Introducing non-linear analysis into sustained speech characterization to improve sleep apnea detection

Jose Luis Blanco[1], Luis A. Hernández[1], Rubén Fernández[1], Daniel Ramos[2]

[1] Signal Processing Applications Group, Universidad Politécnica de Madrid, ETSI Telecomunicacion, 28040 Madrid, Spain
[2] Biometric Recognition Group (ATVS), Universidad Autónoma de Madrid, Escuela Politecnica Superior, 28049 Madrid, Spain

{jlblanco, luis, ruben}@gaps.ssr.upm.es, daniel.ramos@uam.es

Abstract. We present a novel approach for detecting severe obstructive sleep apnea (OSA) cases by introducing non-linear analysis into sustained speech characterization. The proposed scheme was designed for providing additional information into our baseline system, built on top of state-of-the-art cepstral domain modeling techniques, aiming to improve accuracy rates. This new information is lightly correlated with our previous MFCC modeling of sustained speech and uncorrelated with the information in our continuous speech modeling scheme. Tests have been performed to evaluate the improvement for our detection task, based on sustained speech as well as combined with a continuous speech classifier, resulting in a 10% relative reduction in classification for the first and a 33% relative reduction for the fused scheme. Results encourage us to consider the existence of non-linear effects on OSA patients' voices, and to think about tools which could be used to improve short-time analysis.

Keywords: obstructive spleep apnea (OSA), continuous speech, sustained speech, gaussian mixture models (GMMs), classification and regression tree (CART).

1 Introduction

The non-linear analysis of speech signals has recently gained a remarkable interest from the scientific community, particularly in the last decade. Many different applications have been suggested based on this information traditionally neglected from speech and speaker recognition tasks. However, researchers have pointed out that non-linear processes are involved in speech production, and that new features are required to parameterize them [1]. This is particularly relevant when considering the detection of abnormal patterns within individual voices, and even more when those patterns are meant to be caused by physiological evidences found on groups of speakers suffering from a certain condition.

Obstructive sleep apnea (OSA) is one of those conditions, affecting an estimated 2-4% of male population between the ages of 30 and 60 years [2]. It is characterized by

recurring episodes of sleep-related collapse of the upper airway at the level of the pharynx and is usually associated with loud snoring and increased daytime sleepiness. OSA is a serious threat to an individual's health if no treated, as it known to be a risk factor for hypertension and, possibly, cardiovascular diseases [3]. Actually, it has been related to traffic accidents caused by somnolent drivers [2-4], and might lead to a poor quality of life and impaired work performance.

Sleep apnea can be diagnosed on the basis of a characteristic history (snoring, daytime sleepiness) and physical examination (increased neck circumference), but a full overnight sleep study is needed to confirm the diagnosis, involving the recording of neuroelectrophisiological and cardiorespiratory variables (ECG). Excellent performance rates are obtained by this method (ca. 90% [5]); however, this test is quite expensive and time-consuming, which cause most patients to suffer a waiting list of several years before it is done. These considerable delays have motivated the appearance of early diagnosis methods which are meant to reduce them, and to determine patients' priority of need and proper place for the polysomnography test. Clinicians aim to bridge this gap by using non-invasive tests providing useful prior information, additionally to patients' clinical story. Speech analysis appears to have a good opportunity to characterize the alterations/abnormalities of patients' vocal tract, and to be used prior to the polysomnography test aiming clinicians diagnosis.

Few evidences on the effects of OSA on patients' voices have been reported. Most valuable information can be found in a 1989 work from Fox *et al.* [6] in which the results from evaluations of skilled judges on a perceptual study were presented. These evaluations have pointed out several differences which can be perceived when comparing voices from apnea patients to those from a control group (also referred as 'healthy' subjects), and have motivated further research on sleep apnea patients voices. On their 1989 work, *abnormal resonances* (hyponasality and hypernasality), specific *articulatory features* (due to a probable velopharyngeal dysfunction) and *phonation anomalies* were found in OSA patients voices. Moreover, Robb's analysis on the vocal tract resonances of apnea patients [7], stressed the differences in the formant values and bandwidths, particularly for F1 and F2. Continuing with the spectral analysis of speech, Fiz *et al.* [8] considered a different set of measures which are nearer to the standards for non-linear systems' characterization, such as the number of harmonics or their mean/maximum frequencies.

