Application of Automatic Speaker Recognition techniques to pathological voice assessment (dysphonia)

C. Fredouille[1], *G. Pouchoulin*[1], *J.-F. Bonastre*[1], *M. Azzarello*[2], *A. Giovanni*[2], *A. Ghio*[3]

[1] LIA-CNRS, Avignon (France)
[2] LAPC, Marseille (France)
[3] LPL-CNRS, Aix-en-Provence (France)

Abstract

This paper investigates the adaptation of Automatic Speaker Recognition (ASR) techniques to the pathological voice assessment (dysphonic voices). The aim of this study is to provide a novel method, suitable for keeping track of the evolution of the patient's pathology: easy-to-use, fast, non-invasive for the patient, and affordable for the clinicians. This method will be complementary to the existing ones - the perceptual judgment and the usual objective measurement (jitter, airflows...) which remain time and human resource consuming.

The system designed for this particular task relies on the GMM-based approach, which is the state-of-the-art for speaker recognition. It is derived from the open source ASR tools (LIA_SpkDet and ALIZE) of the LIA lab.

Experiments conducted on a dysphonic corpus provide promising results, underlining the interest of such an approach and opening further research investigation.

1. Introduction

In the medical domain, the assessment of pathological voice quality is an important issue, inducing a large amount of research in multidisciplinary domains [1][2]. Concerning dysphonic voices, which this paper is focused on, the voice dysfunction may be assessed following two different approaches: the perceptual judgment or the objective assessment based on jitter, airflows... measurement. On the one hand, the perceptual judgment consists in qualifying and quantifying the vocal dysfunction by listening the speech production of a patient. This method is currently the most used by the clinicians. However, it is largely controversed in voice research and demonstrates various drawbacks. First of all, the perceptual judgment has to be performed by an expert jury to increase the reliability of the analysis because of its intrinsec subjectivity. Nevertheless, due to the lack of universal assessment scales and other factors like professional background and experience of the experts, or knowledge of the patient's history, the perceptual judgment may involve large intra and inter-variability in the judgments. Besides, the perceptual analysis is very costly in time and human resources and cannot be planned regularly. On the other hand, the objective measurement-based analysis consists in qualifying and quantifying the vocal dysfunction by analyzing acoustical, aerodynamic and physiological measurements. These measurements may be directly extracted from patient's speech utterance using a simple computer-based system or may require special devices like *EVA*2^{TM} (Computerised Vocal Assessment, SQLab) [3], designed for the recording and the study of many parameters of the speech and voice production. All the investigations made on the objective measurement-based analysis demonstrate the requirement of combining different measurements in order to cope with the multidimensional nature of the voice and to increase the reliability of the analysis. In [4], discriminant analysis was performed to detect correlation between jury classification and combinations of parameters. Results showed that a nonlinear combination of only six parameters (range, LC, ESGP, MPT, signal-to-noise ratio, and F0) allowed 86% concordance with jury classification. Like the perceptual judgment, the objective analysis has some limitations. First of all, the objective analysis often relies on statistical approaches (like linear discriminant analysis, correlation estimation...) applied on the collection of measurements, which may be strongly dependent on the observed patient population in terms of quality and quantity. Besides, most of the objective analysis relies on the study of some sustained vowels only, which are not representative of the continuous speech [5]. Finally, the use of special devices for measurement gathering may be expensive and costly in time. Therefore, these systems are in limited use in routine examination.

In this paper, we investigate the adaptation of automatic techniques largely used in Automatic Speaker Recognition for dysphonic voice assessment. This study aims at proposing a complementary method to the perceptual judgment and objective analysis, well suited for observing in a regular manner the evolution of voice dysfunction among the patients; to respond to this application, the proposed method has to be easy-to-use, quick, non-invasive for the patients and affordable for the clinicians.

The system designed for this particular task relies on LIA_SpkDet, the Automatic Speaker Recognition (ASR) toolkit of the LIA lab. This toolkit is based on the ALIZE platform [6] designed and developed by the LIA in the framework of the Technolangue program. Both LIA_SpkDet and ALIZE are open source (respectively GPL and LGPL)[7].

