DOUBLE PITCH MARKS IN DIPLOPHONIC VOICE

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

[1]Department of Otorhinolaryngology, Division of Phoniatrics-Logopedics, Medical University of Vienna

[2]Signal Processing and Speech Communication Laboratory, Graz University of Technology

ABSTRACT

Determination of pitch marks (PMs) is necessary in clinical voice assessment for the measurement of fundamental frequency (F0) and perturbation. In voice with ambiguous F0, PM determination is crucial, and its validity needs special attention. The study at hand proposes a new approach for PM determination from Laryngeal High-Speed Videos (LHSVs), rather than from the audio signal. In this novel approach, double PMs are extracted from a diplophonic voice sample, in order to account for ambiguous F0s. The LHSVs are spectrally analyzed in order to extract dominant oscillation frequencies of the vocal folds. Unit pulse trains with these frequencies are created as PM trains and compensated for the phase shift. The PMs are compared to Praat's single audio PMs. It is shown that double PMs are needed in order to analyze diplophonic voice, because traditional single PMs do not explain its double-source characteristic.

Index Terms— Pitch marks, laryngeal high-speed videos, glottal area waveforms, diplophonia, audio and video signal processing

1. INTRODUCTION

Clinical voice assessment is necessary to indicate voice therapy and to document treatment effects. In clinical practice, subjective acoustic expert ratings are commonly used for voice assessment, which is recommended by the European Laryngological Society [1]. But in order to develop objective methods, there have recently been efforts to investigate principles of voice production for advanced voice assessment [2-4]. Still, there is a need for improvement.

Pitch measurement has been discussed recently [5-7]. PMs in voice and speech denote time instances of peak phonatory excitation. From a clinical point of view, PM determination serves as a basis of several methods of voice assessment. Measurements of F0 and perturbation are part of established voice assessment methods [8-9]. To measure F0 and perturbation, it is necessary to determine PMs.

Diplophonia is a frequent phenomenon in pathologic voice, associated with the perception of two simultaneous pitches [10]. A formerly proposed objective acoustic method for clinical voice assessment [11] generalizes all kinds of irregular voice (including diplophonia) into one category (i.e., irregularity). By doing so, a great number of different voice production phenomena are mixed up. As a consequence, the obtained information about laryngeal conditions is poor, and clinical conclusions are difficult. As a long-term goal, different kinds of irregular voice (i.e., diplophonia, vocal fry, rough voice) should be objectively distinguished, in order to provide better evidence in voice assessment.

From a signal processing point of view, diplophonic voice is a type 2 signal with ambiguous F0. Titze [12] postulates that: "Type 2 signals are signals with qualitative changes (bifurcations) in the analysis segment, or signals with subharmonic frequencies or modulating frequencies whose energies approach the energy of the fundamental frequency; there is therefore no obvious single fundamental frequency throughout the segment." Hence, the significance of traditional PMs extracted from diplophonic voice is limited. Nevertheless, there has been an approach of PM determination for diplophonic voice [13]. Praat [14] gives results for PMs in diplophonic voice. However, validity of PMs determination from diplophonic voice still needs special attention. In our work LHSVs are used for PMs determination, because this method does not suffer from vocal tract influences, as compared to PMs determination from audio signals.

It will be shown that in order to provide physically correct information on double pitch signals, it is necessary to extract double PMs rather than single PMs, which is new knowledge both in speech signal analysis and in clinical voice assessment. The double PMs approach explains beat frequency phenomena and ambiguity of F0, period length and pitch. To the authors' knowledge the study contributes the first proposal for extracting double PMs from diplophonic LHSVs. Basic research results for the validity of F0 and perturbation measurement of diplophonic voice are hereby established.

2. METHOD

This section describes the extraction of PMs. The new PMs extraction consists of data acquisition (video and audio), glottal area waveform (GAW) extraction, spectral video analysis, PM train generation and phase compensation via GAW fitting in the least-squares (LS) sense.