In this contribution we intend to characterize the deterministic and stochastic dynamics of speech, aiming to improve our automatic speech detection system for severe obstructive sleep apnea cases detection. In section 2 we describe our database (subsection 2.1), as long as our baseline system based on the characterization, in the cepstral domain, of connected (subsection 2.2) and sustained (subsection 2.3) speech. Revisions on the new features we have selected to characterize sustained speech non-linear dynamics are presented in section 3, including a discussion on the alternatives for the combination of these features with our baseline system. In section 4 we describe several experiments we have carried out to test the improvements on the classification accuracy rates. Finally, in Section 5, some conclusions are provided, as long as a number of open issues regarding the future prospects on the analysis of connected speech dynamics.

2 Automatic detection of apnea based on speech

From a broad perspective the detection of pathological voices can be described as a standard classification problem in which a set of descriptive features are to be selected, while chasing for a set of distinctive patterns which allow us to discriminate among a set of classes. For the detection of severe apnea cases this set is restricted to a unique partition of the range into two different classes, namely control –healthy- and OSA, based on the so called Apnea-Hypoapnea Index (AHI). Conventionally an AHI value below 10 belongs to a healthy subject, while values higher than 30 indicate a really severe case which should go into medical treatment. Additionally, in this problem the set of characteristics describing the acoustic differences between the control group and the one formed by OSA patients is still unclear. However, since the literature enumerates a number of acoustic differences, mel-frequency cepstral coefficients (MFCCs) parameterization was chosen as they relate to the spectral envelope of signals and therefore to the articulation of speech and resonances in the vocal tract. Actually they have been used successfully in a broad number of situations, from speech to speaker recognition and pathological voices detection [9].

Based on this parameterization, Gaussian Mixture Models (GMMs) have proved an enormous potential for modeling the acoustic space of human speech, for both speech and speaker recognition. Several strategies can be followed to estimate a good model depending on the characteristics of the problem and the amount of data available. In our particular situation, we have chosen to begin by training a suitable universal background model (UBM) from a broader database, which is further adapted into two classes (i.e. control and OSA) by means of a *maximium a posteriori* (MAP) adaptation scheme. Other adaptation techniques have also been tested, however, the size of our database seems to be fairly enough for the convergence of the EM algorithm into a suitable class-model. Figure 1 summarizes this procedure, from which we finally have a set of mixture models corresponding to the control and apnea groups respectively.

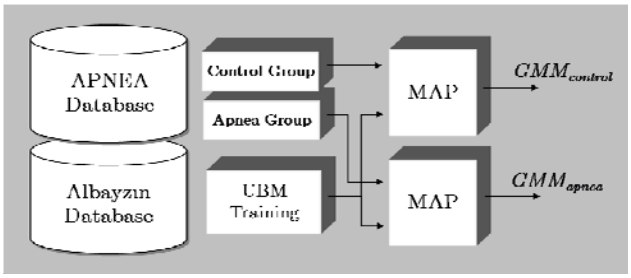

Fig. 1. Apnea and control mixtures training scheme based on MAP adaptation from a UBM model trained on the Albayzin database (extracted from [13]).

Based on these ideas we have designed a set of classifiers which are currently the state-of-the-art technology for sleep apnea detection, both for connected and sustained speech. For all of them both UBM training and GMM adaptation were developed with the BECARS open sour tool [10]. Though formally very similar, intrinsic differences have been found in the development of both classifiers [14], mainly because of the

differences in the nature of these signals, but also because of the limitations of the available database and the influence of the training procedures.

2.1 Speech Corpus

Keeping in mind the results from the perceptual study carried out by Fox and colleagues [6], a set of four phrases was designed to exhibit similar melodic structure and to include instances of the following specific phonetic contexts:

- In relation to *resonance anomalies*, sentences were designed to allow measuring differential voice features for each speaker (e.g. to compare the degree of vowel nasalization).
- Regarding *phonation anomalies*, we included continuous voiced sounds to measure irregular phonation patterns related to muscular fatigue in apnea patients.
- To look at *articulatory anomalies* we collected voiced sounds affected by preceding phonemes that have their primary locus of articulation near the back of the oral cavity (e.g. velar phonemes such as the Spanish velar approximant "g").

A group of 40 healthy male speakers and 40 male patients suffering from OSA in a severe degree were asked to read them three times each, with a specific rhythmic structure under the supervision of an expert. Additionally, sustained vowel /a/ instances were recorded and included in our database. Further information on the design of the database can be found in [11].

