Diplophonie - Definitionen, Modelle und Detektion

Philipp Aichinger[1], Anna Katharina Fuchs[2]

[1] *Klinische Abteilung Phoniatrie-Logopädie, Universitätsklinik für Hals-, Nasen- und Ohrenkrankheiten, Medizinische Universität Wien, 1090 Wien, E-Mail: philipp.aichinger@meduniwien.ac.at*

[2] *Institut für Signalverarbeitung und Sprachkommunikation, Technische Universität Graz, 8010 Graz, E-Mail: anna.fuchs@tugraz.at*

Einleitung

Stimmerkrankungen (Dysphonien) können zu verminderter Lebensqualität, Arbeitsunfähigkeit oder sozialer Isolation führen. Um diese Problematik zu adressieren, werden valide Deskriptoren der Stimmqualität benötigt, welche die Indikation, Selektion, Evaluierung und Optimierung medizinischer Behandlungen unterstützen.

Das Hauptsymptom der Dysphonie ist die Heiserkeit, die sich im veränderten Stimmklang manifestiert. Klinisch werden Abweichungen vom normalen Stimmklang üblicherweise auditiv durch PhoniaterInnen, d.h. Hals-Nasen-Ohren-ÄrztInnen, beschrieben. Hierfür wird häufig die RBH-Skala verwendet, welche die Beschreibung von Rauigkeit, Behauchtheit und Heiserkeit ermöglicht [1]. Normaler Stimmklang wird mit 0 bewertet, geringgradige Abweichung vom normalen Stimmklang mit 1, mittelgradige Abweichung mit 2, und hochgradige Abweichung mit 3.

Obwohl akustische Analysen seit mehreren Jahrzehnten zur Beschreibung pathologischer Stimmklänge eingesetzt werden, sind deren Validität und physiologische Interpretierbarkeit oft limitiert. Insbesondere versagen die Verfahren wenn zugrundeliegende Signalmodelle falsch sind, was bei mittel- und hochgradiger Heiserkeit häufig vorkommt. Dadurch steigt das Risiko klinischer Fehlinterpretationen gerade wenn präzise Beschreibungen benötigt werden. Dies indiziert einen dringenden Handlungsbedarf an der Entwicklung neuer Verfahren.

Diplophonie ist ein Heiserkeitstyp bei dem zwei simultane Tonhöhen im Stimmklang auftreten [2], [3]. Je nach Ursache werden diplophone PatientInnen logopädisch therapiert oder einem chirurgischen Eingriff unterzogen. Die auditive Detektion von Heiserkeit und Diplophonie in der klinischen Praxis ist aus Sicht der evidenzbasierten Medizin und der wissenschaftlichen Methodik problematisch. Eine automatische Methode zur Detektion von Diplophonie wird beschrieben.

Material und Methoden

Es wurden Stimmlippen-Hochgeschwindigkeitsvideos mit simultanen Audio-Aufnahmen von 120 ProbandInnen gemacht. 80 der ProbandInnen waren dysphone PatientInnen, von denen 40 diplophon waren. Die 80 PatientInnen wurden aus dem Ambulanzbetrieb der Klinischen Abteilung Phoniatrie-Logopädie, Universitätsklinik für Hals-, Nasen- und Ohrenkrankheiten der Medizinischen Universität Wien rekrutiert. Die 40 nicht-dysphonen ProbandInnen wurden über öffentlichen Aushang im Allgemeinen Krankenhaus der Stadt Wien rekrutiert.

Die Hochgeschwindigkeitsvideos wurden mit einer HRES ENDOCAM 5562 (Richard Wolf GmbH) aufgenommen. Die ProbandInnen wurde angeleitet ein [i] zu phonieren, damit der Kehldeckel nach vorne kippt und die Sicht auf die Stimmlippen freigibt. Wegen dem starren Endoskop, welches durch den Mund bis zum Rachen eingeführt wurde, und der dadurch nach unten verlagerten Zunge, waren die produzierten Stimmklänge schwa-ähnlich.

Ein Kopfbügel-Mikrophon AKG HC 577 L mit Nierencharakteristik wurde mit dem Poppschutz AKG W77 MP und der Originalkappe ohne Höhenanhebung verwendet. Es wurde über einen Phantomspeiseadapter AKG MPA V L (lineare Einstellung) an einen portablen Rekorder TASCAM DR-100 angeschlossen. Die Quantisierungsauflösung war 24 bit und die Abtastrate 48 kHz.

