Automatic perceptual categorization of disordered connected speech

A. Alpan1, J. Schoentgen1,2, Y. Maryn3, F. Grenez1

1 Laboratory of Images, Signals & Telecommunication Devices, Université Libre de Bruxelles, 50, Avenue F. D. Roosevelt, 1050 Brussels, Belgium
2 National Fund for Scientific Research, Belgium
3 Department of Otorhinolaryngology and Head & Neck Surgery, Department of Speech-Language Pathology and Audiology, Sint-Jan General Hospital, Bruges, Belgium

aalpan@ulb.ac.be, jschoent@ulb.ac.be, Youri.Maryn@azbrugge.be, fgrenez@ulb.ac.be

Abstract

The objective of the presentation is to report experiments involving the automatic classification of disordered connected speech into binary (normal, pathological) or multiple (modal, moderately hoarse, severely hoarse) categories. The multi-category classification according to the perceived degree of hoarseness is considered to be clinically meaningful and desirable given that the reliable perceptual classification by humans of disordered voice stimuli is known to be difficult and time-consuming. The acoustic cues are temporal signal-to-dysperiodicity ratios as well as mel-frequency cepstral coefficients. The classifiers are support vector machines which have been trained and tested on two connected speech corpora. The binary classification accuracy has been high (98%) for both sets of acoustic cues. The multi-category classification accuracy has been 70% when based on signal-to-dysperiodicity ratios and 59% when based on mel-frequency cepstral coefficients.

Index Terms: classification of disordered connected speech, variogram analysis, signal-to-dysperiodicity ratio, mel-frequency cepstral coefficients, support vector machine.

1. Introduction

Within the context of the assessment of laryngeal function, acoustic analysis has a central place because the speech signal is known to be close to the frequency scale of the human ear and often used in automatic speech recognition. The acoustic samples are sustained phonations known to be close to the frequency scale of the human ear and often used in automatic speech recognition. The acoustic samples are sustained phonations (3 – 4 s long) and the first 12 seconds of the Kay Elemetrics Voice Disorder Database developed by the Massachusetts Eye and Ear Infirmary (MEEI) Voice and Speech Labs (Kay Elemetrics Corp., 1994). The acoustic samples are sustained phonations of vowel [a] (3 – 4 s long) and the first 12 seconds of the

results in cycle, harmonic or rhamonic insertion or omission errors that bias the values of dysperiodicity cues. The generalized variogram method that is used here enables tracking cycle-to-cycle dysperiodicities (whatever their cause) in any speech sound produced by any speaker, because it is not based on the assumptions that the signal is locally periodic or that the average cycle length can be known a priori [2]. The signal-to-dysperiodicity ratio (SDR) that summarizes the dysperiodicities has been shown to correlate strongly with the degree of perceived hoarseness. The automatic classification of pathological voice is a topic that has received considerable attention. Many studies have proposed a binary (normal/pathological) classification of voice samples [3]. An automatic categorisation according to perceived degrees of hoarseness appears, however, to be more attractive to both clinicians and technologists and more likely to be clinically relevant. Support Vector Machines (SVMs) have become a popular tool for discriminative classification [4]. SVM classification offers advantages, some of which are listed hereafter [5].

• There are no problems with local minima. One can construct highly nonlinear classifiers without worrying about getting stuck in local minima during training.

• There are few parameters to pick. For example, if one chooses to build a radial basis function (RBF) support vector machine for classification, one must fix two parameters: the penalty parameter for misclassificaion and the width of the gaussian kernel.

• The final results are stable, reproducible, and largely independent of the algorithm used to optimize the SVM model.

In this presentation, binary classification is first reported followed by multi-category classification. One corpus has been sub-divided into three categories with regard to the perceived degree of hoarseness. Also, classification results based on the signal-to-dysperiodicity (SDR) ratio are compared to results obtained with mel-frequency cepstral coefficients (MFCC) known to be close to the frequency scale of the human ear and often used in automatic speech recognition.

