# SPECTRAL ANALYSIS OF LARYNGEAL HIGH-SPEED VIDEOS: CASE STUDIES ON DIPLOPHONIC AND EUPHONIC PHONATION

P. Aichinger[1,2], I. Roesner[1], B. Schneider-Stickler[1], W. Bigenzahn[1], F. Feichter[1], A. K. Fuchs[2], M. Hagmüller[2], G. Kubin[2]

[1]Division of Phoniatrics-Logopedics, Department of Otorhinolaryngology, Medical University of Vienna, Austria
[2]Signal Processing and Speech Communication Laboratory, Graz University of Technology, Austria
philipp.aichinger@meduniwien.ac.at

*Abstract:* **Laryngeal high-speed videos (LHSVs) are analysed in order to provide an objective diagnostic criterion for the detection of diplophonia. Diplophonia is a significant subset of irregular phonation and characterized by the presence of two pitches in the voice sound. In current clinical practice diplophonia is detected by auditive perceptual rating, which should be assisted by objective analysis whenever the presence of diplophonia is doubtful. Ten cases of diplophonic and ten cases of euphonic phonation are analysed by means of spectral video analysis (SVA). The Non-unimodality measure $NUM$ is introduced and serves as a classification feature. Estimates for the sensitivity ($SE$) and specificity ($SP$) of the presented classification paradigm are $SE$: 90% (95%-$CI$: [55.5%, 99.7%]) and $SP$: 80% (95%$-CI$: [44.4%, 97.4%]). The estimators reflect promising results. Taking into account the confidence intervals ($CI$s), increasing the sample size must be considered. As a consequence from these results, LHSVs should be investigated in clinical studies more intensively in order to develop and establish solid interpretation guidelines for clinical issues.**

***Keywords: Laryngeal High-Speed Videos, Diplophonia, Diagnostic Study, Video Signal Processing, Voice Disorders***

## I. INTRODUCTION

Communication disorders may cost 154 to 186 billion dollars per year alone in the US, which equals 2.5 % to 3% of the US Gross National Product [1]. To decrease costs related to communication disorders, accurate and reliable voice assessment methods are needed [2]. Despite great efforts in conducting research on voice production, there is still a lack of such methods [3].

Diplophonia is a phenomenon in disordered voice, which is not well understood. It is characterized by the presence of two simultaneous pitches in the voice sound. Most commonly, diplophonia is understood as irregular phonation [4] or type 2 phonation [5], which does not provide an instrument for differential diagnosis (i.e., distinction between diplophonia and other kinds of voice disorders). The origins of the two pitches often remain hidden for clinicians, because standard stroboscopy does not allow for the investigation of double fundamental frequencies. As a consequence, the treatment decisions and treatment effect measurements are often difficult.

In contrast to stroboscopy, Laryngeal high-speed videos (LHSVs) provide interpretable data, even for irregular phonation. However, clinical interpretation of LHSVs is still difficult, due to the lack of clinical research. It is unclear, how auditive perceptual ratings (e.g., the presence of diplophonia) relate to objective LHSV measures, which will be investigated in this study.

Several cases of diplophonic and euphonic phonation are analysed by means of spectral video analysis (SVA). It will be shown, that the presence of diplophonia relates to the presence of spatially distributed secondary oscillation frequencies in the LHSV. We encourage the use of LHSVs, whenever the presence of diplophonia is unclear.

The paper is structured as follows: The methods section describes the experimental setup, the extraction of $NUM$ via SVA and the classification paradigm. The results section gives the classification results and individual SVA results for four representative subjects. The discussion and conclusion section concludes the paper.