Die Stimmqualität der Audiosignale wurde mit dem Programm Praat [4] annotiert. 125 dysphone Signale mit homogener Stimmqualität wurden zur Analyse ausgewählt und waren zwischen 125.4 und 2048 ms lang. 55 der Stimmsignale waren diplophon und 70 waren nicht-diplophon.

Signalmodell

Es wird angenommen, dass diplophone Signale sich aus zwei harmonischen Oszillatoren und Rauschen zusammensetzen. Die Wellenformen $d_m(n)$ der Oszillatoren mit Index m sind Fourier-Reihen mit P Teiltönen. $a_{m,p}$ und $b_{m,p}$ sind die Fourier Koeffizienten, ω_m die Kreisfrequenzen, p der Teiltonindex und n der Zeitindex. Das akustische Signal $d'(n) = \sum_{m=1}^{M} d_m(n) + \eta(n)$, wobei $\eta(n)$ Rauschen ist, und M die Anzahl der Oszillatoren, d.h. $M = 2$ für Diplophonie.

$$d_m(n) = \sum_{p=1}^{P} \begin{Bmatrix} a_{m,p} \cdot \cos(\omega_m \cdot p \cdot n) + \\ b_{m,p} \cdot \sin(\omega_m \cdot p \cdot n) \end{Bmatrix} \quad \text{[a.u.]} \quad (1)$$

Analyse-durch-Synthese

Ein Analyse-durch-Synthese (AdS) Verfahren zur automatischen Detektion von Diplophonie aus Audiosignalen wird vorgestellt (Abbildung 1). Die Modellparameter werden bestimmt, indem Oszillator-Kandidaten generiert und optimale Oszillatoren-Kombinationen heuristisch identifiziert werden.

Oszillator-Kandidaten $\hat{d}_\gamma(n)$ werden mit Fourier-Synthesizern generiert. Γ diskrete Grundfrequenz-Kandidaten $\hat{k}_\gamma$ werden aus Betragsspektren durch Peak-

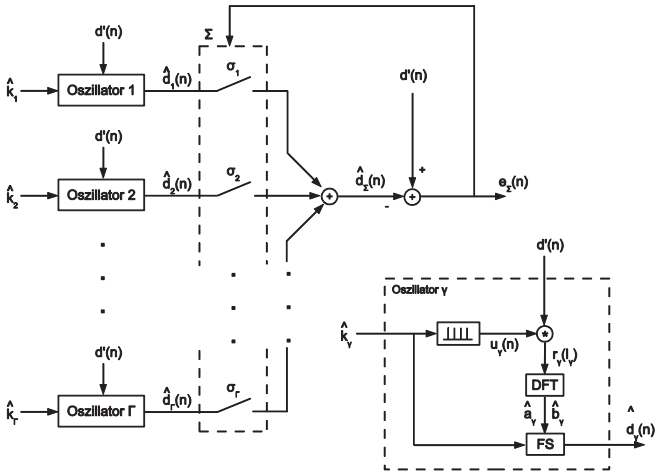

Abbildung 1: Synthesizer für die Detektion von Diplophonie in Audiosignalen. Oszillator-Kandidaten $\hat{d}_\gamma(n)$ werden aus den Grundfrequenz-Kandidaten $\hat{k}_\gamma$ generiert. Die Kreuzkorrelation von Einheitsimpulsketten $u_\gamma(n)$ mit dem Audiosignal wird hierfür Fourier-transformiert (DFT) und gemeinsam mit $\hat{k}_\gamma$ einem Fourier-Synthesizer (FS) zugeführt. Ein Kombinations-Schaltwerk summiert die Oszillator-Kandidaten auf. Die optimalen Oszillator-Kombinationen Σ werden über die Minimierung des Modellfehlers $e_\Sigma(n)$ bestimmt. Die Analyseblöcke sind 64 ms lang.

