

Combining temporal and cepstral features for the automatic perceptual categorization of disordered connected speech

A. Alpan¹, J. Schoentgen^{1,2}, 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
aalpan@ulb.ac.be, jschoent@ulb.ac.be, fgrenex@ulb.ac.be

Abstract

The objective of the presentation is to report experiments involving the automatic classification of disordered connected speech into multiple (modal, moderately hoarse, severely hoarse) categories. Support vector machines, used for the classification, have been fed with temporal signal-to-dysperiodicity ratios, the first rahmonic amplitude as well as mel-frequency cepstral coefficients. The signal-to-dysperiodicity ratio complements the first rahmonic amplitude when categorizing voice samples according to the degree of hoarseness yielding 77% of correct classification.

Index Terms: automatic perceptual categorization of disordered connected speech, variogram analysis, signal-to-dysperiodicity ratio, first amplitude rahmonic, mel-frequency cepstral coefficients, support vector machine.

1. Introduction

A large number of experiments have been devoted to the automatic detection and classification of voice pathologies, by means of automatic pattern recognition or statistical methods. Methods used in the fields of automatic speech recognition or speaker identification, and more generally, automatic pattern recognition have been applied to the classification of pathological voices [1].

Often the classifiers have been trained with acoustic features the clinical relevance of which is not obvious, such as, the mel-frequency cepstral coefficients (*MFCC*), for instance. They are cepstral features that have been widely used in automatic speech recognition or speaker identification [2].

Most studies have proposed a binary (normal/pathological) classification of voice samples. An automatic categorization according to perceived degrees of hoarseness appears, however, to be more attractive to both clinicians and technologists and more likely to be clinically relevant because they report the perceived timbre of a patient's voice, the reporting of which is part of his or her laryngeal assessment. Automatic perceptual categorization may offer advantages over human perceptual scoring, which is a consequence of the observation that the reliability of the scoring can be increased by averaging the scores assigned by several experts. However, the auditory-perceptual evaluation of patients' voice by several judges is time-consuming and costly. Automating that task is therefore desirable in a clinical framework.

The aim of this study is to apply a support vector machine classifier (SVM) [3], which has become a popular tool for discriminative classification, to automatic perceptual scoring of

disordered voices. Support vector machines rely on geometrically intuitive concepts and offer several advantages, which are the lack of problems with local minima, the small number of parameters that must be fixed, etc.

The objective of the presentation is the automatic perceptual categorization of hoarse connected speech. The combination of temporal variogram-based (*SDR*) and cepstral-based (*R1*) acoustic cues is investigated and compared to mel-frequency cepstral coefficients (*MFCC*). The purpose has been to test whether combining temporal and spectral cues enables increasing the accuracy of automatic perceptual scoring compared to categorization based on temporal cues only [4].

In addition, classification results have been compared to those obtained via linear regression.

2. Disordered speech corpus

The corpus comprises the concatenation of two Dutch sentences with a sustained vowel [a] uttered by 251 normal and dysphonic male and female dysphonic speakers. 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 GRBAS 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) [5].

A subset of this corpus has been divided in three classes with similar number of samples. The first class (*L*) contains 56 recordings for which the average grade is 0 or 0.2. The second (*M*) contains 58 recordings for which the average grade is 1.0, 1.2 or 1.4, and the third one (*H*) contains 47 recordings for which the average grade is equal to or larger than 2.

3. Acoustic features

3.1. Generalized variogram analysis

3.1.1. Definition

The generalized variogram (1) 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 [6].

For a locally-stationary signal $x(n)$, the deviation from strict periodicity over an analysis frame of length N can be estimated

by the following expression. Index n positions the samples within the frame.

$$\delta = \min_T \left\{ \sum_{n=0}^{N-1} [x(n) - \alpha x(n+T)]^2 \right\}, \quad \alpha = \frac{\sum_{n=0}^{N-1} x^2(n)}{\sum_{n=0}^{N-1} x^2(n+T)} \quad (1)$$

The expression between accolades in (1) is known as the variogram of the speech signal. Lag T may be positive or negative. In practice, it is permitted to vary between ± 2.5 and ± 20 ms. For each main analysis frame position, lag T is fixed so as to minimize the cumulated squared difference between the main and left and right-shifted frames. The absolute minimum is retained. Signed lags so guarantee that, in connected speech, only intra-segment cycles are compared. Indeed, when a vocal cycle is near the right-hand boundary of a speech sound the segment-internal cycles are expected to be to its left, that is, lag T that gives rise to the absolute minimum is expected to be negative and vice versa for a speech cycle positioned near the left-hand phonetic boundary.