2. Databases

2.1. MEEI corpus

One corpus has been the Kay Elemeitics Voice Disorder Database developed by the Massachusetts Eye and Ear Infirmary (MEEI) Voice and Speech Labs (Kay Elemetrics Corp., 1994). The acoustic samples are sustained phonations of vowel [a] (3 – 4 s long) and the first 12 seconds of the
Rainbow Passage spoken by normophonic subjects and patients with organic, neurological, traumatic, and psychogenic voice disorders at different stages (from early to fully developed). The speech samples have been recorded in a controlled environment at 25 kHz and 16 bits of resolution. Hereafter, the analyses are carried out on the 12-second continuous Rainbow Passage utterances. We have considered a subset comprising 53 normal and 169 pathological voices omitting recordings devoid of a diagnosis and balancing samples with regard to sex and chronological age [6].

2.2. Dutch corpus
The corpus comprises the concatenation of two Dutch sentences with sustained vowel [a] (produced by the same speaker) uttered by 28 normophonic and 223 dysphonic speakers, male and female. The stimuli have been sampled at 44.1 kHz. Five judges have evaluated the stimuli perceptually. Each judge has rated the item “grade”, (G) of the GRABS scale, from 0 (normal) to 3 (severe). The “grade” refers to the overall perceived abnormality of the speech stimuli. The five perceptual scores per stimulus have been averaged. The Spearman correlation coefficients indicate moderate to high intra-rater (0.77 to 0.90) as well as fair to moderate two-by-two inter-rater agreement (0.51 to 0.73) [7]. A subset of this corpus has been divided in three classes. The first class (L) contains 56 files for which the average grade is 0 or 0.2. The second (M) contains 59 files for which the average grade is 1.2, 1.4 or 1.6, and the third one (H) contains 47 files for which the average grade is equal to or larger than 2.

3. Methods

3.1. Generalized variogram analysis
For a periodic signal \( x(n) \) of period \( T_0 \), one may write \( x(n) = x(n-T_0) \). For a locally-stationary signal, the deviation from strict periodicity over an analysis frame of length \( N \) can therefore be estimated by the following expression. Index \( n \) positions the samples within the frame.

\[
\delta = \min_{n} \left\{ \sum_{i}^{N} \left( x(n) - x(n+T) \right)^2 \right\} \quad \text{for} \quad -T_{\text{max}} < T < -T_{\text{min}}, \quad T_{\text{min}} < T < T_{\text{max}}
\]

The expression between accolades in (1) is known as the variogram of the speech signal. It involves the squared difference between a main analysis frame and a shifted auxiliary frame. For voiced sounds, lag \( T \) is expected to be negative and vice versa for a speech cycle positioned near the right-hand phonetic boundary.

For each main analysis frame position, lag \( T \) is fixed so as to minimize the cumulated squared difference between the main and shifted frames. For voiced sounds, lag \( T \) is therefore an integer multiple of the glottal cycle length. For unvoiced sounds, (1) can be still meaningfully computed but the interpretation of lag \( T \) in terms of glottal cycle lengths is not valid anymore [2].

In running speech, the signal amplitude evolves deterministically owing to onsets and offsets, sound-specific loudness as well as accentuation. To remove these clinically non-relevant variations of the signal amplitude, a local gain \( \alpha \) that equalizes the energies of main and auxiliary analysis frames is inserted into (1).

\[
\delta = \min_{n} \left\{ \sum_{i}^{N} \left( x(n) - x(n+T) \right)^2 \right\} \quad \text{for} \quad -T_{\text{max}} < T < -T_{\text{min}}, \quad T_{\text{min}} < T < T_{\text{max}}
\]

The analysis frame length is fixed to 2.5 ms, so that main and lagged frames cannot overlap. The shift between successive analysis frames is also fixed to 2.5 ms, thus enabling the sample-by-sample dysperiodicity to be computed unambiguously (3).

\[
e(n) = x(n) - x(n + T_{\text{opt}})
\]

In (3), lag \( T_{\text{opt}} \) is the positive or negative shift that minimizes the cumulated squared difference in (2). The energy-equaled variogram (2) has been called the generalized variogram to distinguish it from the conventional variogram in (1).