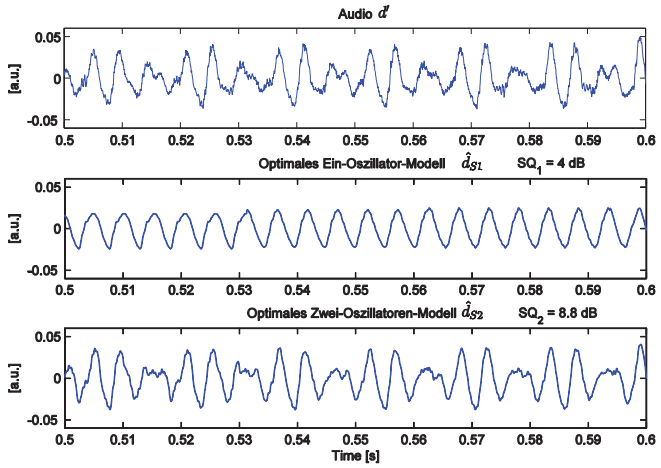

Abbildung 2: Visuelle Evaluierung der Synthese-Qualität für ein diplophones Stimmsignal. Das Audiosignal d' (oben), kann mit einem Ein-Oszillator-Modell $\hat{d}_{\Sigma 1}$ (Mitte) nicht so gut nachgebildet werden wie mit einem Zwei-Oszillatoren-Modell $\hat{d}_{\Sigma 2}$ (unten), was ein Indiz für Diplophonie ist.

picking gewonnen. Der Suchbereich liegt zwischen 70 und 600 Hz und die minimale Peak-Höhe beträgt -15 dB relativ zum Maximum. Aus den Grundfrequenz-Kandidaten werden Einheitsimpulsketten $u_\gamma(n)$ generiert. Die Kreuzkorrelation der Einheitsimpulsketten mit dem Audiosignal $d'(n)$ liefert die Pulsformen $r_\gamma(l_\gamma)$ mit lag-Index l_γ, welche diskret Fourier-transformiert werden (DFT). Die Fourier-Koeffizienten $\hat{a}_\gamma$ und $\hat{b}_\gamma$ steuern mit den Grundfrequenz-Kandidaten die Fourier-Synthesizer (FS), welche die Oszillator-Kandidaten liefern.

Die optimalen Oszillator-Kombinationen $\Sigma_{1,2}$ werden über Minimierung des Modellfehlers $e_\Sigma(n)$ bestimmt. Die Oszillator-Kandidaten werden im Kombinations-Schaltwerk aufsummiert. Alle möglichen Oszillator-Kombinationen $\hat{d}_\Sigma(n)$ mit maximal zwei Oszillatoren werden jeweils vom Audiosignal $d'(n)$ subtrahiert. Das Fehlersignal $e_\Sigma(n)$ wird über das Argument Σ minimiert, um das optimale Ein-Oszillator-Modell $\hat{d}_{\Sigma_1}(n)$ und das optimale Zwei-Oszillatoren-Modell $\hat{d}_{\Sigma_2}(n)$ zu identifizieren.

$SQ_{1,2}$ bilden die Synthese-Qualitäten des optimalen Ein-Oszillator-Modells und des optimalen Zwei-Oszillatoren-Modells ab, d.h. die quantitative Ähnlichkeit der synthetischen Signale $\hat{d}_{\Sigma_1}(n)$ und $\hat{d}_{\Sigma_2}(n)$ mit dem Audiosignal $d'(n)$. $SQ_{1,2}$ sind die root mean square (RMS) Pegelverhältnisse der Fehlersignale $e_{\Sigma_{1,2}}(n)$ mit dem Audiosignal $d'(n)$. Diplophonie wird über logistische Regression mit den Prädiktoren SQ_1 und SQ_2 detektiert.

$$SQ_{1,2} = 20 \cdot \log_{10}\left(\frac{\sqrt{\overline{d'\ (n)^2}}}{\sqrt{\overline{e_{\Sigma_{1,2}}(n)^2}}}\right) \quad \text{[dB]} \qquad (2)$$

Konventionelle Audio-Deskriptoren

Zum Vergleich wurden sechs konventionelle Audio-Deskriptoren untersucht. Diese waren *jitter* [5], *shimmer* [6], [7], Harmonics-to-Noise-Ratio (*HNR*) [4], Göttingen irregularity (*Irr*) und Göttingen noise (*Noi*) [8], und Degree of subharmonics (*DSH*) [9]. Diese Deskriptoren wurden ausgewählt, weil diese in der klinischen Praxis häufig unbesehen zur Analyse aller Heiserkeitsformen verwendet werden und weil Effekte von Diplophonie auf diese Analysen weitgehend unbekannt sind.

Jitter, *shimmer* und *HNR* wurden mit Praat [4] bestimmt. *Jitter* ist die mittlere Zykluslängenschwankung in Prozent. Die verwendeten Subroutinen waren “To Pitch (cc)“, “To PointProcess (cc)” und “Get jitter (local)”. *Shimmer* ist ein Maß für Zyklusenergieschwankung in dB. Es wurden Subroutinen “To Pitch (cc)“, “To PointProcess (cc)” und “Get shimmer (local_dB)” verwendet. *HNR* ist ein Maß für die Prominenz des grundfrequenzassoziierten Peaks in der Autokorrelationsfunktion in dB. *HNR* wurde mit den Subroutinen “To Harmonicity (cc)” und “Get mean” berechnet. Alle Parameter in Praat wurden auf deren Standardwerte gesetzt.