## II. METHODS

Ten diplophonic and ten euphonic phonations are examined by means of LHSVs at a frame rate $fr = 4\,kHz$. The videos are recorded by a phoniatrician with a rigid endoscopic camera (Richard Wolf GmbH., HRES ENDOCAM 5562). The camera is inserted into the mouth of the subject way back to the pharynx. Simultaneously to the video, audio recordings are taken with an AKG HC 577 L microphone and a TASCAM DR-100 recorder. The HC 577 L is an omnidirectional head worn condenser microphone. The DR-100 is a handheld recording device that supplies the microphone with the required phantom power and records the microphone signal as an uncompressed wav-file at a sampling rate of 48 kHz and a resolution of 24 bits. The

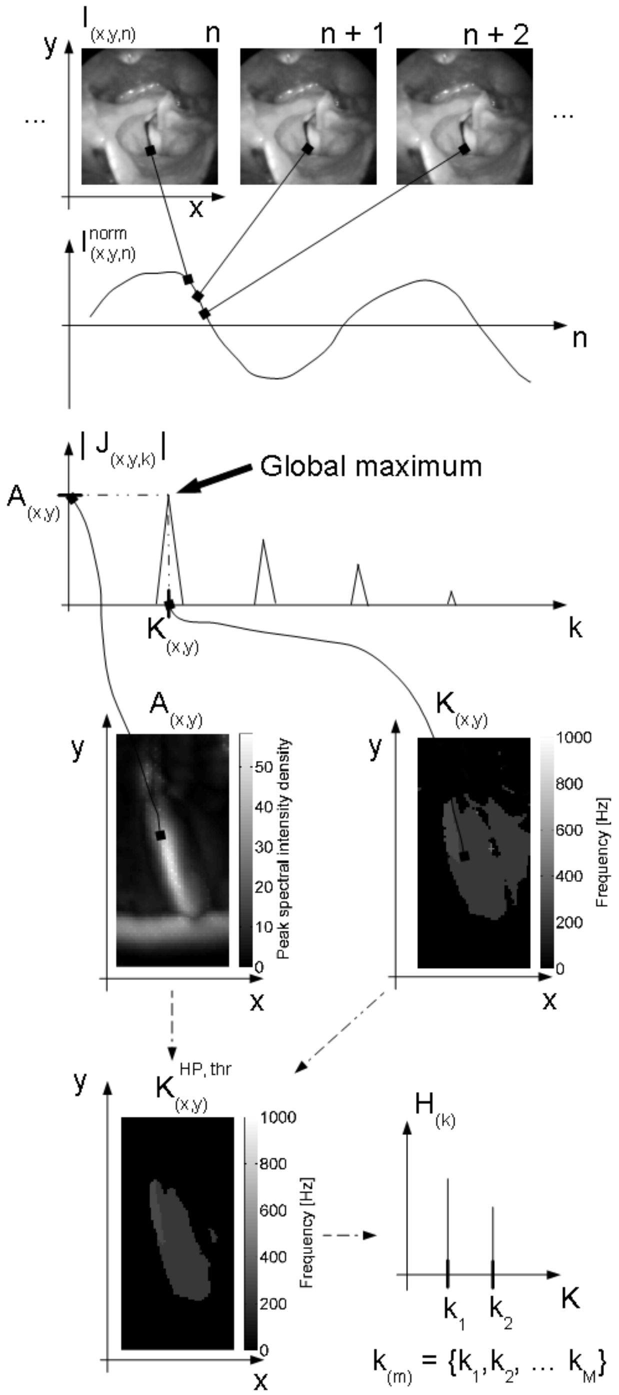

**Fig. 1: Spectral video analysis (SVA): The oscillation frequencies $k_{(m)}$ of the vocal folds are measured from the LHSV.**

audio records are manually synchronized to the videos by visual waveform matching.

For each phonation a stationary segment ($N = 500\,frames, 125\,ms\ at\ 4\,kHz$) is recorded and analysed be means of SVA [6]. Fig. 1 visualizes the methodology of the SVA.

The video is a 3-dimensional array of intensity values $I_{(x,y,n)}$, where $x$ is the lateral position, $y$ is the sagittal position and $n$ is the discrete time. The intensity time series are normalized pixel wise with respect to $n$.

$$I^{norm}_{(x,y,n)} = \frac{I_{(x,y,n)}}{\frac{1}{N} * \sum_{n=1}^{N} I_{(x,y,n)}} - 1 \qquad (1)$$

The normalized time series $I^{norm}_{(x,y,n)}$ are windowed with a Kaiser window ($\beta = 0.5$). The windowed time series are transformed to the frequency domain via discrete Fourier transformation, giving $J_{(x,y,k)}$. For each pixel, the frequency with the maximal intensity spectral density is chosen, resulting in the peak frequency image $K_{(x,y)}$.