The instantaneous value of the dysperiodicity $e(n)$ is estimated framewise, with T_{opt} equal to the lag that minimizes generalized variogram (1) for the current frame position.

$$e(n) = x(n) - \alpha x(n + T_{opt}) \quad (2)$$

3.1.2. Multi-band signal-to-dysperiodicity ratio

The segmental signal-to-dysperiodicity ratio (*SDRSEG*) consists in computing ratio (3) locally over intervals of 5 ms and then taking the average.

$$SDR_{local} = 10 \log \frac{\sum_{n=0}^{L-1} x^2(n)}{\sum_{n=0}^{L-1} e^2(n)} \quad (3)$$

Vocal dysperiodicities have been analyzed in different spectral intervals. For each utterance, the speech signal as well as the corresponding dysperiodicity trace have been filtered by means of three-channel mel-spaced linear-phase filters. The ranges of the three mel bands (B1 – B3) have been (0 – 800 mel), (800 – 1600 mel), (1600 – 2400 mel). These mel-intervals correspond to the frequency bands (0 – 724 Hz), (724 – 2195 Hz) and (2195 – 5188 Hz). The segmental signal-to-dysperiodicity ratio has then been computed for each band.

Table 1 shows the Pearson correlation coefficients between the average perceived grade G (hoarseness) and the segmental signal-to-dysperiodicity ratio (*SDRSEG*) in three frequency bands (B1 – B3) for the whole corpus (section 2).

Table 1: Pearson correlations between the average perceived grade G (hoarseness) and the segmental signal-to-dysperiodicity ratio (*SDRSEG*) in three frequency bands (B1 – B3) for the whole disordered speech corpus.

B1	B2	B3
-0.71	-0.59	-0.45

The correlation decreases as the frequency band increases. The best correlation has been obtained in the first frequency band. Also, when *SDRSEG*s in the three frequency bands are combined

in a stepwise linear regression, the correlation with perceptual scores increases to 0.76 [6].

3.2. Band-limited first rahmonic amplitude (R1)

Several studies have shown that the amplitude of the first rahmonic peak (*R1*) in the cepstrum is a broad descriptor of voice quality. Murphy has provided a theoretical description of the cepstral analysis of voiced speech containing aspiration noise, which suggests that *R1* is proportional to a geometric-mean harmonics-to-noise ratio [7]. He has also shown that limiting the number of harmonics in the spectrum prior to computing the cepstrum enables increasing the correlation between noise levels and *R1*.

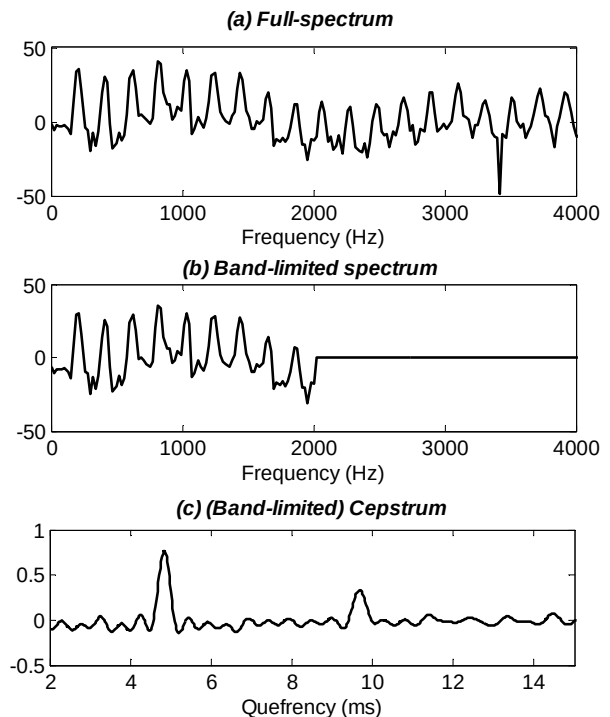


Figure 1: (a) Full spectrum, (b) (Offset removed) band-limited spectrum and (c) (Band-limited) cepstrum for an analysis frame of vowel [a].