Irr und *Noi* wurden mit dem im Internet frei verfügbaren Göttinger Heiserkeitsdiagramm berechnet [10]. *Irr* ist eine Linearkombination von *jitter*, *shimmer* und der mittleren Korrelation aufeinanderfolgender Zyklen. *Irr* bildet die Irregularitäts-Komponenten des Stimmklangs ab, welche häufig mit Rauigkeit assoziiert werden. *Noi* ist ein Korrelationsmaß der Einhüllenden der Subband-Signale, und bildet die Rausch-Anteile des Stimmklanges ab, welche häufig mit Behauchtheit assoziiert werden.

Der *DSH* wurde in MATLAB berechnet und bildet Doppeldeutigkeit in der Grundfrequenzmessung ab. Die

Tabelle 1: Klassifizierungsraten aller Analyseverfahren im Vergleich. Unser Analyse-durch-Synthese (AdS) Verfahren übertrifft alle anderen Verfahren in Sensitivität und Spezifität. Göttingen irregularity (*Irr*) und noise (*Noi*) [8], Harmonics-to-Noise-Ratio (*HNR*) [4], Degree of subharmonics (*DSH*) [9].

	Sensitivität (%)	Spezifität (%)
Jitter	50.9	80.3
Shimmer	67.3	69.7
Irr	57.8	64.0
Noi	60.0	77.8
HNR	60.0	58.2
DSH	69.1	57.1
AdS	80.0	92.9

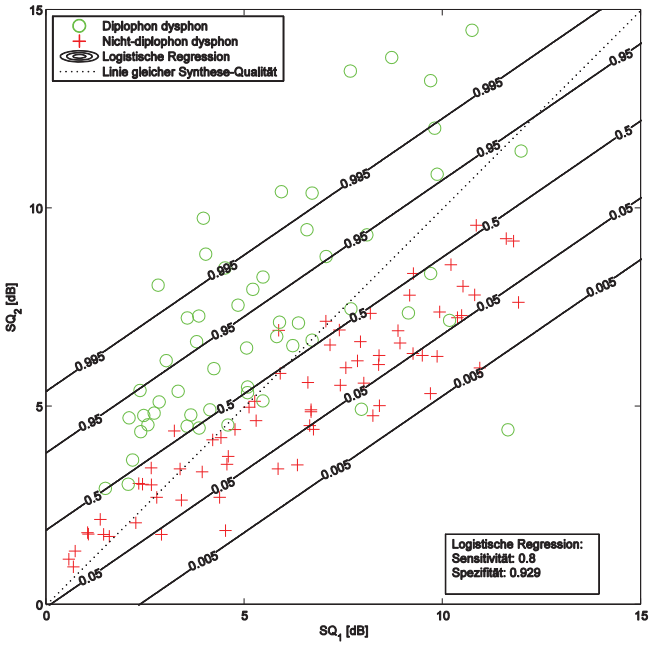

Abbildung 3: Quantitative Evaluierung der Synthese-Qualitäten SQ_1 und SQ_2 für 125 Stimmsignale, respektive Stimmqualität. Die diplophonen Signale spalten sich gut von den nicht-diplophonen Signalen ab.

Signale werden sgn-codiert, d.h. schwellwertabhängig auf -1, 0, oder +1 gesetzt [11]. Ein Analyseblock wird als subharmonisch erkannt, wenn eine kritische Schwellwert-variation zu einer Veränderung der extrahierten Grundfrequenz führt. Der *DSH* entspricht der Rate der subharmonischen Analyseblöcke in Prozent. Insbesondere reagiert der *DSH* auf zyklische Schwankungen der Einhüllenden des Zeitsignals, z.B. auf Schwebungseffekte bei Diplophonie.

Für jeden der konventionellen Deskriptoren wurde ein Schwellwert-Klassifikator gebaut. Receiver-Operating-Characteristic (ROC) Kurven beziehen Klassifikator-Schwellwerte auf Sensitivität und Spezifität und wurden für alle konventionellen Audio-Deskriptoren geplottet. Der Schwellwert, der den Abstand zwischen ROC Kurve und dem Punkt (Sensitivität = 1, Spezifität = 1) minimiert, wird als optimaler Schwellwert angenommen, was einen sinnvollen Kompromiss zwischen bestmöglicher Sensitivität und Spezifität darstellt.