This paper is organized as follows: the dysphonic voice assessment and the corpus used for this first study are described in section 2. The next section (3) is dedicated to a description of the Automatic Speaker Recognition approach proposed in this study. Section 4 presents the experiments including protocols and results. Conclusion and perspectives on this study are finally provided in section 5.

2. Dysphonic voice assessment

This work aims at studying the behaviour of classical techniques used in the Automatic Speaker Recognition (ASR) domain when applied to pathological voices. Different studies have investigated the use of automated speech analysis for pathological voice detection [8][9][10], focusing on the feature analysis only.

In this paper, we propose to adapt the overall ASR techniques required for the pathological voice assessment task. Basically, these ASR techniques rely on Gaussian Mixture Models (GMM)[11] and adaptation algorithms, which require few constrained speech material, are very fast and well adapted for keeping track of the pathology evolution of a patient.

In this study, the focus is made on functional dysphonias (nodules, polyps, oedema, cysts...). These dysphonias are classified according to the G parameter of the GRBAS scale[1] proposed by [12]. On this G-based scale, a normal voice is rated as 0, a slight dysphonia as 1, a moderate dysphonia as 2 and finally, a severe dysphonia as 3.

The dysphonic corpus used in this paper (named DV) is composed of 80 voices of females aged 17 to 50 (mean: 32.2). The speech material is obtained by reading the same short text which the duration varies from 13.5 to 77.7 seconds (mean: 18.7s). Among the 80 voices, 20 voices are normal (G_0 voices), 20 are associated with a dysphonia of grade 1 (G_1 voices), 20 with a dysphonia of grade 2 (G_2 voices) and 20 with a dysphonia of grade 3 (G_3 voices). These perceptual grades were determined by a jury composed of 3 expert listeners. This perceptual judgment was carried out by consensus between the different members of the jury in order to limit inter-listener variability. Besides, the judgment was done during one session only in order to limit intra-listener variability.

3. Dysphonic voice classification system

The principle retained in this study consists in adapting a classical speaker recognition system to the dysphonic voice classification. A speaker recognition system is a supervised classification system able to differentiate speech signals into classes (two classes for speaker verification, n-classes for speaker identification). Each class of signal belongs to a given speaker and is learnt using a set of examples from this speaker. A composite class could be also proposed by grouping several classes learnt independently or by learning an unique class on speaker's signals belonging to this class. Two adaptations are made to suit the speaker recognition system to this task. Firstly, a class does not longer correspond to a given speaker but to the specific pathology targeted (or to a grade of this pathology). The class is then learnt using data from a set of speakers presenting this pathology. Obviously, the voices used for training a pathological class could not be included in the test set, in order to differentiate pathology detection and speaker recognition. The second modification applied to the speaker recognition system is the representation of the audio data, which could be optimized for pathology discrimination.

The speaker recognition technique used in this study is based on GMM-based approach, which is the state-of-the-art for speaker recognition [13]. This approach consists in three phases:

- a parameterization phase;
- a model training phase;
- a classification phase.

3.1. Parameterization phase

All the speech material is parameterized as follows: each signal is characterized by 16 MEL frequency cepstral coefficients (MFCC). These MFCC coefficients are obtained from 24 filter bank coefficients applied on 20ms Hamming windowed frames at a 10ms frame rate. The first derivatives of the MFCC coefficients are added to the parameter vectors.
A frame removal processing, based on the energy, is applied on each speech signal in order to delete silence segments. The parameter vectors are then normalized to fit a 0-mean and 1-variance distribution.
The MFCC computation is done thanks to the (GPL) SPRO toolkit [14].

3.2. Training phase

The GMM models representing the pathological classes are built as follows:

- a generic GMM model is first estimated thanks to the EM algorithm, maximizing the Maximum Likelihood criterion (ML) on a French read-speech corpus composed of 76 female speech utterances of 2 minutes each. This female population is extracted from the BREF corpus [15], which is entirely separate from the dysphonic corpus and the targeted task.
- the pathological class models are then derived from the generic GMM model thanks to the MAP adaptation technique [16]. Only means are adapted (as classically done in speaker verification). This technique increases the robustness of the models, especially when few speech material is available for the model training.

