On classification between normal and pathological voices using the MEEI-KayPENTAX database: Issues and consequences

Khalid Daoudi, Blaise Bertrac

On classification between normal and pathological voices using the MEEI-KayPENTAX database: Issues and consequences

Khalid Daoudi[1], *Blaise Bertrac*[2]

[1]INRIA Bordeaux-Sud Ouest. GeoStat team
200 av. de la vieille tour. 33405 Talence. France.
[2]University of Bordeaux 1
351 cours de la libération. 33405 Talence. France.
khalid.daoudi@inria.fr, blaise.bertrac@etu.u-bordeaux1.fr

Abstract

A large amount of research in pathological voice classification consider the task of feature extraction for discrimination between normal and dysphonic sustained vowels. The most widely used dataset for this purpose is the Massachusetts Eye & Ear Infirmary (MEEI) Voice Disorders Database commercialized by KayPENTAX Corp. During the last two decades, dozens of methods have been proposed to extract discriminative features from these signals in order to design accurate classifiers between the two classes of this database. The main contribution of this paper is to show that the normal and dysphonic sustained vowels of the KayPENTAX database are actually perfectly separable. This implies that this dataset is not suited for the normal-vs-dysphonic classification task, as long as the only concern is to achieve high classification accuracy. Indeed, we show that a single scalar parameter extracted from a matching pursuit decomposition of these signals (with a Gabor dictionary) yields a prefect classification accuracy (100 % with a large margin). We then discuss the implication of this finding on the precaution that should be taken with this database and on research in pathological voice detection in general.

Index Terms: Pathological voice classification, speech perturbation measure, dysphonia, matching pursuit, MEEI-KayPENTAX Voice Disorders Database.

1. Introduction

Quality assessment of pathological voices have gained an ever increasing interest in speech research. One reason is its practical impact in many areas of biomedical engineering related to voice disorders diagnosis and monitoring [1]. Another reason is the scientific challenges it raises as many common hypothesis and methods in "classical" speech processing become less effective. For instance, the use of nonlinear methods in this area keeps growing in order to overcome the limitation of standard linear methods [2, 3]. In this paper we are interested in the task which concentrates a large amount of research: classification between normal and pathological voices. There exists a broad spectrum of methods and systems which address this task using a wide range of databases and algorithms. A good review of such algorithms can be found in [4], a more recent review is provided in [3]. In [4], an interesting constructive discussion is provided on the methodological issues in existing methods to address this task. In particular, many methods use personal and inaccessible databases, with a disparity in recording and patients conditions. Moreover, from the algorithmic point of view, there exists a disparity in training/testing strategies. This makes it hard to draw consistent conclusions about the validity of the proposed methods. It is thus argue in [4] that a good option is to use the the Massachusetts Eye & Ear Infirmary Voice Disorders Database (KPdb) commercialized by KayPENTAX Corp. [5], because of its availability. The authors of [4] then stress the fact that even when KPdb is used, the is a disparity in data selection and experimental set ups, which renders impossible serious comparisons. Most of the time, unspecified subsets of KPdb are selected to run different kind of unreproducible experiments. They thus proposed a methodology, inspired from speaker recognition evaluation standards, to carry out training and testing of classifiers.

[4] highlights some methodological issues which (unfortunately) still exist in pathological voice detection research. In this paper, we reveal another major issue by arguing that even KPdb is not suited for the task of classification between normal and pathological voices, as long as the only concern is to achieve high classification accuracy. Many methods for pathological voice classification have been proposed using this database. The majority consists in defining several (more or less involved) features and then using feature selection/fusion algorithms to design the best possible classifier. A good example is the work in [6] which achieves one of the best accuracy scores (98.3% for vowels) using the full KPdb, but with a relatively high number of features and HMM training. Another example is [7] which reports 100 % accuracy but using only an unspecified subset of 67 pathological vowels, and using a rather heavy and highly tuned method. In this paper, we show that a **single** scalar parameter derived from a classical matching pursuit [8] decomposition of these signals (with a Gabor dictionary) yields a prefect classification (100 % accuracy with a large margin) between the normal and pathological sustained vowels of KPdb. This parameter was introduced in [9] but was used (surprisingly) only on KPdb sentences. Our main contribution is to show that by using this parameter, a major discrepancy of the KPdb vowels dataset is revealed. Based on this finding, we present some key points that should be considered when using this database. We then discuss the urgency for the development of standard corpora and evaluations in pathological voice research.

