel, where the low bit-rate causes a high reduction in the accuracy. It seems like it is better to use VoIP systems than mobile recordings. Mismatched conditions severely impact the performance of the algorithms. Although we achieved some initial promising results with data augmentation techniques (not reported), we believe the topic deserves further and more systematic investigation.

6. ACKNOWLEDGEMENTS

To the COLCIENCIAS project # 111556933858, to CODI from University of Antioquia, and Telefónica Research.

7. REFERENCES

[1] O. Hornykiewicz, "Biochemical aspects of Parkinson's disease," *Neurology*, vol. 51, no. 2, pp. S2–S9, 1998.

[2] J. A. Logemann, H. B. Fisher, B. Boshes, and E. R. Blonsky, "Frequency and cooccurrence of vocal tract dysfunctions in the speech of a large sample of Parkinson patients," *Journal of Speech and Hearing Disorders*, vol. 43, no. 1, pp. 47–57, 1978.

[3] A. Tsanas, M. Little, P. E. McSharry, and L. O. Ramig, "Using the cellular mobile telephone network to remotely monitor Parkinsons disease symptom severity," *IEEE Transactions on Biomedical Engineering*, 2012.

[4] A. Tsanas, M. Little, P. E. McSharry, J. Spielman, and L. O. Ramig, "Novel speech signal processing algorithms for high-accuracy classification of Parkinson's disease," *IEEE Transactions on Biomedical Engineering*, vol. 59, no. 5, pp. 1264–1271, 2012.

[5] J. Rusz, R. Cmejla, T. Tykalova, H. Ruzickova, J. Klempir, V. Majerova, J. Picmausova, J. Roth, and E. Ruzicka, "Imprecise vowel articulation as a potential early marker of Parkinson's disease: Effect of speaking task," *The Journal of the Acoustical Society of America*, vol. 134, no. 3, pp. 2171–2181, 2013.

[6] J. R. Orozco-Arroyave, F. Hönig, J. D. Arias-Londono, J. F. Vargas-Bonilla, S. Skodda, J. Rusz, and E. Nöth, "Voiced/unvoiced transitions in speech as a potential bio-marker to detect Parkinsons disease," in *Anual Conference of the Speech and Communication Association (INTERSPEECH)*, 2015, pp. 95–99.

[7] J. R. Orozco-Arroyave, F. Hönig, J. D. Arias-Londoño, S. Skodda, J. Rusz, and E. Nöth, "Automatic detection of Parkinson's disease from words uttered in three different languages.," in *Anual Conference of the Speech and Communication Association (INTERSPEECH)*, 2014, pp. 1573–1577.

[8] F. Eyben and B. Schuller, "openSMILE:): the Munich open-source large-scale multimedia feature extractor," *ACM SIGMultimedia Records*, vol. 6, no. 4, pp. 4–13, 2015.

[9] T. Bocklet, S. Steidl, E. Nöth, and S. Skodda, "Automatic evaluation of Parkinson's speech-acoustic, prosodic and voice related cues," in *Anual Conference of the Speech and Communication Association (INTERSPEECH)*, 2013, pp. 1149–1153.

[10] C. G. Goetz and et al., "Movement Disorder Society-sponsored revision of the Unified Parkinson's Disease Rating Scale (MDS-UPDRS): Scale presentation and clinimetric testing results," *Movement Disorders*, vol. 23, no. 15, pp. 2129–2170, 2008.

[11] J. C. Vásquez-Correa, T. Arias-Vergara, J. R. Orozco-Arroyave, J. F. Vargas-Bonilla, J. D. Arias-Londoño, and E. Nöth, "Automatic detection of Parkinson's disease from continuous speech recorded in non-controlled noise conditions," in *Anual Conference of the Speech and Communication Association (INTERSPEECH)*, 2015, pp. 105–109.

[12] J. R. Orozco-Arroyave, N. García, J. F. Vargas-Bonilla, and E. Nöth, "Automatic detection of Parkinsons disease from compressed speech recordings," *Lecture Notes in Computer Science*, vol. 9302, pp. 88–95, 2015.

[13] J. R. Orozco-Arroyave, J. D. Arias-Londoño, J. F. Vargas-Bonilla, M. C. Gonzalez-Rátiva, and E. Nöth, "New spanish speech corpus database for the analysis of people suffering from Parkinson's disease.," in *Language Resources and Evaluation Conference, (LREC)*, 2014, pp. 342–347.

[14] J. C. Vásquez-Correa, N. García, J. R. Orozco-Arroyave, J. D. Arias-Londoño, J. F. Vargas-Bonilla, and E. Nöth, "Emotion recognition from speech under environmental noise conditions using wavelet decomposition," in *IEEE International Carnahan Conference on Security Technology (ICCST)*, 2015, pp. 247–252.

[15] D. B. Dean, S. Sridharan, R. J. Vogt, and M. W. Mason, "The QUT-NOISE-TIMIT corpus for the evaluation of voice activity detection algorithms," in *Anual Conference of the Speech and Communication Association (INTERSPEECH)*, 2010, pp. 26–30.

[16] S. Skodda, W. Visser, and U. Schlegel, "Vowel articulation in Parkinson's disease," *Journal of Voice*, vol. 25, no. 4, pp. 467–472, 2011.

[17] B. Schuller, S. Steidl, A. Batliner, S. Hantke, F. Hönig, J. R. Orozco-Arroyave, E. Nöth, Y. Zhang, and F. Weninger, "The INTERSEECH 2015 computational paralinguistics challenge: Nativeness, Parkinsons & eating condition," in *Anual Conference of the Speech and Communication Association (INTERSPEECH)*, 2015, pp. 478–482.