All the GMM models are composed of 128 gaussian components with diagonal covariance matrices.

3.3. Classification phase

During this phase, an input signal is presented to the system and compared to the N GMM-models depending on the targeted task. This comparison relies on an averaged frame-based likelihood computation - $L(y|M)$ - between a given model M and the input signal y.

4. Experiments

In this paper, two tasks have been investigated:

- Task1: normal and dysphonic voice classification;
- Task2: dysphonic voice assessment according to the G-based scale.

4.1. Experimental protocols

As seen in section 3, the training data used for learning the pathological classes should not be used for testing. In other words, the speakers including in the training set should not be present in the testing set. As the dysphonic voice (DV) corpus available for this study is relatively small (80 voices), it is not well suited to split it into two separate subsets. Consequently, some special protocols have been designed for each task (Task1 and Task2) in order to respect this constraint while providing more statistically significant results. These protocols rely on the *leave-x-out* techniques. In this paper, it consists in discarding x speakers from the experimental set, learning some models on the remaining data and testing the x speaker data using the models. This scheme is repeated until reaching a sufficient number of tests.

[1] the G parameter refers to as the degree of hoarseness, which is considered as the global grade for assessing a dysphonic voice.

4.1.1. *Task1 protocol (P1)*

For this task, two different models have to be estimated: the $M_{\overline{d}}$ and M_d GMM models corresponding to the normal and dysphonic voices (the Task1 consists in determining whether a given voice is normal or dysphonic). Therefore, the 20 normal voices (G_0 voices) and 60 dysphonic ones (G_1, G_2 and G_3 voices) available in the DV corpus are used as follows:

- All the subsets of 18 voices among the G_0 set are used to estimate a normal voice GMM model per each;
- Different subsets of 18 dysphonic voices[2], equally-balanced over the G_1, G_2 and G_3 voices, are used to estimate a dysphonic GMM model per each.
- During the test, each of the 80 voices available in the DV corpus are first compared to $N_{\overline{d}}$ normal voice GMM models, resulting in an averaged normal voice likelihood $L(y|M_{\overline{d}})$ and secondly compared to N_d dysphonic voice models, resulting in the averaged likelihood $L(y|M_d)$;
- The decision relies on the maximum between the two likelihoods.

In summary, 80 tests are performed (20 "normal" tests vs 60 "dysphonic" ones). All the models are estimated on 18 voices and a voice (a speaker) is systematically discarded from the models while being tested.

4.1.2. *Task2 protocol (P2)*

The Task2 consists in assessing a given voice following the G-scale rates. Four classes are in competition in the recognizer and four corresponding models (M_{G_0}, M_{G_1}, M_{G_2} and M_{G_3}), one by G-scale grade, are needed. In this context, the 20 normal voices (G_0 voices) and the 60 dysphonic ones (G_1, G_2 and G_3 voices) available in the DV corpus are used as follows:

- All the subsets of 19 voices among the G_0 set are used to estimate a model per each - $M_{G_0}^{-x}$ - with x the discarded voice; The same process is applied on the set G_1, G_2 and G_3; This results in 20 different models available per grade.
- When testing voice y relating to grade i, y is first compared to model $M_{g_i}^{-y}$ leading to the likelihood $L(y|M_{g_i}^{-y})$ computation. Then, an averaged likelihood is computed for all the other grades (different from i), by using the grade dependent model sets (average on 20 likelihoods per grade).
- The decision relies on the maximum over the four likelihoods.

4.2. Results and discussion

Table 1 provides the correct classification rates obtained for the first task (Task1 - normal and dysphonic voice classification). An overall correct classification rate of 85% is reached on this task. It is interesting to observe the behavior of the system, regarding the grade of the 80 voices tested, illustrated by the confusion matrix (table 2). Indeed, we can note that the confusion comes mainly from the intermediate grades (1 and 2), the grades 1 and 3 behaving quite well.