The paper is organized as follows. In the next section, we give a brief description of PKdb. In section 3, we present the basic of the matching pursuit algorithm and the work of [9]. We then show that the experimental set up of [9] leads to misleading conclusion. In section 4, we use a feature introduced in [9] to

reveal a major issue in PKdb. Finally, in section 5, we discuss some implications of this finding.

2. The MEEI-KayPENTAX Voice Disorders database (KPdb)

The MEEI-KayPENTAX Voice Disorders database [5] was released in 1994 and has been developed by the MEEI Voice and Speech lab and the Kay Elemetrics (now KayPENTAX) Corp. The recordings consist in sustained phonation of the vowel /ah/ (53 normal and 657 pathological) and utterance of the first sentence of the rainbow passage (53 normal and 662 pathological). All normal vowels and 77 pathological vowels are sampled at 50 kHz, while the remaining 580 pathological vowels are sampled at 25 kHz. 36 of the normal rainbow sentences are sampled at 25 kHz and 17 at 10 kHz. 648 of the pathological sentences are sampled at 25 kHz, 13 at 10 kHz and one at 50 kHz. More details about KPdb can be found in [5] , and [4] lists some key points to be careful about when handling it. In the last years, KPdb has been the most widely used dataset for research in pathological voice classification.

3. An application of matching pursuit to pathological voice classification

In this section, we first recall the basics of the matching pursuit (MP) algorithm. We then present our own analysis of the work [9] which used MP in pathological voice classification.

3.1. The Matching Pursuit (MP) algorithm

During the last two decades, the Matching Pursuit (MP) algorithm [8] has been widely used as a powerful tool for spare representation of signals using redundant dictionaries D of time-frequency functions ϕ_j (called atoms) generated by translation, dilatation/scaling and modulation of complex sinusoids:

$$\phi_j(t) = g(\frac{t-p_j}{s_j})exp\{j(2\pi f_j t + \omega_j)\}$$

where p_j is the atom position, f_j its central frequency, s_j its scale (or length) and g is the modulating function. When g is a Gaussian, D is the Gabor dictionary which we will use in all the experiments of this paper.

MP is a greedy algorithm which iteratively approximates a signal $x(t)$ by a projecting it onto the overcomplete dictionary D:

$$R_x^n(t) = \langle R_x^n(t), \phi_j \rangle \phi_j + R_x^{n+1}(t), \quad (1)$$

with $R_x^0(t) = x(t)$ at the first iteration $n = 0$. At each iteration n, a single atom ϕ_n is selected such that:

$$\phi_n = \arg\max_{\phi_j \in D} |\langle R_x^n(t), \phi_j \rangle| \quad (2)$$

where $\langle \cdot, \cdot \rangle$ is the Hermitian inner product. After M iterations, the signal $x(t)$ is thus decomposed as:

$$x(t) = \sum_{n=1}^{M} a_n \phi_n(t) + e(t), \quad (3)$$

where a_n is the amplitude of atom ϕ_n and e is the residual error after M iteration.

Recently a toolkit which efficiently implements the matching pursuit algorithm has been released: the Matching Pursuit ToolKit (MPTK) which is based on [10] and can be downloaded from http://mptk.irisa.fr. It can be installed on various platforms (Windows, Linux and Mac OSX) and is now massively used as it is the best available toolkit for MP analysis. MPTK provides fast implementation of different kind of dictionaries, including the Gabor dictionary. Another major advantage of using MPTK is that all the results presented in this paper can be easily reproduced.