$$K_{(x,y)} = \underset{k}{\operatorname{argmax}}\{|J_{(x,y,k)}|\} \qquad (2)$$

The peak intensity spectral density image $A_{(x,y)}$ relates the peak intensity spectral density to the $x/y$ coordinates.

$$A_{(x,y)} = \max_{k}\{|J_{(x,y,k)}|\} \qquad (3)$$

In order to suppress oscillations with irrelevant small amplitudes, the relevance threshold $thr$ is introduced. Additionally, low frequency components are removed.

$$K^{HP,thr}_{(x,y)} = \begin{cases} K_{(x,y)} \cdots K_{(x,y)} \geq 70\,Hz \cap A_{(x,y)} \geq thr \\ NaN \cdots K_{(x,y)} < 70\,Hz \cup A_{(x,y)} < thr \end{cases} \qquad (4)$$

The frequencies in $K^{HP,thr}_{(x,y)}$ are counted into the peak frequency histogram $H_{(K)}$. The distinct peaks in the histogram represent spatially distributed oscillation frequencies of the laryngeal tissue, which are referred to as “modes” in a SVA sense. $M$ is the number of modes, which is a function of $thr$.

To provide a scale measure as a diagnostic criterion, the Non-unimodality measure $NUM$ is introduced. $NUM$ is the minimal $thr$ for which $M = 1$. We hypothesize that $NUM$ can predict the presence of diplophonia, which will be tested.

$$NUM = min\{thr \in \mathbb{R}^{+} | M_{(thr)} = 1\} \qquad (5)$$

In order to generate a baseline classification for the presence of diplophonia, a listening test on the audio records has been conducted. An expert classified the audio segments played back on an AKG K 271 MK II headphone. The experiment was blinded and randomized.

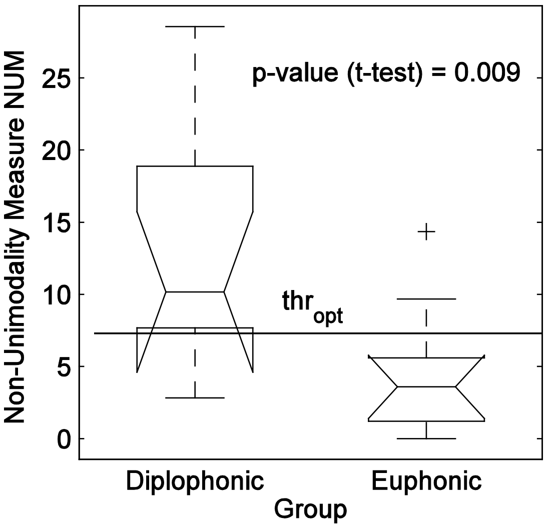

**Fig. 2: Boxplot: Non-unimodality measure $NUM$, healthy versus diplophonic phonation.**

Fig. 2 shows the boxplot of $NUM$ for groups diplophonic and euphonic. $NUM$ is higher in the diplophonic group, which indicates that it is a potential diagnostic criterion. The t-test (p = 0.009) reveals a significant difference of the means. However, the distributions do overlap, so it is impossible to realize perfect classification without false decisions. The optimal cut-off threshold $thr_{opt}$ is found via a receiver operating characteristic (ROC) curve. At $thr_{opt}$, the sensitivity ($SE$) and the specificity ($SP$) of the test are optimal.

Fig. 3 shows the ROC curve of predicting diplophonia by means of a simple cut-off threshold classifier. The ROC curve depicts the $SE$ (y-axis) and 1 - $SP$ (x-axis) with respect to the cut-off threshold. Choosing a low cut-off threshold (e.g., 1) makes all of the presented cases classified as diplophonic, whereas a high threshold (e.g., 19) makes all of the cases classified as euphonic. The optimal threshold $thr_{opt}$ is chosen as a trade-off between these two extremes of high $SE$ and high $SP$. It is found by minimizing the distance from the curve to the virtual optimum point (i.e., $SE$ = 1, $SP$ = 1, at the upper left corner), and settles at 7.