Table 1: Normal and dysphonic voice classification (Task1): correct classification rate (in %) on the DV corpus - 80 tests (20 normal and 60 dysphonic).

	Correct classification rate in % (Succeeded Test Nb/Total Test Nb)		
System	Normal	Dysphonic	Overall
32 MFCC + Δ	95.0 (19/20)	81.7 (49/60)	85.0 (68/80)

Table 2: Confusion matrix built from normal and dysphonic classification tests (Task1): Classification system response according to the grade of the test voices - 80 tests (20 normal and 60 dysphonic)

	Classification system response	
Test Gr.	Normal	Dysphonic
0	**19**	1
1	6	**14**
2	5	**15**
3	0	**20**

Concerning the second task (Task2 - dysphonic voice assessment), results are provided in table 3. The high performance obtained on grade 0 (normal voices) in the first task is confirmed here. Nevertheless, a significant loss of performance is observed for the other grades related to the dysphonic voices, especially for the grade 2. The confusion matrix, provided in table 4, shows that the confusions involve, in most of the cases, the adjacent grades[3]. This observation is particularly true for the extreme grades 0 and 3 for which all the misclassification errors concerns grades 1 and 2 respectively. Regarding the grade 2, for which the performance is very poor, it can be observed that the associated voices are distributed almost uniformely between grades 1 and 2.

These results are encouraging for several reasons:

- the first studies on objective measurement based systems like *EVA*^{TM} for instance reached 66% of correct classification for dysphonic voice assessment, which is quite similar to the performance obtained by the GMM-based system;
- they show that dysphonic information may be caught by a GMM-based system, even if very few speech material is available for the training phase, thanks to the adaptation techniques;
- no particular attention was given, in this paper, to the choice of acoustic parameters. For instance, prosodic information which seems to be suitable for dysphonic voice discrimination has to be integrated to the current system;
- they show that intermediate classes should help in grade assessment when confusion is present for a single voice between two adjacent grades;
- obviously, these experiments have to be validated on a larger corpus since it is quite difficult to bring conclusions on 80 test samples.

[2] These subsets are built randomly, under the constraint that all the voices are used at least once.

[3] except 2 confusions between the grades 2 and 0, and 1 between grades 1 and 3

Table 3: Dysphonic voice assessment (Task2): correct classification rate (in %) on the DV corpus - 80 tests (20 normal and 60 dysphonic).

	Correct classification rate in % (Succeeded Test Nb/Total Test Nb)				
System	Gr. 0	Gr. 1	Gr. 2	Gr. 3	All
32 MFCC + Δ	95 (19/20)	70 (14/20)	45 (9/20)	65 (13/20)	69 (55/80)

Table 4: Confusion matrix built from dysphonic voice assessment tests (Task2): Classification system response according to the grade of the test voices - 80 tests (20 tests per grade)

	Classification system Response			
Test Gr.	0	1	2	3
0	**19**	1	0	0
1	3	**14**	2	1
2	2	7	**9**	2
3	0	0	7	**13**

5. Conclusion

This paper is concerned with the pathological voice assessment (functional dysphonic voices). Two main approaches are currently used for this particular task: the perceptual judgment and the objective measurement-based analysis; Nevertheless, both present some limitations. This paper investigates an original method, derived from the Speaker Recognition domain to cope with part of these limitations. Indeed, the aim of this study is to propose a system well adapted to keep track of the evolution of the patient's pathology in a regular manner.

The proposed system is based on the GMM approach, state-of-the-art in the Speaker Recognition domain. It was evaluated on a dysphonic corpus for two different tasks: normal and dysphonic voice classification and dysphonic voice assessment. The performance obtained for both tasks was very promising (despite the limited size of the corpus) and very closed to those obtained by the first studies on the objective measurement-based analysis. They have pointed out the interest of such an approach, especially in the case of few speech material available and have opened further research investigation like the integration of acoustic parameters more suitable for pathological voice discrimination (prosodic information for instance). Furthermore, it is important to notice that the current approach allows to investigate in smaller units than an entire voice utterance for the decision making. Therefore, future work will be focused on the behavior of the classification system at a segmental level (phoneme or shorter events) in order to evaluate if the dysphonic phenomenas are uniformely spread over the speech production or more located at some specific zones of the speech.