3.2. Matching pursuit on KPdb's sentences

In [9], MP with a Gabor dictionary is used to discriminate between the normal and pathological "rainbow" sentences of KPdb. The authors used their own implementation of MP based on [11] (MPTK did not exist at that time). In a Gabor dictionary, the atom length s_j is generally taken as a power of 2 : $s_j = 2, ..., 2^J$. In [9], three features are defined:

- $Ocmax = max\{O_j, j = 1, ..., J\}$, where O_j is the number of occurrences of selected atoms with length s_j in M iterations of MP (eq. (3)).
- $Ocmean = \frac{\sum_{j=1}^{J/2} O_j}{J/2}$
- $Fr = \frac{M_{lf}}{M}$, where is M_{lf} the number of selected atoms whose center frequencies f_n are below half the sampling frequency.

The authors then used the 3-dimensional vector $[Ocmax, Ocmean, Fr]$ as the input feature for a linear discriminant analysis (LDA) classifier, with $J = 14$ and $M = 2000$. They used the one-leave-on-out method for training and testing the classifier. Their experiments were however carried out only on 51 normal sentences (2 missing without justification) and a subset of unspecified 161 pathological sentences (as is unfortunately the case in many research papers). Table 1 shows the classification scores reported in [9]. N-N (resp. P-P) stands for normal (resp. pathological) voices correctly classified.

Table 1: *Classification scores reported in [9]*

Feature/Accuracy	N-N %	P-P %
$[Ocmax, Ocmean, Fr]$	96.1	92.5

As argued before, the lack of transparency in experiments procedures can yield completely misleading conclusions. A typical example are the conclusions of [9]. Indeed, we have conducted the same experiment as [9] using MPTK. The only difference is that, this time, we use **all** the sentences of KPdb. We also checked the individual discriminative power of each of the 3 features. Table 2 shows the classification scores we obtain.

Table 2: *Classification scores on* **all** *KPdb sentences*

Feature/Accuracy	N-N %	P-P %
$Ocmax$	69.4	95.9
$Ocmean$	77.7	82.1
Fr	0	100
$[Ocmax, Ocmean, Fr]$	88.8	88.1

The results we obtain are in complete contradiction with those reported in [9]. First, the Fr feature does not provide any

discriminative information. Second, the accuracy score of the $[Ocmax, Ocmean, Fr]$ feature vector, which actually collaps to $[Ocmax, Ocmean]$, is significantly lower than the one reported in [9]. This example highlights what has been already reported in [4], namely the lack of transparency/consistency in data selection and classification methodology does not permit to assess validity. This is unfortunate because, from the methodological point view, the work of [9] is actually very interesting and we will use it to reveal a major discrepancy in the PKdb. This is the purpose of the next section.

4. Matching pursuit on KPdb's vowels

We have shown that experimental set up used in [9] lead to misleading conclusions (from the classification accuracy point of view). However, the conceptual methodology of that work is very interesting. Indeed, surprisingly the authors of [9] did not apply their methodology on the KPdb vowels dataset. From our point view, global features such as $Ocmax$ are rather more suited to analyze dysphonia than dysarthria. We thus proceeded to evaluate the discriminative power of this feature on the *full* vowels dataset, and the results are striking. Figure 1 displays the histograms of $Ocmax$ on the normal and pathological vowels, in blue and red, respectively. For sake of representation quality, we chose the number of bins so that their ratio equals the ratio between the size of the normal dataset (53) and the pathological one (657). The most important observation is that the support of $Ocmax$ is $[723; 1144]$ for normal vowels and $[219; 607]$ for pathological ones. We have used here $J = 13$ given that there exists 5 files in the dataset which cannot be processed by MPTK with $J = 14$, because they are too short. However, the same behavior holds if one excludes these 5 files and uses $J = 14$ in MPTK. Note also that MPTK takes into account the difference in sampling frequency. The latter is an input parameter to the algorithm. However, to avoid any potential doubt, we have antialias-filtered and down-sampled to 25 kHz all the normal vowels and the 77 pathological vowels sampled at 50 kHz. Figure 2 shows that the same behavior holds. In this case, the support of $Ocmax$ is $[520; 776]$ for normal vowels and $[219; 445]$ for pathological ones. In order to make sure that the difference in duration between normal vowels ($\sim 2 - 3s$) and pathological ones ($\sim 0.4 - 1.4s$) has no influence on these results (given that $M = 2000$ for all), we define the following duration-independent feature:

$$\frac{O^+cmax}{O^-cmax} = \frac{max\{Oj, j = 1 + \lceil J \rceil /2, ..., J\}}{max\{Oj, j = 1, ..., \lceil J \rceil /2\}}$$

This feature measures the ratio between the weight of dominant long and short atoms. The histograms of $\frac{O^+cmax}{O^-cmax}$, without downsampling, are shown on Figure 3. In this case, the support of $\frac{O^+cmax}{O^-cmax}$ is $[7.5; 59.7]$ for normal vowels and $[0.7; 4.9]$ for pathological ones. The same behavior holds when downsampling the 50 kHz files to 25 kHz, as shown in 4. In that case, the support of $\frac{O^+cmax}{O^-cmax}$ is $[4.6; 35.8]$ for normal vowels and $[0.7; 4.1]$ for pathological ones. Note finally that the choice of $M = 2000$ is ad-hoc and the same behavior persists for a large range of M values.

These results show that a single saclar parameter, derived from MP, allows perfect discrimination between the normal and pathological sustained vowels of KPdb. Indeed, any classical classifier and any training/testing strategy would lead to perfect accuracy. Thus, by considering for instance a simple thresholding on $\frac{O^+cmax}{O^-cmax}$, one gets 100% accuracy with a large confidence interval:

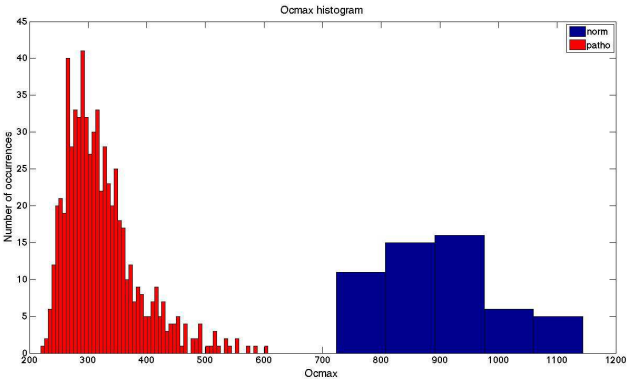

Figure 1: *$Ocmax$ histograms without changing sampling frequency.*

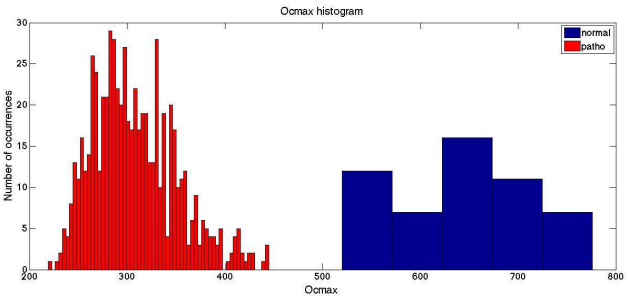

Figure 2: *$Ocmax$ histograms after downsampling all the 50 kHz files to 25 kHz.*

Table 3: *Classification accuracy scores with $Ocmax$ and $\frac{O^+cmax}{O^-cmax}$ on all KPdb vowels.*

Feature/Accuracy	N-N %	P-P %
$Ocmax$	**100**	**100**
$\frac{O^+cmax}{O^-cmax}$	**100**	**100**

