Burst-based Features for the Classification of Pathological Voices

Julie Mauclair†  Lionel Koenig*  Marina Robert, Peggy Gatignol‡

† Université Paris Descartes, 45 rue des Saints Pères 75006 Paris, France
*IRIT - Université de Toulouse, 118 route de Narbonne, 31062 Toulouse Cedex 9, France
‡Service ORL et de chirurgie de la face et du cou, Hôpital de la Pitié Salpêtrière, 75013 Paris, France

firstname.lastname@parisdescartes.fr, †irit.fr, ‡psl.aphp.fr}

Abstract

Classification of pathological voices is an important problem for early detection and diagnosis. The automatic analysis is a useful complementary tool to other methods based on direct observation. The pathological voices that are studied in this paper are the voices of patients who have a peripheral facial paralysis which, among other pronunciation impairments, affects their ability to pronounce bilabial sounds. The idea is then to use a burst detector to compute acoustic features and provide them to a classifier in order to automatically determine the degree of the voice impairment. The speech database that is used through the paper is unique and was recorded in a soundproof cabin at the La Pitie Salpêtrière Hospital in Paris, France. Even if the database is in french, the features that are used in those experiments are independent from the language. The speech recordings used for the experiments are isolated sentences. Two kinds of artificial neural networks are studied for the classification task, a multilayer perceptron and a neural network based on learning vector quantization. Our results show a correlation between the burst-based acoustic features computed from the voices and the degree of the impairment that affects patients.

Index Terms: Burst detection, classification, pathological voices, automatic speech processing, peripheral facial paralysis, neural network

1. Introduction

Peripheral facial paralysis is a frequent pathology with numerous consequences in the patient’s life. Among them, the paralysis impairs their articulation ability [1]. Currently, there is a large set of tools that are available to measure the degree of impairment but most of them only take care about the physical aspect of the face and do not measure the patient’s ability to pronounce words and sentences.

The clinical grading scale that is internationally used is the House and Brackmann scale [2]. It rates the degree of facial impairment and integrates the evaluation of sequelae from stage I (normal/no paralysis) to stage VI (no movement of the face). It is a six-point subjective grading scale and it only rates the physical aspects of the paralysis (mainly the forehead, the eyes and the mouth) according to a direct observation by a therapist. Table 1 focuses on the mouth abilities reported through the various stages of the House and Brackmann grading scale. For patients with peripheral facial paralysis from stage III, literature reports dramatic psychological and functional disorders such as ocular disorders, synkinesis, hemi-spasm, impaired chewing and swallowing, and moreover, articulatory difficulties [1].

Within the field of speech processing, the analysis of voice disorders is used to help medical diagnosis. The automatic acoustic analysis of voices is a non invasive method and is widely investigated for laryngeal pathologies and dysphonic voices [3, 4] to provide an objective assessment for speech therapists. For facial impairment, an automatic voice analysis could help the patient in measuring their own progress while being in therapy. It is a very long process and it is difficult for them to be objective in their recovery, even if the improvement is stated by relatives or by the doctors.

Most of the time in the literature, the acoustic analysis of pathological voices is carried out on sustained vowels [4] and uses either linear signal processing methods and/or nonlinear methods [4, 5, 6]. Usually, the aim is to discriminate pathological from non pathological voices [3, 4, 5]. If the paper deals with the classification between various stages of voice pathology, the scale that is used is the GRBAS scale [7] which is a scale for dysphonic voices. As the voices we collected are not dysphonic, we will prefer the House and Brackmann scale and try to find out if there is a correlation between the physical facial impairment that is taken into account in this scale and the articulatory difficulties of the patients. In order to apprehend those articulatory difficulties, the aim is then to find a set of acoustic features that will represent the degree of impairment of the patient. When listening to the patient’s speech, the clearest impact of the paralysis is that the plosives are affected [1]. Patients are unable to utter the burst needed in order to pronounce those sounds precisely. In order to verify this hypothesis, a previous experiment was carried out on isolated words uttered by the patients of our database [8]. On 24 words uttered by 48 subjects where the target phonemes were the stops /p/, /b/, /t/ and /d/, the number of bursts that were observed on the spectrogram was manually counted. The bursts caused by dental occlusives /l/ and /d/ were all visible for all the subjects on the spectrogram unlike those caused by the phonemes /p/ and /b/. For bilabial stops, the number of bursts visible decreased with the increase of the stage severity. The results were analysed thanks to the Wilcoxon signed-rank test [9] and showed a significant difference between the number of bursts uttered by healthy subjects and those uttered by the other subjects in all other stages. The difficulty to pronounce bilabials precisely was also experimented by Professor Bruce E. Murdoch, in [10] with patients suffering from trigeminal, facial and hypoglossal nerve lesions.

For the classification of the acoustic vectors computed from the features based on the burst detection, two kinds of neural networks are studied. Neural networks are often used in classification tasks and also in voice disorders classification tasks [11]. Multilayer Perceptron (MLP) and Learning Vector Quantization (LVQ) algorithms are often experimented as they
allow a non linear classifications on data sets. In [11], the authors use MLP and LVQ to detect voice impairments by means of Mel Frequency Cepstral Coefficients (MFCC). The vocal pathologies studied are pathologies affecting vocal folds and the recorded sounds are sustained vowels. On a frame basis, the LVQ net yields to 94% of accuracy on discriminating pathological/non pathological voices. Through our paper, those two network structures will be computed for our classification task.