An endoscopic high-speed camera (HRES ENDOCAM 5562, Richard Wolf GmbH., $fs_0 = 4000$ *frames/s*) operated by a phoniatrician is used for data acquisition. The camera endoscope is inserted into the oral cavity of the subject, way back to the pharynx. Led through a flexible light conductor, a Xenon light (AUTO LP HIGH LIGHT 5132, Richard Wolf GmbH.) is used to illuminate the larynx. A Sennheiser microphone is fixedly mounted on the endoscope, in approximately 16 cm distance from its tip. While letting the proband phonate, a video of the vocal folds'

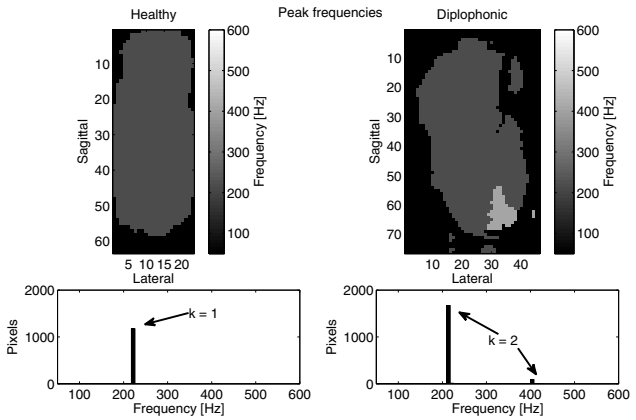

Fig. 1. Spectral video analysis

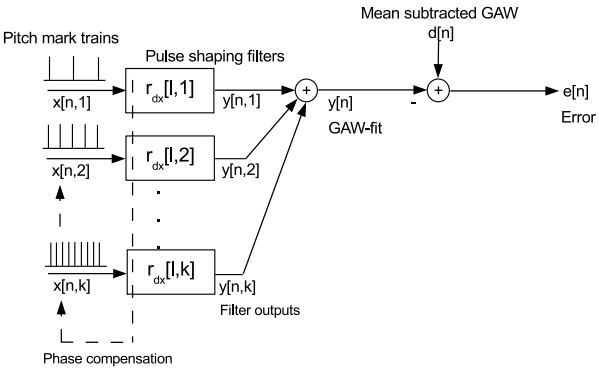

Fig. 2. GAW fitting in the least-squares (LS) sense

movement and an audio recording are taken. For this study, one healthy and one diplophonic subject are examined.

The glottal area is the 2D-projection of the glottal gap and contains relevant information about the vocal folds movement (i.e., opening and closing), when examined over time. While the vocal folds reflect irradiated light, the glottal gap does not; thus, the light intensity in the video frames is used for extracting the glottal area. In a simple threshold segmentation algorithm, the glottal area matrix $GAM[x,y,n_0]$ is determined.

$$GAM[x,y,n_0] = \begin{cases} 0, & i[x,y,n_0] > thr_1 \\ 1, & i[x,y,n_0] \le thr_1 \end{cases} \quad (1)$$

where x and y are the spatial indices, n_0 is the discrete time index, $i[x,y,n_0]$ is the light intensity and thr_1 is the segmentation threshold. Summing $GAM[x,y,n_0]$ over x and y yields the glottal area waveform in time $GAW[n_0]$.

$$GAW[n_0] = \sum_x \sum_y GAM[x,y,n_0] \quad (2)$$

$GAW[n_0]$ is upsampled from $fs_0 = 4\ kHz$ to $fs = 200\ kHz$, resulting in $GAW[n]$.

The vocal folds' oscillation frequencies are extracted from the video by spectral video analysis [15]. Granqvist and Lindestad proposed Fourier analysis on video signals with respect to time, but did not combine the method with PMs determination. The DFT of $i[x,y,n_0]$ is calculated with respect to the time index n_0, on a rectangular window of size of $M = 506\ samples\ @\ 4\ kHz\ (126.5\ ms)$. The window size is chosen in order to capture 3 meta cycles (see below) of the diplophonic voice sample. The DFTs of pixel wise intensity time series $i[x,y,n_0]$ result in pixel wise spectra $I[x,y,f]$, with a frequency resolution of 7.9 Hz. For each x and y, the frequency with the highest amplitude is picked. This gives the peak frequency matrix $PFM[x,y]$. Relevant frequency peaks are considered to lie between 50 Hz and 600 Hz, with an amplitude greater than $thr_2 = 800$ (i.e., a manually chosen relevance threshold). The values of the peak frequency matrix are summarized in a peak frequency histogram.

$$I[x,y,f] = \left| DFT_{n_0 \to f}\{i[x,y,n_0]\} \right| \quad (3)$$

$$PFM[x,y] = \begin{cases} \underset{50 \le f \le 600}{\operatorname{argmax}}\{I[x,y,f]\}, I[x,y,f] \ge thr_2 \\ undefined, \quad I[x,y,f] < thr_2 \end{cases} \quad (4)$$