These results definitely confirm the (already known) usefulness of MP in pathological voice detection. However, a naive interpretation would be to overestimate its strength in this setting. A more realistic interpretation is that MP acts as a "nonlinear mirror" which readily reflects the strong difference between the normal and pathological datasets. $Ocmax$ is indeed a measure of the weight of the dominant structures (atoms having a particular length) in a signal. The results simply reflect that the latter are significantly heavier in the normal vowels that in the pathological ones. It is thus fair to expect (for instance by focusing on this property) that other straightforward features, derived from other techniques, would achieve the same results. Consequently, a realistic interpretation is that dysphonia of the recorded patients is so pronounced that it is straightforward to detect it by an objective (automatic) evaluation. Thus, the task of normal-vs-dysphonic classification on this dataset would be inconsistent if this fact is not taken into account. We discuss this matter in the next section.

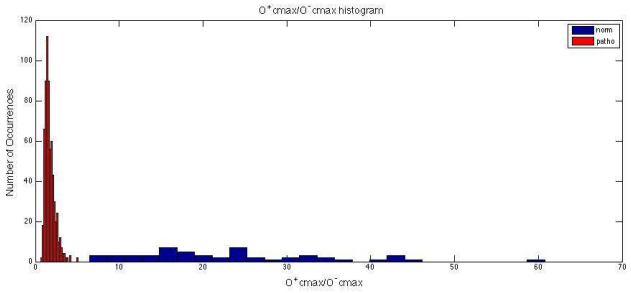

Figure 3: $\frac{O^+cmax}{O^-cmax}$ *histograms without changing sampling frequency.*

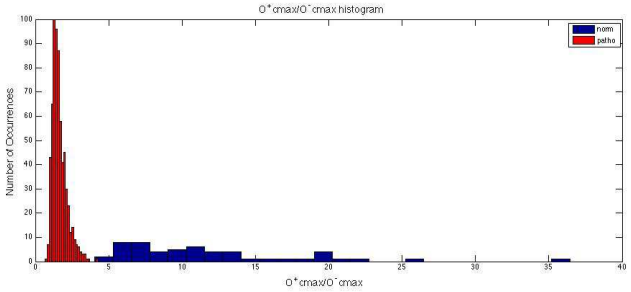

Figure 4: $\frac{O^+cmax}{O^-cmax}$ *histograms after downsampling all the 50 kHz files to 25 kHz.*

5. Discussion

We now start a discussion on some implications of the results of the previous section. We first list some key points that should be considered when using KPdb:

- The first implication of our results is that the KPdb vowels dataset is not suited for the normal-vs-dysphonic classification task, **but only if** the only concern is to achieve high classification accuracy on this dataset. Indeed, in that case, any method which does not achieve perfect classification would be irrelevant. From this perspective, KPdb can be used as a "toy example" dataset. That is, one starts by checking whether perfect accuracy is reached on KPdb before proceeding further with other datasets.
- Achieving high classification accuracy should not be (and is not) always the only concern. Research in this area is anyway still far away from having such an objective central, as compared to speech/speaker recognition for example. KPdb can thus still be used if, for instance, the main goal is to develop features which improve knowledge about pathological voices from the acoustic and/or physiological perspective. A typical example is the standard perturbation measures which are widely used (Jitter, Shimmer, HNR,...). Alone or combined, these features do not achieve perfect classification on KPdb, however they are acoustically and physiologically meaningful which makes them useful in practice and easy to interpret by clinicians. Thus, any effort in this direction (and there exist many) should not worry about classification scores.
- One can fairly expect that the KPdb sentences data set also exhibits the same behavior, because the speakers of the vowels and the sentences are the same. If proved, then the last two points hold also for this dataset.
- As reported in [12], most existing features and systems focus on for classification between normal and pathological voices, while there have been only few research in discriminating between different categories of pathologies. We believe more research like [12] is required and that much more effort should be put on this problem which is scientifically more challenging than the classical normal-vs-pathological task. Moreover, it has many direct applications in biomedical engineering, such as differential diagnosis assist. The different pathology groups of KPdb can thus serve as an exploratory ground to develop discriminative features/classifiers between pathologies. The latter could indeed be used/adapted later in real-world biomedical applications.