Ergebnisse

Das AdS Verfahren wird visuell und quantitativ evaluiert. Sensitivität und Spezifität aller Audio-Analysen werden verglichen.

Abbildung 2 zeigt einen Abschnitt eines diplophonen Audiosignals (d', oben), das optimale Ein-Oszillator-Modell ($\hat{d}_{\Sigma_1}$, Mitte) und das optimale Zwei-Oszillatoren-Modell ($\hat{d}_{\Sigma_2}$, unten). Das optimale Ein-Oszillator-Modell ist nicht geeignet um die Wellenform nachzubilden, wohingegen das optimale Zwei-Oszillatoren-Modell dem Audiosignal ähnlich ist.

Abbildung 3 zeigt das Streudiagramm von SQ_1 mit SQ_2, gruppiert nach Stimmqualität. Die Kreise korrespondieren mit diplophonen Audiosignalen, und die Kreuze mit nicht-diplophonen Signalen. Die diplophonen Signale spalten sich links oben im Diagramm ab, während sich die nicht-diplophonen Signale rechts unten scharen. Dieser Trend entspricht grob dem Kriterium $SQ_2 > SQ_1$, wobei die gepunktete Linie die Linie gleicher Synthese-Qualitäten ($SQ_1 = SQ_2$) ist.

Zur genaueren Analyse des Parameterraumes wurde ein logistisches Regressionsmodell trainiert, welches in Abbildung 3 durch Wahrscheinlichkeitskonturen dargestellt ist. Die 50 % Kontur entspricht der Entscheidungsgrenze, welche geringfügig flacher ist als die Linie gleicher Synthese-Qualität. Diese Abflachung wird durch aperiodische nicht-diplophone Signale hervorgerufen, die mit harmonischen Oszillatoren nicht gültig modelliert werden können. Diese scharen sich links unten und liegen per Zufall leicht über oder unter der Linie gleicher Synthese-Qualität. Die Sensitivität des Verfahrens wird auf 80.0 % geschätzt, da 44 der 55 diplophonen Audiosignale richtig erkannt werden. Die Spezifität wird auf 92.9 % geschätzt, da 65 der 70 nicht-diplophonen Audiosignale richtig erkannt werden.

Tabelle 1 zeigt die Sensitivitäten und Spezifitäten aller Audio-Analysen. Unter den konventionellen Deskriptoren erreicht der *DSH* die beste Sensitivität (69.1 %) und *jitter* die beste Spezifität (80.3 %), während AdS überlegen ist.

Diskussion, Schlussfolgerungen und Ausblick

Das komplexe akustische Phänomen der Diplophonie wurde in [3] aus mehreren Blickwinkeln erforscht. Dieser Konferenzbeitrag bietet einen exemplarischen Einblick in diese Forschungsarbeit. Im Folgenden werden der erreichte Mehrwert erläutert, die Definition von Diplophonie diskutiert, die aktuellen Ansätze der Diplophonie-Forschung erwähnt und ein Ausblick gegeben.

Dieser Beitrag zeigt am Beispiel der Diplophonie, wie akustische Stimmklangphänomene mit einem AdS Verfahren beschrieben werden können. Sensitivität und Spezifität übertreffen jene der getesteten konventionellen Features. Weitere Vorteile gegenüber herkömmlichen Verfahren ergeben sich aus der automatischen

Modellstrukturoptimierung. Erstens wird durch Automatismus der Weg zu einer vereinheitlichten und objektiven Diagnostik geebnet, indem subjektive Stimmklangphänomene festgemacht werden. Zweitens wird eine physiologische Interpretation ermöglicht, da Stimmproduktionsmechanismen qualitativ abgebildet werden. Bei der Diplophonie werden zwei räumlich separierte Stimmlippenoszillatoren mit unterschiedlichen Grundfrequenzen detektiert. Drittens wird das Problem der invaliden Signalmodelle explizit adressiert, indem beim Testen die Modellgüte evaluiert wird. Hohe Modellgüte ist ein Indikator für das Vorhandensein der Zieleigenschaft (z.B. Diplophonie) und kann ausschließlich erreicht werden, wenn das Signalmodell dem getesteten Signal entspricht. Diese Vorgehensweise folgt dem Paradigma des modellabhängigen Realismus, in dem das beste Modell die angenommene Wirklichkeit bestimmt [14].