2.2 Connected speech classifier

Regarding the amount of data needed to estimate a suitable GMM, we decided to train a UBM from phonetically balanced utterances in the Albayzin database [12], and use MAP adaptation to derive the specific GMMs for the different classes to be trained. Only the means were adapted, as is classically done for speaker verification.

In the end, a GMM-based classification system was trained and tested, based on the 12MFCC, plus Energy, velocity and acceleration coefficients feature vectors and the adapted models. The resulting system provided an overall 76.9% accuracy rate in the detection of severe apnea cases on a leave-one-out cross-validation scheme with 256 mixture components. Further information on this classifier can be found in [13].

2.3 Sustained speech classifier

On an attempt to improve the results from the previous classifier, and noticing that sustained speech analysis is actually the standard for pathological voices detection, we reproduced the same idea for the sustained vowel we had collected in our database. However, we lacked of a database which could fit our needs in the same way as the Albayzin did for or connected speech classifier. This fact motivated a number of questions to decide which was the best way to train and adapt our models, and raised a deep discussion on the balance between models' complexity and the dependency of the final model's convergence with the trained UBM. Interested readers would refer to a recently published work [14] for further information on the design of this classifier.

A set of two GMMs with 16 gaussian components was adapted from a UBM trained from the sustained vowels in Childers database [15]. The feature vector included the same 12 MFCCs, energy, velocity and acceleration parameters used in

the connected speech classifier. A poor 39.4% EER classification rate was achieved, which is actually really far from the result for the connected speech classifier.

Nevertheless, the information coming from sustained speech was proved to be uncorrelated with that coming from connected speech. This seems quite sensible as the kind of the information being modeled is quite different on each of these classifiers. Still, results from this second classifier can be improved by taking into account the differences on the dynamics of speech from apnea patients compared to healthy people.

3 Measures on sustained speech dynamics

Though most efforts in speech processing have been put into the analysis of speech signals as a response to a linear system, the modeling of these signals is currently being enhanced by introducing additional measures derived from the fact that speech production is actually a dynamic process. Throughout the literature many different measures have been suggested which can be used in our classification problem. Those algorithms can be broadly classified into three groups attending to the characteristics of speech signals on which they rely. Those are: (1) the deviations of the vocal tract from an ideal resonator, (2) the recurrent structure of the signals and the self-similarity property, (3) the existence of noisy components within speech signals.

The first group includes a subset of features which are based on the cycle-to-cycle variations of fundamental frequency and waveform amplitude, namely, *jitter* and *shimmer*. However, both magnitudes can be measured according to different *criteria* resulting in a set of jitter and shimmer measurements. In the present contribution we have chosen to include a subset of those (see Table 1) which have been successfully used for speaker verification [16] in order to test their discriminative power.

Table 1. Jitter and shimmer measurements considered in this work (group 1) on sustained vowels recordings (T_i stands for the estimated glottal closure instants –estimated using Matlab's VOICEBOX toolbox [17], and A_i for the extracted peak-to-peak amplitude on each cycle).

Jitter Measurements		**Shimmer Measurements**	
Absolute jitter	$\frac{1}{N-1}\sum_{i=1}^{N-1}\lvert T_i - T_{i-1}\rvert$	Shimmer	$\frac{1}{N-1}\sum_{i=1}^{N-1} 20\log\lvert A_{i+1} - A_i\rvert$
Relative jitter	$\left(\frac{1}{N-1}\sum_{i=1}^{N-1}\lvert T_i - T_{i-1}\rvert\right) \Big/ \left(\frac{1}{N}\sum_{i=1}^{N} T_i\right)$	Relative shimmer	$\left(\frac{1}{N-1}\sum_{i=1}^{N-1}\lvert A_i - A_{i-1}\rvert\right) \Big/ \left(\frac{1}{N}\sum_{i=1}^{N} A_i\right)$
RAP jitter	$\frac{\frac{1}{N-1}\sum_{i=1}^{N-1}\lvert T_i - (T_{i-1}+T_i+T_{i+1})/3\rvert}{\frac{1}{N}\sum_{i=1}^{N} T_i}$ *3-point period perturbation quotient based 2 closest neighbours averaging.*	APQ3 shimmer	*3-point period perturbation quotient, based on the 2 closest neighbours averaging.*
PPQ5 jitter	*5-point period perturbation quotient requires averaging in the period and the four closest neighbours.*	APQ5 shimmer	*5-point period perturbation quotient requires averaging in the period and the four closest neighbours.*
		APQ11 shimmer	*11-point period perturbation quotient requires averaging in the period and the four closest neighbours.*

In this same group, a novel measure for pitch period uncertainty estimation was included: the *pitch period entropy* (PPE) [18]. PPE quantifies the inefficiency in speaker's voice frequency control in terms of the unpredictability of the fundamental frequency evolution curve while uttering a sustained sound.