The amplitude of first rahmonic *R1* is computed framewise.

1. The log-magnitude spectrum of the Hamming-windowed frame is calculated first (Figure 1.a).
2. The dB spectrum is limited to a fixed cut-off frequency or a fixed number of harmonics and the offset is removed (Figure 1.b).
3. The cepstrum is obtained via the inverse Fourier transform of the log-amplitude band-limited spectrum (Figure 1.c). The first rahmonic is located using a peak-picking algorithm, which selects the maximum in the queffrequency range corresponding to a frequency interval between 50Hz and 400Hz. No attempt has been made to check whether the position of the cepstral peak corresponds to the fundamental period.
4. A global *R1* amplitude is obtained by averaging *R1* amplitudes over all the windows.

For each utterance, an average fundamental period has been extracted with Praat [8] on the vowel [a] fragment.

In a previous study [9], a number of spectral analysis alternatives have been implemented, including period-synchronous (a fixed number of fundamental periods per frame), period-asynchronous, harmonic-synchronous (the spectrum is limited to a fixed number of harmonics) and harmonic-asynchronous band-limited analyses of speech.

Table 2 summarizes the Pearson correlation coefficients between average perceived grade G (hoarseness) and first harmonic amplitude ($R1$) for different analysis options. For period-asynchronous analyses, a window length of 2048 samples (46.4ms) has been used. For period-synchronous analyses, a window length containing 6 glottal cycles has been used. Also, the cut-off frequencies and the number of harmonics for the harmonic-asynchronous and -synchronous analyses that have obtained the best correlations are shown in Table 2 [9].

Table 2: Pearson's correlation coefficients between average grade scores of G and $R1$ for different analysis option.

Period-asynchronous (2048 samples)		Period-synchronous (6 cycles)	
Full-band $R1$	-0.61	Full-band $R1$	-0.63
Harmonic-asynchronous $R1$ (2000 Hz)	-0.73	Harmonic-asynchronous $R1$ (2000 Hz)	-0.73
Harmonic-synchronous $R1$ (6 harm)	-0.69	Harmonic-synchronous $R1$ (6 harm)	-0.76

The best correlation between acoustic cues and perceptual scores has been obtained for the period-synchronous (6 cycles) harmonic-synchronously (6 harmonics) band-limited $R1$.

3.3. Temporal finite differences

A representation reporting the dynamics of speech can be obtained by including the finite differences between neighboring frames. The root mean square of the differences over all the frames is used as an acoustic feature of the utterance.

4. Classification

4.1. Two-category SVM classifier

A support vector machine (SVM) is a supervised classifier solving, via a quadratic program, non-linear classification problems by transformation into a higher dimensional space. The idea is to find the optimal hyperplane that maximizes the margin between the two classes [2]. The hyperplane is fully specified by a subset of training samples, which are called the support vectors. The mapping of the non-linear low-dimensional decision surface to a higher dimension in which the two classes are separable by a hyperplane is carried out by means of the so-called "kernel trick".

4.2. Multi-category SVM classifier

The support vector machine classifier enables a binary classification. A direct solution of multi-class problems using a single SVM formulation is usually avoided [10]. A combination

of several binary SVM classifiers to solve a given multiclass problem is used instead.

The one-against-one method has been used, in the framework of which one classifier is constructed for every pair of different classes [11]. The total number of binary classifications then is $K(K-1)/2$ with K the number of categories. The final decision is based on a majority rule [12].

4.3. Cross-validation

Results are obtained for a 6-fold cross-validation. 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. The 6-fold cross validation scheme is repeated 10 times and the classification results are averaged for more accuracy.