## III. Results

Table 1 illustrates the LHSV based classification of diplophonia, with the optimized cut-off threshold $thr_{opt}$. Cases with $NUM > 7$ are counted in the upper row (9 + 2 cases). Cases with $NUM < 7$ are counted in the lower row (1 + 8). The columns of the table represent the perceptual decision of the expert rater. Diplophonia is present in 9 + 1 cases, and absent in 2 + 8 cases. The presented table demonstrates high interrelation between the perceptual classification and $NUM$.

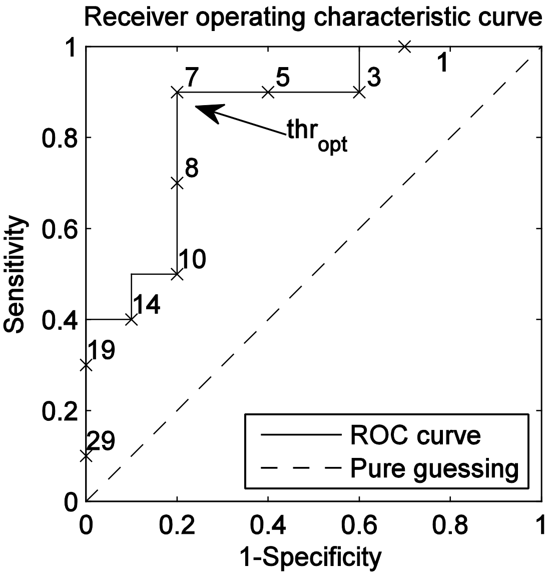

**Fig. 3: Non-unimodality measure $NUM$ as a predictor for diplophonia. The optimal cut-off threshold $thr_{opt}$ is found by minimizing the distance from the curve to the upper left corner.**

The $SE$ and $SP$ of the classifier are estimated as 90 % and 80 %. The $CI$s are found by using an iterative algorithm, based on the binomial distribution [7].

$$\begin{aligned} SE &= 9 \,/\, (9+1) = 90\,\% \\ CI(95\%) &= [55.5\%, 99.7\%] \end{aligned} \tag{6}$$

$$\begin{aligned} SP &= 8 \,/\, (8+2) = 80\,\% \\ CI(95\%) &= [44.4\%, 97.4\%] \end{aligned} \tag{7}$$

**Table 1: Classification table: Non-unimodality measure $NUM$ versus presence of diplophonia.**

| | | **Group:** | |
|---|---|---|---|
| | | Diplophonic | Euphonic |
| **Test result:** | $NUM > 7$ | 9 | 2 |
| | $NUM < 7$ | 1 | 8 |

Figs. 4 and 5 show the SVA of four representative cases. Fig. 4 shows the peak frequency images $K_{(x,y)}^{HP,thr_{opt}}$. In the true positive case, the major part oscillates at 224 Hz. The upper left part of the image shows a secondary oscillation at 288 Hz, which is non-unimodal in a SVA sense.

The true negative case shows one relevant oscillation frequency (120 Hz) along the entire glottal region, which is unimodal in a SVA sense. The false positive case with a primary frequency of 320 Hz shows secondary oscillations (648 Hz, 968 Hz) at the upper right area of