Regarding the second group, a whole set of measures have been proposed which are based on ideas from the dynamic systems theory. All of them are somehow related to the recurrence and self-similarity properties assumed for the sustained production of speech sounds. For this contribution we have chosen to use two of the most common and well-known measurements [18], briefly described on Table 2.

Table 2. Brief description of the RPDE and DFA measures ($x(n)$ is the speech sample, (a,b) result from a first-order approximation to the windowed series, N_{max} is the maximum recurrence time in the series attractor, and $R(i)$ is the normalized histogram of the recurrence time).

Measure		**Definition**	**Equation**
RPDE	*Recurrence Period Density Entropy*	Extends the conventional concept of periodicity and substitutes it by the idea of recurrence, and represents the uncertainty in the estimate of the pitch period.	$RPDE = \frac{-\sum_{i=1}^{N_{\max}} R(i)\ln R(i)}{\ln N_{\max}}$
DFA	*Detrended Fluctuation Analysis*	Calculates the scaling exponent in non-stationary time series. A least-square straight-line approximation is carried out on each and at every time scale	$DFA = \sqrt{\frac{1}{N}\sum_{n=1}^{N}\left[y(n)-(an+b)\right]^2}$ $y(n) = \sum_{m=1}^{n} x(m)$

Finally, the third group includes different measures which are all based on the idea of estimating the fraction of noise in the recorded speech signals. This fraction is usually presented by means of the estimated *signal-to-noise* ratio (SNR), or its converse the *noise-to-signal* ratio (NSR), attending to a certain definition of energy (e.g. the squared energy operator, SEO, or the Teager-Kaiser energy operator, TKEO). In this contribution we have selected two measures [18] based on noise level estimation which decompose speech signals attending to:

- an hypothetically invariant excitation: the *vocal fold excitation ratio*, VFER
- a decomposition of the signal into a set of AM-FM contributors –*intrinsic mode functions*: the *empirical mode decomposition excitation ratios*, EMD-ER.

All these measures were estimated on each record of a sustained vowel in the apnea database, properly segmented in advance to guarantee their successful estimation and avoid abnormal effects in the speakers' utterances.

4 Experiments and results

In order to achieve a final decision on whether a certain speaker is suffering or not from OSA through the automatic processing of his utterances, and aiming to improve the results from the classifiers described in Section 2, feature sets are to be built by

the combination of the measures presented in Section 3 and the averaged scores assigned to cepstral feature vectors by the log-likelihood ratios using the apnea and control mixture models (i.e. GMMs). Figure 2 depicts a block diagram summarizing the overall scheme of the designed system. The resulting parallel scheme suggests an incremental improvement of the classifiers, which fits our initial motivation for improving our baseline system.

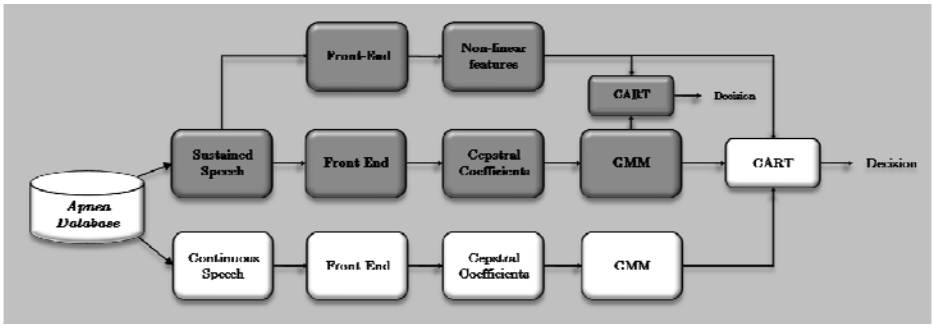

Fig. 2. Combined scheme for the two systems developed to the automatic detection of severe obstructive sleep apnea cases (based only on sustained speech – dark grey-, based on the combination of connected and sustained speech -white)

Two main branches can be identified on Figure 2, the upper one corresponding to the sustained speech processing and classification; and the lower one summarizing the testing procedure for our continuous speech classifier (see subsection 2.2), which is finally combined with the upper branch aiming to improve classification results.