Figure 1 shows the peak frequency matrices of healthy and diplophonic phonation as spatial images and the peak frequency histograms. The spectral video analysis results in k above threshold peak frequencies f_j with period lengths N_j. In the healthy phonation example, a large area is dominated by the 221.8 Hz peak. Some peripheral pixels do not show above threshold oscillation. Those areas in the images are black. In diplophonic phonation, there is a small region at the bottom right of the image, where the 404 Hz peak dominates. This region corresponds to the left vocal fold, anterior (front) position of the subject. The rest of the area is either dominated by the 213.9 Hz peak, or does not contain above threshold oscillation.

The peak frequency histograms summarize the frequency matrices. In healthy phonation the peak frequency distribution is unimodal ($k = 1$). In diplophonic phonation, the peak frequency distribution is bimodal ($k = 2$). At the present stage of this work the decision about modality is subjective. Future work will incorporate a statistical test to automatically determine the modality of the vocal folds' oscillation.

The extracted peak frequencies f_j serve as input for the PM train generation. The uncompensated phase PM trains $x_u[n,j]$ are unit pulse trains with period lengths N_j.

$$N_j = \left\lfloor \frac{fs}{f_j} \right\rceil, \quad where\ j = 1,2,\dots,k \quad (5)$$

$$x_u[n,j] = \sum_m \delta[n - m * N_j], \quad m \in \mathbb{Z} \quad (6)$$

Figure 2 shows the block diagram of the GAW fitting in the LS sense. The desired signal $d[n]$ is the mean subtracted $GAW[n]$.

$$d[n] = GAW[n] - \frac{1}{M}\sum_{n=1}^{M} GAW[n] \quad (7)$$

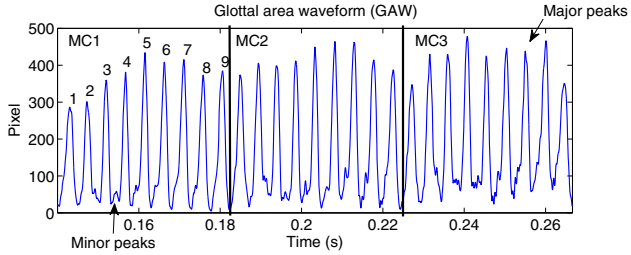

Fig. 3. GAW of diplophonic phonation

The FIR-filters with coefficients $r_{dxu}[l_j,j]$ are designed from cross correlating $x_u[n,j]$ and $d[n]$. l_j is the filter coefficient index.

$$r_{dxu}[l_j,j] = \sum_n x_u[n,j] * d[n+l_j] \quad with \quad l_j = 1-\frac{N_j}{2}, 2-\frac{N_j}{2}, \dots, -1,0,1, \dots, \frac{N_j}{2}-2, \frac{N_j}{2}-1 \tag{8}$$

Given the uncompensated phase filter coefficients r_{dxu}, the phase shift is calculated. The filter coefficient vector is shifted so that the maximal filter coefficient is in the middle of the vector, i.e., at time lag $l_j = 0$ *s*.

$$\Delta\varphi[j] = \underset{l_j}{\operatorname{argmax}}\{r_{dxu}[l_j,j]\} \tag{9}$$

$$x[n,j] = \sum_m \delta\left[n - m * N_j - \Delta\varphi[j]\right], \quad m \in \mathbb{Z} \tag{10}$$

The final FIR-coefficients for the fixed phase pulse shaping filters $r_{dx}[l_j,j]$ are calculated by cross correlating $x[n,j]$ and $d[n]$. In order to perform pulse shape filtering, the filter output signals $y[n,j]$ are computed by convolving $x[n,j]$ with $r_{dx}[l_j,j]$. The filter output signals $y[n,j]$ are summed up to $y[n]$, which results in the optimal fit of the $GAW[n]$ in a LS sense.