3.2. Multi-band segmental signal-to-dysperiodicity ratios
The segmental signal-to-dysperiodicity ratio \( \text{SDRSEG} \) consists in computing ratio (4) locally over intervals of 5 ms and then taking the average.

\[
\text{SDR}_{\text{seg}} = 10 \log \frac{\sum_{i} x_i^2}{\sum_{i} e_i^2}
\]

For each utterance, the speech signal as well as the corresponding dysperiodicity trace have been filtered by means of four-channel mel-spaced linear-phase filters. The ranges of the four mel bands (B1 – B4) have been (0 – 800 mel), (800 – 1600 mel), (1600 – 2400 mel) and 2400 mel and beyond. These mel-intervals correspond to the frequency bands (0 – 724 Hz), (724 – 2195 Hz), (2195 – 5188 Hz) and 5188 Hz and beyond. The filterbank was designed by means of the Parks-McClellan method. The segmental signal-to-dysperiodicity ratio has then been computed for each band.

3.3. Mel-Frequency Cepstral Coefficients (MFCC)
MFCC parameters are obtained via the Discrete Cosine Transform of the logarithm of the energy in several frequency bands (5).

\[
c_m = \sum_{i=0}^{B} \log(S_i) \cos(m(k-0.5)\frac{\pi}{B})
\]

whith \( m = (0:12) \) and \( B \) the number of bands in the mel scale; \( S_i \) is given by (6),

\[
S_i = \sum_{j=1}^{NFFT} W(j) X(j)
\]

and \( W(j) \) is the triangular weighting function associated with the \( k \)th mel band, and \( X(j) \) is the NFFT point magnitude spectrum \((j=1:NFFT)\) [8].

The MFCCs have been computed over 40ms Hamming-windowed frames with 50% overlap. The MFCCs averaged over all the frames of an utterance are used as acoustic features of that utterance.

3.4. Temporal finite differences
A representation reporting the dynamics of speech can be obtained by including the finite differences between
neighboring frames of the acoustic features \( p \). First (\( \Delta \)) and second (\( \Delta \Delta \)) differences (7) have been used.

\[
\Delta p(i) = \sum_{k=k}^{K} p(i+k) / (\sum_{k=k}^{K} k^2),
\]

where \( K = 4 \) and \( K = 1 \) for the first and second-order differences taking into account that (7) is applied twice to obtain the second-order difference. The root mean square of the differences over all the frames are used as acoustic features of the utterance.

### 3.5. Two-category SVM classifier

Support vector machines have been used for the automatic classification of normal/pathological voices [4]. In the linearly separable case, the SVM optimization algorithm maximizes the margin between the two classes (Figure 1).

![Figure 1: SVM classification: hyperplane maximizing the margin between the two classes.](image)

In linearly non-separable cases, a mapping of the input data to some higher-dimensional space, where the data are linearly separable, is carried out by means of a Kernel function \( K \) that has to satisfy certain properties (the Mercer conditions). The margin maximization algorithm leads to the following classifier, with \( x \) a data point to be assigned to one of two classes, according to the sign of function (8).

\[
f(x) = \sum_{i=1}^{N} \alpha_i y_i K(x_i, x) + b
\]

where \( y_i \in \{+1,-1\} \) are the class labels, \( b \) the bias term and

\[
\sum_{i=1}^{N} \alpha_i y_i = 0, \quad \alpha_i > 0.
\]

In this paper, a Radial Basis Function (RBF) (9) kernel has been used.

\[
K(x, y) = e^{-\gamma ||x-y||^2}, \gamma > 0
\]

The number \( N \) of RBF centers, the centers \( x_i \), the weights \( \alpha_i \), and the bias \( b \) are calculated automatically during SVM training via an optimization procedure. Prior to training, inverse kernel width \( \gamma \) and a penalty parameter \( C \) that is part of the cost function must be fixed. A larger \( C \) value corresponds to assigning a higher penalty to classification errors. A grid-search within intervals defined by the user is carried out to identify \((C, \gamma)\) pairs that enable the classifier to predict unknown data as accurately as possible. The LIBSVM software has been used for SVM training and classification [9].