The fusion of the different branches in order to reach a decision on the presence or absence of speech patterns which could be related to OSA, was carried out by means of a standard classification and regression tree (CART). The regression trees estimated for the fusion on the purely sustained speech classifier and the combined connected and speech classification were trained and tested according to a standard leave-one-out cross-validation scheme, just as the prior GMM classifiers. The optimization of the regression tree was defined as a conventional misclassification rate minimization problem in order to be implemented. Additionally, prior to the CART training, tests based on minimum-redundancy and maximum-relevance feature selection tests were developed to identify features with negligible interclass variability. The result is quite similar to a posterior pruning of the lower branches, though this solution reduces the number of features beforehand, and therefore simplifies and accelerates the estimation of the regression trees. The reduced feature set includes: VFER and IMF –in different implementations- RPDE, absolute shimmer and MFCCs (static coefficients, velocity and acceleration fusion at score level).

Table 3. Classification accuracy for each of the classifiers.

	Classifier	Sensitivity (%)	Specificity (%)	Positive Predictive Value	Negative Predictive Value	Accuracy (%)
1	Sustained	66.0	63.0	66.0	63.0	64.6%
2	Sust. & Connected	86.7	90.6	90.9	86.2	88.5%

Finally, classification rates were obtained for each configuration (Table 3), achieving a 35.4% error rate for control/OSA classification based on the information modeled by sustained speech analysis (upper branch), which happens to be a 10% relative reduction from the previously reported rate 39.4% [14] (p=0.075 binom. test). On the other hand, the combined scheme including scores from previous GMM-based classifiers and the complexity measures introduced in Section 3 had an estimated misclassification rate of 11.5%, over a 33% relative reduction from our previous best-performing system for which the EER was estimated to reach a 17.3% [14].

5 Conclusions and future work

The incremental methodology presented in this paper to improve our baseline system by introducing complementary information for severe OSA cases detection has achieved excellent results. The inclusion of non-linear measures describing speech dynamics in the production of sustained vowels has enhanced our characterization of the acoustic space and improved classification rates when only sustained sounds are analyzed. Though still far from the rates achieved by modeling connected speech, this result encourages us to explore other possible parameterizations suitable to describe the complexity of coarticulated sounds. However, the particular pathogenesis of OSA regrets following such approach for a future full-AHI-range OSA detection system, as abnormal patterns caused by vocal folds irregular vibration can only be expected in a severe stage of the syndrome.

Moreover, the combination of three branches of measures and information (i.e. linear connected and sustained, and non-linear sustained speech) to make a single decision has produced a significant improvement in the classification rate for the control/severe-OSA problem. This reinforces the fact that information coming from sustained and connected speech is poorly correlated

We are quite enthusiastic on these results, though some improvement can be expected by introducing new sets of features. The short-time changes in the non-linear dynamics of connected speech still require further analysis in order to be included into our features sets. Following the proposed scheme, those could be introduced into our classification system in the same way as we just did for sustained vowels.

Acknowledgments. The activities described in this paper were funded by the Spanish Ministry of Science and Innovation as part of the TEC2009-14719-C02-02 (PriorSpeech) project. The corresponding author also acknowledges the support from Universidad Politécnica de Madrid full-time PhD scholarship program. Finally, authors will like thank Athanasios Tsanas, Max Little and Professor J. I. Godino-Llorente, for their opinions and recommendations.