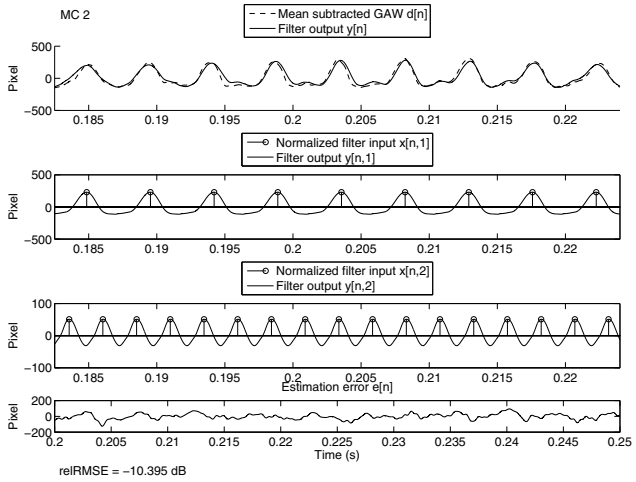

Fig. 4. GAW fitting summary, diplophonic phonation, meta cycle 2

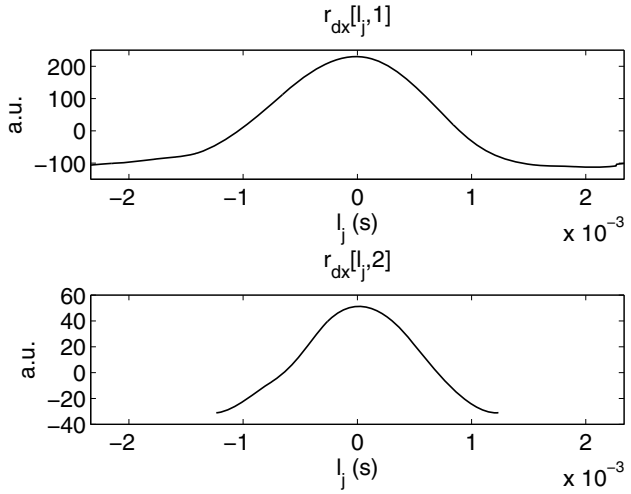

Fig. 5. Pulse shape filtering coefficients, diplophonic phonation

$$r_{dx}[l_j,j] = \sum_n x[n,j] * d[n+l_j] \tag{11}$$

$$y[n,j] = \sum_{l_j} x[n,j] * r_{dx}[n-l_j,j] \tag{12}$$

$$y[n] = \sum_{j=1}^{k} y[n,j] \tag{13}$$

The error signal $e[n]$ is obtained by subtracting $y[n]$ from $d[n]$. As a quality-of-fit criterion, the relative root mean square error *relRMSE(dB)* is calculated. More negative log-values mean better fits.

$$e[n] = d[n] - y[n] \tag{14}$$

$$relRMSE(dB) = 20 * log_{10} \frac{\sqrt{\frac{1}{M}\sum_{n=1}^{M} e[n]^2}}{\sqrt{\frac{1}{M}\sum_{n=1}^{M} d[n]^2}} \tag{15}$$

The description of the relative root mean square error concludes the description of our approach for PMs determination.

In order to compare the results of our method to the results obtained from Praat, the audio signals and videos of one healthy and one diplophonic phonation are analyzed. Praat needs input of the audio data, whereas the novel approach works with video files.

3. RESULTS

This section shows the analysis results of the new approach for PMs determination for healthy phonation and for disordered voice, compared to PMs obtained from Praat's audio analysis. Figure 3 shows the GAW from diplophonic phonation. There are several major peaks in the GAW, corresponding to the main glottal cycle. The signal shows a periodic beat phenomenon, resulting in times where peak heights are minimal (e.g., at 0.185 s) and times where peak heights are maximal (e.g., at 0.205 s). The