### 3.6. Multi-category SVM classifier

The one-against-one method has been used for multi-category classification in the framework of which one classifier is constructed for every pair of different classes [9], [10]. The total number of binary classifications is \( K(K-1)/2 \) with \( K \) the number of categories. The final decision is made using a majority rule. For each binary classification, the vote of the category in which the unknown sample has been classified is incremented by one. The sample is assigned to the class with the largest vote [11].

### 3.7. Cross-validation

Results are obtained for a 6-fold cross-validation that is repeated 25 times. For each repetition, the corpus is split randomly in 6 subsets. One of the 6 subsets is used as the test set and the other 5 subsets are pooled to form the training set. This is carried out 6 times so that each subset is used once as test set. For each test and training set, the ratios of the number of samples belonging to different categories are the same as in the original data.

### 4. Results

#### 4.1. Binary classification: MEEI corpus

Tables 1 and 2 show the confusion matrixes for two sets of acoustic cues. The efficiency (correct detection and rejection rate), sensitivity (correct detection rate) and specificity (correct rejection rate) are also shown.

Table 1: Confusion matrix of the classifier fed with SDRSEG and the RMS values of the first and second finite differences of SDRSEG. The pair \((C, \gamma) = (10^3, 10^3)\) has given the best results.

<table>
<thead>
<tr>
<th>Observed</th>
<th>Predicted</th>
</tr>
</thead>
<tbody>
<tr>
<td>Pathological</td>
<td>Positive</td>
</tr>
<tr>
<td>Normal</td>
<td>98.2%</td>
</tr>
</tbody>
</table>

Table 2: Confusion matrix of the classifier fed with MFCCs and the energy and their first and second finite differences. The pair \((C, \gamma) = (10^3, 10^3)\) has given the best results.

<table>
<thead>
<tr>
<th>Observed</th>
<th>Predicted</th>
</tr>
</thead>
<tbody>
<tr>
<td>Pathological</td>
<td>Positive</td>
</tr>
<tr>
<td>Normal</td>
<td>98.7%</td>
</tr>
</tbody>
</table>

It can be seen that for both feature sets (SDRSEG and MFCCs & energy), the SVM classifier performs well. An accuracy of about 98% is obtained. However, only 12 features have been used in the first case (SDRSEG in 4 bands and finite differences) and 42 in the second case (MFCCs). With regard to the classification task that is investigated here, one observes that the choice of the \( C \) and \( \gamma \) values is not critical and that very large parameter intervals give rise to maximal classification accuracies. To decrease the number of cues, a principal component analysis has been carried out on the 12 dysperiodicity cues (4 SDRSEG & 1st and 2nd finite differences) and 42 in the second case (MFCCs). With regard to the classification task that is investigated here, one observes that the choice of the \( C \) and \( \gamma \) values is not critical and that very large parameter intervals give rise to maximal classification accuracies. To decrease the number of cues, a principal component analysis has been carried out on the 12 dysperiodicity cues (4 SDRSEG & 1st and 2nd finite differences). The first two principal components, which explain 85% of the total variance, have been used for classification and an accuracy of 97% has been obtained. When the SVM classifier is based on the SDRSEGs in the four frequency bands only (omitting the finite differences), the
accuracy is 93.5%. This suggests that the finite differences genuinely contribute to the classification and that the increase of classification accuracy is not only due to the increase in the number of features.

4.2. Multi-category classification: Dutch corpus

Tables 3 and 4 show the confusion matrices of the multi-category SVM-classifier for two sets of acoustic cues.