References

1. Arias-Londoño, J.D., Godino-Llorente, J.I., Sáenz-Lechón, N., Osma-Ruiz, V. and Castellanos-Domínguez, G.: Automatic detection of pathological voices using complexity

measures, noise parameters, and mel-cepstral coefficients. IEEE Transactions on Biomedical Engineering, Vol. 58, No. 2 (2011).
2. Puertas, F.J., Pin, G., María, J.M., & Durán, J: *Documento de consenso Nacional sobre el síndrome de Apneas-hipopneas del sueño*. Grupo Español De Sueño (2005).
3. Coccagna, G., Pollini, A. and Provini, F.: Cardiovascular disorders and obstructive sleep apnea syndrome. Clinical and Experimental Hypertension, vol. 28, pp. 217–224 (2006).
4. Lloberes, P., Levy, G., Descals, C. et al.: Self-reported sleepiness while driving as a risk factor for traffic accidents in patients with obstructive sleep apnoea syndrome and in non-apnoeic snorers. Respiratory Medicine, vol. 94, no. 10, pp. 971–976 (2000).
5. Penzel, T., McNames, J., de Chazal, P., Raymond, B., Murray, A. and Moody, G.: Systematic comparison of different algorithms for apnoea detection based on electrocardiogram recordings. Medical and Biological Engineering and Computing, vol. 40, no. 4, pp. 402–407 (2002).
6. Fox, A. W., Monoson, P. K., and Morgan, C. D.: Speech dysfunction of obstructive sleep apnea. A discriminant analysis of its descriptors," Chest, vol. 96, no. 3, pp. 589–595 (1989).
7. Robb, M. P., Yates, J. and Morgan, E. J.: Vocal tract resonance characteristics of adults with obstructive sleep apnea. Acta Oto-Laryngologica, vol. 117, no. 5, pp. 760–763 (1997).
8. Fiz, J. A., Morera, J., Abad, J. et al.: Acoustic analysis of vowel emission in obstructive sleep apnea, Chest, vol. 104, no. 4, pp. 1093–1096 (1993).
9. Godino-Llorente, J. I., Gomes-Vilda, P., and Blanco-Velasco, M.: Dimensionality reduction of a pathological voice quality assessment system based on gaussian mixture models and short-term cepstral parameters. IEEE Transactions on Biomedical Engineering, vol. 53, no. 10, pp. 1943–1953 (2006).
10. Blouet, R., Mokbel, C., Mokbel, H., Sanchez Soto, E., Chollet, G., and Greige, H.: BECARS: a Free Software for Speaker Verification. In Proceedings of the Speaker and Language Recognition Workshop, ODYSSEY, pp 145-148 (2004).
11. Fernandez R., Hernández L. A., López E., Alcázar J., Portillo G., and Toledano D. T.: Design of a Multimodal Database for Research on Automatic Detection of Severe Apnoea Cases. In Proceedings of 6th Language Resources and Evaluation Conference. LREC, Marrakech (2008).
12. Moreno, A., Poch, D., Bonafonte, A., Lleida, E., Llisterri, J., Mariño, J.B., & Nadeu, C.: ALBAYZIN Speech Database: Design of the Phonetic Corpus. In Proceedings of Eurospecch 93. Berlin, Germany, 21-23. Vol. 1 pp. 175-178 (1993).
13. Fernández Pozo, R., Blanco Murillo, J.L., Hernández Gómez, L., López Gonzalo, E., Alcázar Ramírez, J. and T. Toledano, D.: Assessment of severe apnoea through voice analysis, automatic speech, and speaker recognition techniques. EURASIP Journal on Advances in Signal Processing, Volume 2009, Article ID 982531, doi:10.1155/2009/982531
14. Blanco, J.L., Fernández, R., Torre, D., Caminero, F.J. and López, E.: Analyzing training dependencies and posterior fusion in discriminative classification of apnea patients base don sustained and connected speech. To appear in the Proceedings of the 12th Annual Conference of the International Speech Communication Association (2011).
15. Childers, D.G.: Speech Processing and Synthesis Toolboxes. John Wiley & Sons, 2000.
16. Farrús, M. and Hernando, J.: Using jitter and shimmer in speaker verification. IET Signal processing journal, Special Issue on Biometric Recognition, 2008. DOI: 10.1049/iet-spr.2008.0147, (2008)
17. Brookes, M.: VOICEBOX: Speech processing toolbox for Matlab. At personal are on Imperial College's website: http://www.ee.ic.ac.uk/hp/staff/dmb/voicebox/voicebox.html
18. Tsanas, A., Little, M.A., McSharry, P.E. and Ramig, L.O.: Nonlinear speech analysis algorithms mapped to a standard metric achieve clinically useful quantification of average Parkinson's disease symptom severity. Journal of the Royal Society Interface, Vol. 8, pp. 842-855 (2010).