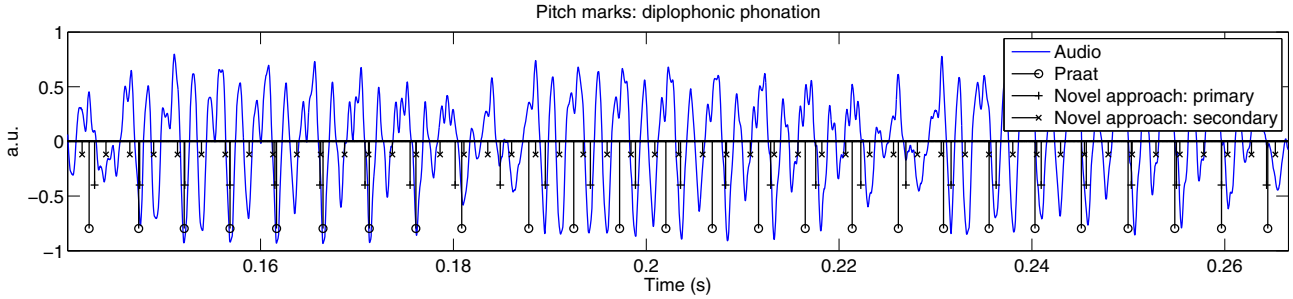

Fig. 7. Audio and pitch marks: diplophonic phonation

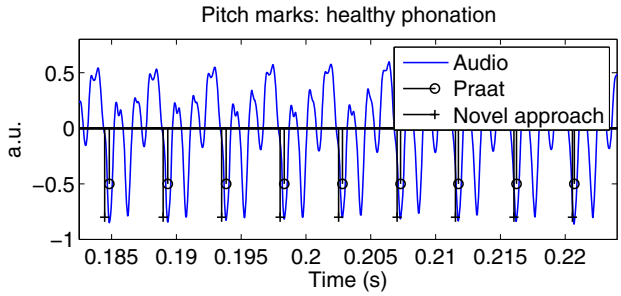

Fig. 6. Audio and pitch marks: healthy phonation

GAW is segmented into 3 meta cycles (MCs, i.e., the shortest time interval in which the two intrinsic frequencies are commensurable or the least common multiple of the two intrinsic periods).

The limits of the MCs are set to the minima in GAW peak height. Between the major peaks, there are several minor peaks, belonging to the secondary glottal oscillation. In times where the primary and secondary oscillations are out of phase, there are minimal peak heights of the GAW (i.e., beat frequency phenomenon). Determination of the PMs shows that one MC consists of 17 secondary oscillation cycles, compared to 9 primary oscillations. This corresponds to the frequency ratio of the frequencies measured in the spectral video analysis (213.9 Hz/404 Hz = 0.5295, 9/17 = 0.5294).

Figure 4 shows the GAW fitting summary for diplophonic phonation, according to the block diagram in figure 2. The desired signal $d[n]$ (mean subtracted $GAW[n]$) and the filter output $y[n]$ are shown in subplot 1. $y[n]$ represents the LS fit of the mean subtracted GAW. Subplots 2 and 3 show the filter input signals $x[n,j]$ (i.e., the PMs) and the filter output signals $y[n,j]$, being the summands of the GAW fit. The filter output signals are achieved by convolving $x[n,j]$ with the pulse shaping filter coefficients $r_{dx}[l_j,j]$ (see figure 5). Subplot 4 shows the fitting error $e[n]$, with $relRMSE$ of -10.40 dB.

Figure 5 shows the pulse shaping filter coefficients. In subplot 1, the filter coefficients $r_{dx}[l_1,1]$ show the pulse shape of the average primary GAW pulse, i.e., signal components correlated with $x[n,1]$. The shape looks like a typical GAW pulse of modal phonation. Subplot 2 shows the pulse shaping filter coefficients of the average secondary GAW pulse $r_{dx}[l_2,2]$. This secondary contribution to the GAW-fit is smaller and looks like a sinus, rather than a typical GAW pulse. The shape comes from the anterior left (bottom right in the peak frequency matrix in figure 1) oscillation of the vocal folds at 404 Hz.

Figures 6 and 7 show the audio signals and the PMs of healthy and diplophonic phonation, determined with Praat versus the novel approach. The delay between the audio and the video is compensated. Figure 6 shows the results from healthy phonation, with the PMs' positions determined via our approach compared to Praat's PMs. The results do qualitatively fit, with the PMs at the negative sound pressure peak of each period. Figure 7 shows the audio signal from diplophonic phonation, with the PMs' positions determined via the novel approach compared to Praat's PMs. It is shown that Praat's method is inappropriate for the analysis of a double pitch signal, because Praat assumes a single pitch track. Thus, Praat loses sync to the audio signal in MC 2. The PMs obtained with our approach show consistent patterns in each MC, i.e., out-of-phase PMs at MC borders, and in-phase PMs at MC centers. There is no established ground truth for double pitch mark analysis of diplophonic voice, which is herewith originally proposed.

4. CONCLUSIONS

The GAW of diplophonic phonation has been modeled with a 2 source model ($k = 2$). The fitting error ($relRMSE = -10.40$ dB) is slightly greater than the fitting error of healthy phonation ($relRMSE = -12.27$ dB), even with double source fitting. Thus, it is hypothesized that fitting the GAW of diplophonic phonations is more complex than fitting healthy phonation, taking into account the number of sources.