Table 3: Confusion matrix of the multi-category SVM-classifier fed with SDRSEG and the RMS values of the first and second finite differences of SDRboys. The pair \((C, γ) = (1, 1)\) has given the best results.

<table>
<thead>
<tr>
<th>Observed</th>
<th>Predicted</th>
</tr>
</thead>
<tbody>
<tr>
<td>L</td>
<td>M</td>
</tr>
<tr>
<td>82.0%</td>
<td>18.0%</td>
</tr>
<tr>
<td>M</td>
<td>21.9%</td>
</tr>
<tr>
<td>H</td>
<td>4.2%</td>
</tr>
</tbody>
</table>

Efficiency: 70.2%

Table 4: Confusion matrix of the multi-category SVM-classifier fed with MFCCs and energy and their first and second differences. The pair \((C, γ) = (1, 10^{-4})\) has given the best results.

<table>
<thead>
<tr>
<th>Observed</th>
<th>Predicted</th>
</tr>
</thead>
<tbody>
<tr>
<td>L</td>
<td>M</td>
</tr>
<tr>
<td>70.2%</td>
<td>27.5%</td>
</tr>
<tr>
<td>M</td>
<td>24.4%</td>
</tr>
<tr>
<td>H</td>
<td>14.5%</td>
</tr>
</tbody>
</table>

Efficiency: 59.2%

In Table 3, it can be seen that 82% of stimuli in class L (normal voices) are well classified. However, almost 40% of stimuli in class H (very hoarse voices) are assigned to class M (moderately hoarse voice). One also observes that if classes M and H are merged 86% of the stimuli belonging to that class are classified correctly. Therefore, the classifier appears to have difficulties discriminating between classes M and H. This may be related to the choice of classes. Indeed, classes M and H are closer in terms of perceived hoarseness than L and M. An accuracy of 59% (Table 4) is obtained when the SVM-classifier is founded on the MFCCs (& energy and first and second differences), which is considerably less than the accuracy obtained with dysperiodicity cues (Table 3).

Also, when the SVM-classifier is based on the first three principal components obtained for the SDRSegs (and 1st and 2nd differences), which explain 85% of the total variance, an accuracy of 67.3% is obtained, which is more the 65.7% obtained with the SDRSeg in the four frequency bands only, which agrees with the previous observation that the finite differences genuinely contribute to the classification results.

5. Discussion and conclusion

SVM-based binary and multi-category classifications have been carried out on connected speech. The SVM-classifier has been based on dysperiodicity cues which are the SDRSegs in four frequency bands. These cues have been compared to MFCCs and spectral energy. The first and second differences have also been considered.

For binary classification, the two categories of acoustic cues (temporal and cepstral) show similar results; 98% of accuracy has been obtained for each. In addition, when the SVM classifier is based only on the first two principal components obtained from SDRSegs (1st and finite differences), an almost as accurate classification of 97% has been observed. However, these results must be taken with a grain of salt. Indeed, it is known that some of the normal speakers in the MEEI database were recorded at sites and over channels different from the pathological speakers. Therefore, classification results published in the literature are often very high for that database.

A three category classification has therefore also been carried out on a subset of a Dutch corpus. This task is more challenging and more interesting from a clinical point of view compared to the normal/pathological classification. An SVM-classifier resting on dysperiodicity cues enables obtaining 70% of correct classification. This is higher than the 59% obtained with the MFCCs.

These results are preliminary. Further work will focus on the agreement of an automatic categorization into four classes with the perceptual classification performed by human experts.

6. Acknowledgements

This research was supported by the “Région Wallonne”, Belgium, in the framework of the “WALEO II” programme. We are grateful to Dr. Marc Remacle, Department of Otorhinolaryngology and Head and Neck Surgery, University Hospital of Louvain at Mont-Godinne, Belgium, for providing the Kay Emetricals Database.

7. References