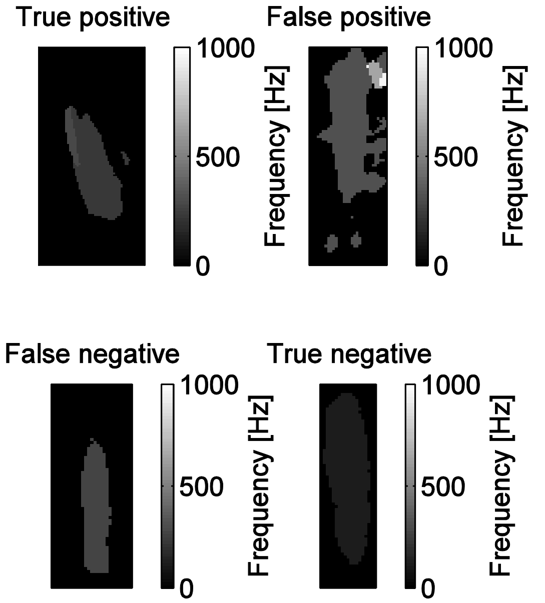

**Fig. 4: Peak frequency images $K_{(x,y)}^{HP,thr_{opt}}$of four representative cases.**

the image. After visual inspection of the video, these secondary oscillation frequencies were revealed as artifacts from mucus reflections. The false negative case shows a very small area with above threshold oscillation (6 pixels at 368 Hz). Visual inspection revealed a slow glottis drift (i.e., movement of the glottis along $x$ and $y$), resulting in spectral intensity density adversely spread. The false decisions of the presented cases are likely to be compensable in future versions of the analysis prototype by incorporating additional image processing techniques. Fig. 5 summarizes the observations from the peak frequency images into the peak frequency histograms $H_{(K)}$.

## IV. Discussion and Conclusion

In this study, an objective diagnostic criterion for the detection of diplophonia was tested on a case group of ten diplophonic and ten euphonic subjects. It was shown that the $NUM$ highly relates to the presence of diplophonia, as determined perceptually. The estimated values for $SE$ and $SP$ are promising, but hold large $CI$s because of the limited sample size. Thus, the analysis method must be further validated on a larger sample size before assisting the clinician in determining the presence of diplophonia.

The investigated method is more invasive than auditive perceptual expert ratings or objective acoustic methods (i.e., computer analysis of microphone signals). Nevertheless it is worthwhile to examine vocal fold oscillations by means of LHSVs. Compared to objective acoustic methods, LHSVs do not suffer from vocal tract resonances, which makes LHSV analyses in general more accurate and less error prone. Besides, LHSVs provide spatial information on laryngeal vibration, which is not contained in the audio signal. Compared to auditive perceptual ratings, it is likely that the analysis of LHSVs will achieve more objective results in terms of reliability and validity, which must be tested in more extensive studies.

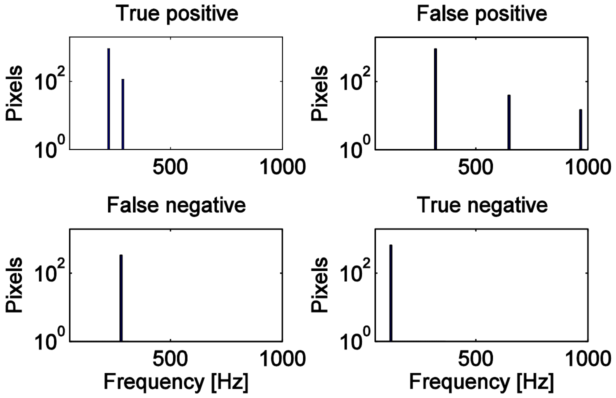

**Fig. 5: Peak frequency histograms $H_{(K)}$ of four representative cases.**

## References

[1] R.J. Ruben, “Redefining the survival of the Fittest: Communication Disorders in the 21st Century,” *The Larnygoscope*, vol. 110, pp. 241–245, 2000.

[2] M. Kob and P. Dejonckere, “Advanced Voice Function Assessment - Goals and Activities of COST Action 2103,” *Biomedical Signal Processing and Control*, vol. 4, no. 3, pp. 173–175, 2009.

[3] P. Aichinger, F. Feichter, B. Aichstill, W. Bigenzahn, and B. Schneider-Stickler, „Inter-device reliability of DSI measurement,” *Logopedics Phoniatrics Vocology*, vol 37, no. 4, pp. 167–173, 2012.

[4] D. Michaelis, M. Fröhlich, and H. Strube, “Selection and combination of acoustic features for the description of pathologic voices,” *Journal of the Acoustical Society of America*, vol. 103, no. 3, pp. 1628–1639, 1998.

[5] I. R. Titze, “Workshop on acoustic voice analysis: Summary statement,” National Center for Voice and Speech, Free books, pp. 1–36, 1995.

[6] S. Granqvist and P. Lindestad, “A method of applying Fourier analysis to high-speed laryngoscopy,” *The Journal of the Acoustical Society of America*, vol. 110, no. 6, pp. 3193–3197, 2001.

[7] M. Bland, “An Introduction to Medical Statistics,” Oxford University Press, 2000.