4.4. Classification based on linear regression

For the classification based on linear regression, a 6-fold cross validation procedure has been adopted as well. In the training phase, linear regression coefficients have been determined by predicting the grade scores for each training set. In the test phase, the category assignment for an unknown sample is carried out via a linear combination of the acoustic features via the regression coefficients obtained in the training phase. If the (grade) score is lower than 1.0 then the unknown sample is assigned to class L. If the (grade) score is comprised between 1.0 and 2.0 then the unknown sample is assigned to class M, and if the score is equal to or larger than 2.0 then the unknown sample is assigned to class H.

5. Multi-category classification results

Table 3: Correct classification rates of the SVM-classifier fed with feature $SDRSEG$ s in the first three frequency bands, the first five $MFCC$ s and period-synchronous $R1$. The first line reports results obtained with the features (without finite differences). The second and third lines report results obtained when including the first and the first two finite differences.

	$SDRSEG$ (B1, B2, B3)	$R1$	$MFCC$	$SDRSEG$ (B1) & $R1$
Features	67.6	68.4	53.2	71.7
Features + Δ	71.4	72.3	53.9	74.7
Features + Δ + $\Delta\Delta$	73.5	72.6	56.4	77.4

The choice of acoustic features to feed the support vector machine classifier has been based on the correlation analyses between perceptual scores and temporal and cepstral acoustic cues (section 3.1 and 3.2). Therefore, the signal-to-dysperiodicity ratio in the first three frequency bands, the period-synchronous (6 cycles) harmonic-synchronous (6 harmonics) band-limited $R1$ are used. $SDRSEG$ and $R1$ are combined to test whether the combination enables increasing the accuracy of automatic perceptual scoring. In addition, the first five (including 0th) $MFCC$ s are used for comparison purposes ($MFCC$ s are averaged over all the windows).

Table 3 summarizes the correct classification rates in % (averaged over the 10 repetitions) for different sets of acoustic features.

Ratio *SDRSEG* in the first frequency band combined with period-synchronous rahmonic amplitude *R1* classifies the voice samples best. While, classification rates of 77% are obtained when the SVM-classifier is fed together with their first and second inter-frame differences, the use of ratio *SDRSEG* only and rahmonic *R1* obtains a rate of 72%. Results for features *SDRSEGs* are similar to those obtained with feature *R1* but largely exceed those obtained with cepstral coefficients *MFCC*.

Table 4 shows the average (over 10 repetitions) confusion matrix obtained when the SVM-classifier is fed by ratio *SDRSEG* in band 1 and rahmonic *R1* together with their first and second inter-frame differences.

Table 4: Confusion matrix obtained with *SDRSEG* in band 1 and period-synchronous *R1* together with their first and second finite differences.

Observed		Predicted		
		L	M	H
Observed	L	80.9%	19.1%	0.0%
	M	14.1%	79.1%	6.8%
	H	4.4%	24.5%	71.1%

The combination of these features has given the best overall classification rate of 77.4 %. It can be seen that 81% of the stimuli in class L (normal) are correctly classified. For class M (moderately hoarse), the correct classification rate (79%) is almost as high. The classification accuracy is the least for class H (very hoarse) (71%). Almost 25% of the stimuli in class H are assigned to class M (moderately hoarse). One also observes that if classes M and H are merged, 90% of the stimuli belonging to these classes are classified correctly. Also, the confusion between extreme classes (L and H) is negligible. Indeed, no modal voice sample (class L) is classified as very hoarse (class H) and only 2 very hoarse voice samples (4%) are classified as modal.

Classification results obtained with support vector machines have been compared to those obtained via linear regression.

One observes that classification rates obtained with the SVM-classifier are systematically larger (up to 10%). For instance, a correct classification rate of 68% has been obtained when combining *SDRSEG* in band 1 and rahmonic *R1* together with their first and second inter-frame differences.