Es existieren divergente Definitionen pathologischer Stimmklangphänomene. Definitionen von Diplophonie verorten sich auf den Ebenen der auditiven Perzeption, der akustischen Wellenform und der Stimmlippenschwingung. Die auditive Definition ist die pragmatischste, klinisch am praktikabelsten und für die Beurteilung kommunikativer Aspekte relevant. Die Definition auf der Ebene der akustischen Wellenformen ermöglicht die vollautomatische synthesebasierte Detektion von Diplophonie. Auf der Ebene der Wellenform existiert auch eine alternative Definition von Diplophonie, nämlich „alternierende Zyklusparameter“, d.h. z.B. abwechselnd hohe und niedrige Spitzenamplituden oder lange und kurze Zyklen [12]. Diese Definition umfasst weder alle Signale, die zu auditiver Diplophonie führen, noch bewirken derartige Signale auditive Diplophonie notwendigerweise. Ein passenderer Begriff zur Beschreibung alternierender Wellenformmuster ist etwa „Bizyklik“ [13]. Eine Definition auf der Ebene der Stimmlippenschwingung steht im Zusammenhang mit der mechanischen Ursache der Diplophonie. Diese glottale Diplophonie ist klinisch von großem Interesse aber schwer zu untersuchen, weil Hochgeschwindigkeitsvideos gebraucht werden.

Weitere aktuelle Ansätze der Diplophonie-Forschung sind die qualitative und quantitative Modellierung und Analyse der Stimmlippenschwingung mithilfe von Hochgeschwindigkeitsvideos, die Untersuchung auditiver Effekte in psychoakustischen Experimenten, die Grundfrequenzmessung bei Diplophonie, sowie latente Klassenanalyse zur probabilistischen Evaluierung der diagnostischen Basisklassifizierung. In Zukunft soll die für Diplophonie erschlossene Methodik für andere Formen der Heiserkeit adaptiert werden.

Danksagung

Die Autoren bedanken sich bei der Thesis-Jury Gernot Kubin, Jean Schoentgen und Berit Schneider-Stickler, bei den beteiligten Institutionen, bei allen weiteren Teammitgliedern, Beratern und Diskutanten [3], sowie bei der Richard Wolf GmbH für die Bereitstellung der Hochgeschwindigkeitskamera.

Literatur

[1] Werth K., Voigt D., Döllinger M., Eysholdt U. und Lohscheller J.: Clinical value of acoustic voice measures: a retrospective study. European Archives of Oto-Rhino-Laryngology 267 (2010), 1261-1271

[2] Dejonckere, P.H. und Lebacq, J.: An analysis of the diplophonia phenomenon. Speech Communication 2 (1983), 47-56

[3] Aichinger, P. Diplophonic Voice - Definitions, models, and detection. PhD Dissertation, Technische Universität Graz, 2015

[4] Free software Praat, URL: `http://www.fon.hum.uva.nl/praat/download_win.html`

[5] Lieberman, P.: Perturbations in vocal pitch. The Journal of the Acoustical Society of America 33 (1961), 597-603

[6] Wendahl, R.W.: Some parameters of auditory roughness. Folia Phoniatrica et Logopaedica 18 (1966), 26-32

[7] Wendahl, R.W.: Laryngeal analog synthesis of jitter and shimmer auditory parameters of harshness. Folia Phoniatrica et Logopaedica 18 (1966), 98-108

[8] Michaelis, D., Fröhlich, M. und Strube, H.: Selection and combination of acoustic features for the description of pathologic voices. Journal of the Acoustical Society of America 103 (1998) 1628-1639

[9] Deliyski, D.D.: Acoustic model and evaluation of pathological voice production. In Third European Conference on Speech Communication and Technology (1993) 1969-1972

[10] Free software The Hoarseness Diagram, URL: `http://www.physik3.gwdg.de/~micha/hd.html`

[11] Rabiner, L.: On the use of autocorrelation analysis for pitch detection. IEEE Transactions on Acoustics, Speech and Signal Processing 25 (1977) 24-33

[12] Titze, I. Workshop on acoustic voice analysis: Summary statement. National Center for Voice and Speech, Iowa, 1995

[13] Kreiman, J., Gerratt, B.R., Precoda, K. und Berke, G.S.: Perception of supraperiodic voices. The Journal of the Acoustical Society of America 93 (1993) 2337

[14] Hawking S. und Mlodinow L. The Grand Design. Bantam Books, New York, 2010