The primary pulse shape filter coefficients of diplophonic phonation show intervals at the filter edges where the pulse is relatively flat, as well as a positive pulse in its center (representing the closed and open phase). In contrast the secondary pulse shape filter coefficients do not have a clear closed phase. The vocal folds do not fully close at the anterior part, and so the vocal fold is moving sinusoidally. The comparison of traditional PM determination and our method confirms that the validity of the traditional method used on type 2 phonation is not given. On the other hand, the double PMs extracted with our approach represent the two pitches of the voice. The new double PM approach explains the periodic peak height fluctuations (beat frequency) and the minor peaks in the GAW. The minor peaks must be extracted from the video, because they are dispersed in the vocal tract and not visible in the audio signal. Concluding, we suggest validating audio based PM determination methods with our new LHSV based approach. Future work will investigate if our approach can be used as a front end for the automatic detection of diplophonic voice, more patients will be examined and analyzed.

5. ACKNOWLEDGEMENT

The authors thank Jean Schoentgen for a fruitful discussion and comments on the manuscript, as well as Richard Wolf GmbH. for providing the HRES ENDOCAM 5562.

6. REFERENCES

[1] P. Dejonckere, P. Bradley, P. Clemente, G. Cornut, L. Crevier-Buchman, G. Friedrich, P. Van De Heyning, et al., "A basic protocol for functional assessment of voice pathology, especially for investigating the efficacy of (phonosurgical) treatments and evaluating new assessment techniques," European Archives of Oto-Rhino-Laryngology, vol. 258(2), pp. 77-82, 2001.

[2] D. D. Deliyski, "Acoustic model and evaluation of pathological voice production," Third European Conference on Speech Communication and Technology, pp. 1969-1972, 1993.

[3] M. Kob, P. Dejonckere, "Advanced Voice Function Assessment - Goals and activities of COST Action 2103," Biomedical Signal Processing and Control, vol. 4(3), pp. 173-175, 2009.

[4] P. Aichinger, F. Feichter, B. Aichstill, W. Bigenzahn, B. Schneider-Stickler, "Inter-device reliability of DSI measurement," Logopedics Phoniatrics Vocology, vol. 37(4), pp. 167-173, 2012.

[5] L. N. Tan, A. Alwan, “Noise-robust F0 estimation using SNR-weighted summary correlograms from multi-band comb filters,” in Proc. IEEE Int. Conf. on Acoustics, Speech, and Signal Processing (ICASSP), pp. 4464-4467, 2011.

[6] R. Kumaresan, V. K. Peddinti, P. Cariani, “Multiple pitch identification using cochlear-like frequency capture and harmonic grouping,” in Proc. IEEE Int. Conf. on Acoustics, Speech, and Signal Processing (ICASSP), pp. 613-616, 2011.

[7] F. Huang, T. Lee, “Sparsity-based confidence measure for pitch estimation in noisy speech,” in Proc. IEEE Int. Conf. on Acoustics, Speech, and Signal Processing (ICASSP), pp. 4601-4604, 2012.

[8] I. Titze, "Acoustic interpretation of the voice range profile (phonetogram)," Journal of Speech and Hearing Research, vol. 35(1), pp. 21-34, 1992.

[9] F. Wuyts, M. de Bodt, "The dysphonia severity index: an objective measure of vocal quality based on a multiparameter approach," Journal of Speech, Language and Hearing Research, vol. 43, pp. 796-809, 2000.

[10] P. Dejonckere, J. Lebacq, "An analysis of the diplophonia phenomenon," Speech Communication, vol. 2, pp. 47-56, 1983.

[11] D. Michaelis, "Selection and combination of acoustic features for description of pathologic voices," Journal of the Acoustical Society of America, vol. 103(3), pp. 1628-1639, 1998.

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

[13] M. Hagmüller, G. Kubin, "Poincaré pitch marks," Speech Communication, vol. 48(12), pp. 1650-1665, 2006.

[14] P. Boersma, D. Weenink, "Praat: doing phonetics by computer," [Computer program], Version 5.3.34, retrieved 21 November 2012 from http://www.praat.org/

[15] S. Granqvist, P. Lindestad, "A method of applying Fourier analysis to high-speed laryngoscopy," The Journal of the Acoustical Society of America, vol. 110(6), pp. 3193-3197, 2001.