6. Discussion and conclusion

SVM-based multi-category classifications have been carried out on connected speech. The SVM-classifier has been based on dysperiodicity cues which are the *SDRSEGs* in three frequency bands and first rahmonic amplitude *R1*. These cues have been compared to *MFCCs*. The first and second differences have also been included. An SVM-classifier resting on a combination of temporal and cepstral cues enables obtaining 77 % of correct classification. We may therefore conclude that temporal signal-to-dysperiodicity ratio and cepstral first rahmonic amplitude cues are genuinely complementary. Indeed, simulation experiments have shown that rahmonic amplitude *R1* mainly reports modulation noise and segmental *SDR* mainly reports additive noise [13].

Normal and moderately hoarse voices have been well classified. Correct classification for severely hoarse voices has been lower, because some have been classified as moderately hoarse. A confusion of extreme classes (modal vs severely hoarse) has been very rare.

Variogram-based and cepstrum-based features alone yield an accuracy of 73%, which is higher than the 56% accuracy obtained with the cepstral *MFCC* coefficients.

One also observes that accurate classification rates obtained with SVM-classifiers are up to 10 % larger than those obtained via linear regression.

7. Acknowledgements

This research has been supported by COST ACTION 2103 “Advanced Voice Function Assessment” in the framework of a short-term scientific mission at the University of Limerick with Prof. Peter Murphy, and by the “Région Wallonne”, Belgium, in the framework of the “WALEO II” programme. We are grateful to Dr. Youri Maryn who kindly provided the disordered speech corpus.

8. References

- [1] Godino-Llorente J. I., Gómez-Vilda P, Sáenz-Lechón N, Blanco-Velasco M, Cruz-Roldán F, and Ferrer-Ballester MA., “Support vector machines applied to the detection of voice disorders”, In: M. Faundez-Zanuy et al. (Eds.). Lecture Notes in Computer Science: Nonlinear Analyses and Algorithms for Speech Processing. Heidelberg: Springer Berlin, pp: 219-230, 2005.
- [2] Reynolds, D., “Experimental evaluation of features for robust speaker identification”, Speech and Audio Processing, IEEE Transactions, 2(4), pp. 639–643, 1994.
- [3] Vapnik, V., “An overview of statistical learning theory”, IEEE Transactions on Neural Networks, vol. 10, no. 5, pp. 988-1000, Sept.1999.
- [4] Alpan, A., Schoentgen, J., Maryn, M., Greniez, F., “Automatic perceptual categorization of disordered connected speech”, Proc. Interspeech, Makuhari (Japan), pp. 2574 – 2577, Sept 2010.
- [5] Maryn, Y., Corthals, P., Van Cauwenberge, P., Roy, N., and De Bodt, M., “Toward improved ecological validity in the acoustic measurement of overall voice quality: Combining continuous speech and sustained vowels”, J. of Voice., vol. 43(3), pp. 161-174, 2010.
- [6] Alpan, A., Maryn, Y., Kacha, A., Greniez, F., Schoentgen, J., “Multi-band dysperiodicity analyses of disordered connected speech”, Speech Com., Vol. 53(1), pp. 131-141, 2011.
- [7] Murphy, P. J., “On first rahmonic amplitude in the analysis of synthesized aperiodic voice signals,” J. Acoust. Soc. Am., vol. 120(5), pp. 2896–2907, 2006.
- [8] Boersma, P., Weenink, D., “Praat: doing phonetics by computer (Version 4.6.09) [Computer program]”, Retrieved June 26, 2007, from <http://www.praat.org/>
- [9] Alpan, A., Schoentgen, J., Maryn, Y., Greniez, F., Murphy, P., Assessment of disordered voice via the first rahmonic, Speech Com., Vol. 54(5), pp. 655-663, 2012.
- [10] Hsu, C.-W. and Lin, C.-J., “A comparison of methods for multi-class support vector machines”, IEEE Transactions on Neural Networks, vol. 13, no. 2, pp: 415-425, 2002.
- [11] Chang, C.-C. and Lin, C.-J., “LIBSVM : a library for support vector machines”, 2001. Software available at <http://www.csie.ntu.edu.tw/~cjlin/libsvm>
- [12] Friedman, J., “Another Approach to Polychotomous Classification”, Technical report, Department of Statistics, Stanford University, 1996.
- [13] Alpan, A., “Objective assessment of disordered connected speech”, PhD Thesis, Université Libre de Bruxelles, 2012.