Finally, this work has to be validated on a larger dysphonic corpus to increase the result reliability.

6. References

[1] F. L. Wuyts, M. S. De Bodt, G. Molenberghs, M. Remacle, L. Heylen, B. Millet, K. Van Lierde, J. Raes, P. H. Van de Heyning, The dysphonia severity index: an objective measure of vocal quality based on a multiparameter approach, Journal of Speech, Language, and Hearing Research 43 (2000) 796–809.

[2] J. Revis, L'analyse perceptive des dysphonies : approche phonétique de l'évaluation vocale, Phd thesis, Université de la Méditerranée (2004).

[3] B. Teston, B. Galindo, A diagnosis of rehabilitation aid workstation for speech and voice pathologies, in: Proceedings of European Conference on Speech Communication and Technology (Eurospeech 95), 1995, pp. 1883–1886.

[4] P. Yu, M. Ouakine, J. Revis, A. Giovanni, Objective voice analysis for dysphonic patients: a multiparametric protocol including acoustic and aerodynamic measurements, Journal Voice 15 (2001) 529–542.

[5] V. Parsa, D. G. Jamieson, Acoustic discrimination of pathological voice: sustained vowels versus continuous speech, Journal of Speech, Language, and Hearing Research 14 (2001) 327–339.

[6] J.-F. Bonastre, F. Wils, S. Meignier, Alize, a free toolkit for speaker recognition, in: Proceedings of International Conference on Acoustics Speech and Signal Processing (ICASSP 2005), Philadelphia, USA, 2005.

[7] ALIZE and LIA_SpkDet tools, `http://www.lia.univ-avignon.fr/heberges/ALIZE/`.

[8] M. Wester, Automatic classification of voice quality: Comparing regression models and hidden markov models, in: VOICEDATA98, Symposium on Databases in Voice Quality Research and Education, 1998, pp. 92–97.

[9] A. A. Dibazar, S. Narayanan, T. W. Berger, Feature analysis for automatic detection of pathological speech, in: Engineering Medicine and Biology Symposium02, Vol. 1, 2002, pp. 182–183.

[10] C. Maguire, P. de Chazal, R. B. Reilly, P. Lacy, Identification of voice pathology using automated speech analysis, in: Third International Workshop on Models and Analysis of Vocal Emission for Biomedical Applications, 2003.

[11] D. A. Reynolds, A gaussian mixture modeling approach to text-independent speaker identification, Phd thesis, Georgia Institute of Technology (1992).

[12] M. Hirano, Psycho-acoustic evaluation of voice : GRBAS Scale for evaluating the hoarse voice. Clinical Examination of voice, Springer Verlag, 1981.

[13] F. Bimbot, J.-F. Bonastre, C. Fredouille, G. Gravier, I. Magrin-Chagnolleau, S. Meignier, T. Merlin, J. Ortega-Garcia, D. Petrovska, D. A. Reynolds, A tutorial on text-independent speaker verification, EURASIP Journal on Applied Signal Processing 2004 (4) (2004) 430–451.

[14] G. Gravier, Spro: a free speech signal processing toolkit, `http://www.irisa.fr/metiss/guig/spro/`.

[15] L. Lamel, J. Gauvain, L. Eskénazi, "BREF, a large vocabulary spoken corpus for french", in: Proceedings of European Conference on Speech Communication and Technology (Eurospeech 99), 1991.

[16] D. A. Reynolds, T. F. Quatieri, R. B. Dunn, Speaker verification using adapted gaussian mixture models, Digital Signal Processing (DSP), a review journal - Special issue on NIST 1999 speaker recognition workshop 10 (1-3) (2000) 19–41.