This paper is organized as follows. The database that was recorded from voices of patients with facial paralysis is described in section 2. In section 3, we will present the burst detector that is used for the experiments and the features that will be given as input for the two classifiers will be detailed. Then, the experiments developed for this article using the MLP and the LVQ structures neural networks and the obtained results are explained in section 4. Finally, we conclude and discuss a future research.

2. Database

The experiments have been carried out on a speech database recorded at the La Pitié Salpêtrière Hospital in Paris, France. The database contains 70 speech recordings. Some of the recordings come from the same patient at the same or at various stages of the paralysis. The patients were recorded in a soundproof booth with a supercardioid microphone and on a numerical recorder using a 16 bits/44.1kHz linear PCM WAV audio resolution.

Various type of sentences were recorded. The patients were given various texts to read: isolated words, isolated sentences, newspaper texts, and they had to spontaneously answer to the question: "How do you make an omelet?". In the whole protocol, the database consists also in the measurement of the tense in the patients lips with a dynamometer, a phonetical test of intelligibility, a form that evaluates the motricity of the lips, the tongue and the face and the patients have to place their articulation satisfaction on a scale.

For this study, only the set of isolated sentences is exploited. There are 48 subjects that is to say 48 speech recordings, equally distributed in 4 classes. Each subject are pronouncing 17 different sentences. There are only one recording of a patient per class and there are the same number of recordings in each four classes. There are 6 women and 6 men per class. The recorded sentences used for those experiments consist of approximately 1 minute of speech per patient. 14 sentences are proposed for their place of articulation regarding the uttered phonemes: 2 for bilabial realisations, 6 for alveolars, 2 for labiodentals and 2 for palatals. There are three more sentences regarding the articulation mode: voiceless and voiced stops, voiceless and voiced fricatives and one sentence for voiced nasals.

Four classes corresponding to five stages of facial paralysis (see Table 1) are represented in the database part that is used here. Most of the time, patients with a facial paralysis never go back to stage I. So, all the 12 people labeled stage I were never diagnosed with a facial paralysis and are all healthy subjects. Their is no stage II facial paralysis patients according to the House and Brackmann scale in the database because usually, those patients do not feel the need to come for a consultation anymore. From the lips movement point of view, there is no difference between stage I patients and stage II patients. They are only different regarding if they were once affected by a facial paralysis or not. There are 12 people with stage III paralysis, 12 people with stage IV paralysis and 12 patients with stage V or VI. The difference between patients in stage V and VI does not concern the lips movements so their voices are then considered in the same class for the voice impairment classification.

3. Burst detection

Detection of burst impulses are usually used in speech recognition for accurately measuring acoustic features and for robust phonetic alignment. For this paper, the idea is to detect the burst for another purpose. Indeed, when people develops a facial paralysis, the accuracy in pronouncing plosives is altered. The hypothesis is that the presence of bursts decreases when the speech is altered and the peaks of energy are less marked and more dispersed. Another aspect of using features based on a burst detector is that the features are independent from the language of the database used.

In order to automatically process the bursts uttered by the various subjects, we use the burst detection method developed in [12]. The authors propose an intensity discrimination applied to bark-scale frequency bands. They combine the separate
frequency-band information into a single measurement using Baye’s rule. Then, they use a threshold to select a number of candidate frames for further spectral-domain artificial neural network classification. This burst detection method has a total error rate of 13.20% on the TIMIT corpus [13]. In our dataset, there are 84 bursts and the total error rate for the healthy subjects is 16%. There is a high insertion rate that is most of the time due to tongue clipping. Therefore, the aim here is more to see the evolution of the burst detection between the various stages than to observe the performance of the detection in itself. Table 2 summarizes the average number of bursts that were detected for each stages thanks to the burst detector.

Table 2: Average number of bursts in each class

<table>
<thead>
<tr>
<th>Stage</th>
<th>Stage III</th>
<th>Stage IV</th>
<th>Stage V-VI</th>
</tr>
</thead>
<tbody>
<tr>
<td>96.75</td>
<td>95.75</td>
<td>91.5</td>
<td>83.5</td>
</tr>
</tbody>
</table>

Table 2 shows that there is a correlation in the number of bursts that are detected and the various stages of the paralysis. There are more bursts detected in the higher stages than in the lower stages. This statistic let us think that features based on the burst detection could help a classification task on speech from patients with a facial paralysis.

The two figures 1 and 2 present the result of the burst detection for a healthy male patient and a stage VI male patient saying the sentence “Pierre et Paul préparent la purée de petits pois”. This sentence is one of the 17 sentences that were pronounced in our database and contains 9 bursts that should normally happen on phonemes /pl/, /t/ and /d/. For the healthy patient, 8 bursts are detected and only two for the stage VI patient. 