In our view, the most important issue that this work highlights is the absence of well-designed standard corpora in pathological voice detection. The authors of [4] argued that KPdb is a good choice. Our results show that, at best, KPdb is a default choice on which serious carefulness should be taken. This renders the situation more complicated than it was. It is then urgent that the research community in this field gathers its efforts to face this major problem of data. Existing personal databases needs to be made available when there is no legal/ethical/technical problems. Providing free datasets is always the best option, however commercial corpora are also welcome given the huge lack of data. Experts, from academia and industry, should set up standards and means for data collection, share and evaluation. Speech/Speaker recognition would have never achieved their current level of progress without the strong effort on corpora and evaluation standards. Pathological voice research has no choice but to follow the same steps, otherwise it will struggle to follow and profit from the scientific and technological progress. The need for pathological voice analysis tools from the medical sector keeps growing (surgery, phoniatry, neurodegenerative diseases,...). It is thus necessary and urgent to create the best possible research environment in order to fulfill this need. Meanwhile, any proposed method, claiming improvement w.r.t. other techniques, has to provide the necessary material to allow fair comparisons. Otherwise, it would only reinforce the existing confusion and should be considered irrelevant.

6. Conclusions

We showed that a single scalar parameter derived from a matching pursuit decomposition allow perfect discrimination between the normal and the pathological sustained vowels of KPdb. Consequently, we argued that the KPdb vowels dataset is not suited for the normal-vs-dysphonic classification task, if the only concern is to achieve high classification accuracy on this dataset. We then listed some elements that should taken into consideration when using KPdb and proposed some scenarios where this database can still be useful. Finally, we discussed the major problem of lack of corpora in pathological voice detection research.

7. References

[1] Freed, D.,"Motor speech disorders", in Thomson Learning [Ed], 2000.

[2] Titze, I.R.,"Principles of Voice Production", in National Center for Voice and Speech, Iowa City, USA [2nd Ed], 2000.

[3] Tsanas A. et al, "Novel speech signal processing algorithms for high accuracy classification of Parkinsons disease", IEEE Trans. Biomedical Engineering, 59(5):1264–1271, 2012.

[4] Saenz-Lechon, N. et al, "Methodological issues in the development of automatic systems for voice pathology detection", Biomedical Signal Processing and Control 1:120–128, 2006.

[5] Massachusetts. Eye and Ear Infirmary,"Voice disorders database, (Version 1.03 cd-rom)", Kay Elemetrics Corp., Lincoln Park, NJ, 1994.

[6] Dibazar, A.A. et al "Feature Analysis for Automatic Detection of Pathological Speech", in Proc. IEEE Engineering in Medicine and Biology Society (EMBS) Meeting, 182–183, 2002.

[7] Arjmandi, M.K. and Pooyan, M., "An optimum algorithm in pathological voice quality assessment using wavelet-packet-based features, linear discriminant analysis and support vector machine", Biomedical Signal Processing and Control 7:3–19, 2012.

[8] Mallat, S. and Zhang, Z., "Matching pursuit with time-frequency dictionaries", IEEE Trans. Signal Processing, 41(12):3397–3415, 1993.

[9] Umapathy, K. et al, "Discrimination of Pathological Voices Using a Time-Frequency Approach", IEEE Trans. Biomedical Engineering, 52(3):421–430, 2005.

[10] Krstulovic, S. and Gribonval, R. "MPTK: Matching Pursuit made Tractable", in Proc. ICASSP'2006, (3):496–499, 2006.

[11] Gribonval, R. "Fast Matching Pursuit with a multiscale dictionary of Gaussian Chirps", IEEE Trans. Signal Processing, 49(5):994–1001, 2001.

[12] Markaki, M. and Stylianou, Y., "Voice Pathology Detection and Discrimination based on Modulation Spectral Features", IEEE Trans. Audio, Speech and Language Proc., 19(7):1938–1948, 2011.