Thanks to this burst detector, various acoustic features have been computed. We choose the features by analysing the shape of the different bursts. The shape of the bursts in the higher stages seemed larger and lower in amplitude than in the healthiest stages. The idea is then to extract from each file some features that model those observations. In those experiments, the features that are processed are the number of bursts per second, the number of bursts that are greater in intensity than the average burst intensity per second, the number of bursts that are greater in intensity than the average burst intensity over the number of bursts, the number of bursts which have their standard deviation lower than the average standard deviation over the number of bursts, and the number of bursts which have their standard deviation lower than the average standard deviation per second. All those features are the inputs of two neural networks. The first neural network is based on a Multi-Layer Perceptron [14] and the second, on a Learning Vector Quantization method [15]. The experiments will be detailed in the next section.

4. Experiments

All the experiments have been carried out using a 10-fold cross validation process. In a 10-fold cross-validation, the original dataset is randomly partitioned into 10 equal size subsets. Among the 10 subsets created, a single subset is retained for testing the model, and the remaining 9 are used for training the model. The cross-validation process is then repeated 10 times, in order to test each one of the 10 subsets exactly once. For this paper, the dataset consists of 48 speech recordings, 12 in each four classes detailed in section 2).

Two neural networks are computed for the classification task. They are developed thanks to the WEKA open source machine learning software [16]. Those neural networks are based on a MLP and a LVQ algorithm. The difference between the two neural networks, in terms of classification accuracy, is mainly the ability of the LVQ to correctly classify the given recordings even when classes are close to each other. There are several versions of LVQ described in [15] which have slightly different properties. The one that is used in this paper is the LVQ2.1, which is efficient for fine tuning the decision boundaries between the competing classes. The number of neurons in the input layer is adjusted to the number of input parameters and there are two hidden layers for the MLP. The transfer function for the MLP is a sigmoidal function. The architecture of the LVQ network is composed by an input and a Kohonen layer, fully connected between them. The output layer of both structures has four neurons that can be activated depending on which stage is associated to the speech recording when testing the data.

4.1. Classification according to the Multilayer Perceptron network

In the results shown in Table 3, only 29% of the voices are correctly classified by the MLP neural network. Nevertheless, the important aspect to observe is the results in the classes that are close to the target class. For example, half of the stage I voices that are misclassified are labeled as stage III voices, which is the closest stage in terms of articulatory impairment (stage II voices are only different from stage I voices depending on if the subject had once the paralysis or not). 4 stage III voices are misclassified as stage IV and 2 are misclassified as
stage I. Furthermore, 6 stage IV voices are misclassified as stage III and 7 over 12 stage V-VI voices are misclassified as stage IV. In summary, 27 voices are misclassified in a class that is the closest neighbour to the target class.

According to those results, the burst-based features used as input for a multilayer perceptron network really seem to globally spot the articulatory impairment of the pathological speech.

4.2. Classification according to the Learning Vector Quantization network

As the various classes seem to have very close boundaries, the next experiment uses the LVQ neural network described earlier at the beginning of section 4. This neural network is indeed known to help the classification when the boundaries between them are fuzzy. The results are shown in table 4.

In this table, one can observe that 27 voices are correctly classified which correspond to a correct classification rate of 56%. The LVQ based neural network performs better than the MLP neural network. This method is more relevant for the speech impairment classification because the aim of the LVQ algorithm is not to have a perfect modelling of the classes but rather to minimize the classifications errors made on the training data set.

The same tendency of classifying the recordings in a class that is the closest to the target class can be observed as well with the LVQ network classification. 16 misclassified voices are classified in a class that is the closest to the target class. The burst-based features seem to allow a global modeling of the speech impairment that affects patients gradually on the four stages.

The results obtained in this paper demonstrate the correlation between the burst-based acoustic features and the stages of the House and Brackmann grading scale that uses only a physical description of the patient. Moreover, those burst-based features are independent from the language spoken by the patients and can be applied to other languages. In a future work, those burst-based features will be associated with other non-linear parameters such as phase-based features, pitch-based features, etc. and/or traditional Mel frequency cepstral coefficients in order to improve the classification task. Another perspective is also to perform an unsupervised clustering of the speech recordings in order to check if the stages of the House and Brackmann scale can be found in the final clustering.

5. Discussion and conclusion

In this paper, a new application in speech processing for vocal impairments is introduced. The database consisting of voices coming from people with peripheral facial paralysis is unique and composed of various types of sentences. The purpose of the study is to find acoustic features that are able to model the speech impairment that affects the patients and that can be correlated to the House and Brackmann grading scale used by speech therapists.

The experiments are carried out on isolated sentences and lie on original features that are computed thanks to a burst detector. It is interesting to note that, even if the classification using a basic MLP neural network shows only a 29% correct classification, there is an interesting tendency to classify the other recordings in a class that is close to the target class. Using a LVQ based neural network leads to better results with a correct classification rate of 56%. The same tendency of classifying the other recordings in a class that is close to the target class is still observed. This method is more relevant for speech impairment classification because the final aim which is also the philosophy of the LVQ algorithm is to minimize the misclassification rate in the training data.

6. References


