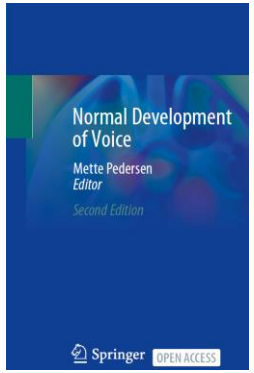
Minimum Standard Measures ! of the Singing Voice

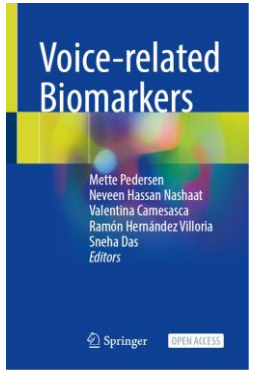
PEVOC, 25-28 March 2026

Authors,

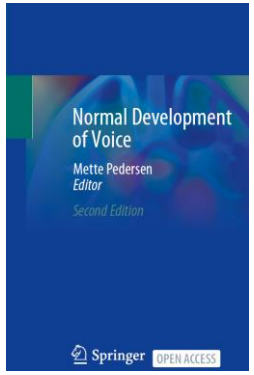
Mette Pedersen, Vitus Girelli Meiner, & Sneha Das
Denmark

From the Medical Center Østergade 18, Copenhagen 1100
IT-University, Rued Langgaards Vej 7, Copenhagen 2300
Technical University of Denmark, Anker Engelunds Vej 1, Bygning 101A, Kongens Lyngby 2800



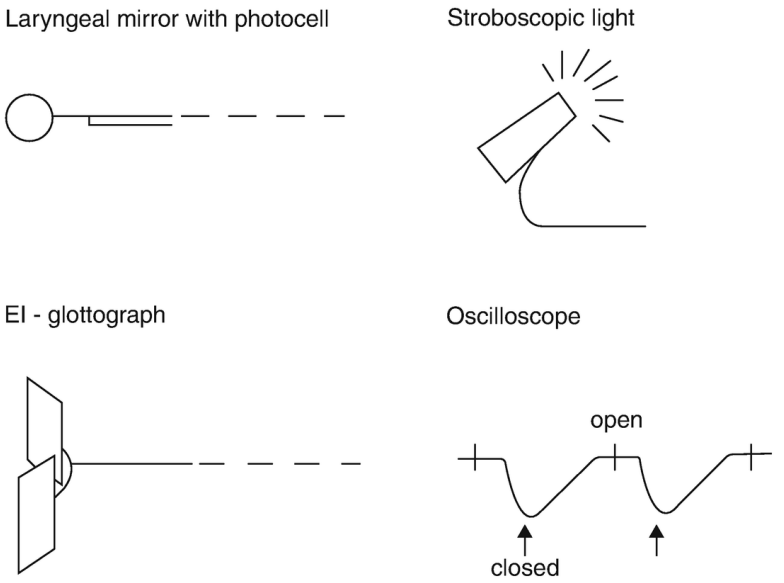


- A presentation is made of two new book on
- Basic science in laryngology related to the normal voice development
- Voice-related Biomarkers for use as standards based on AI



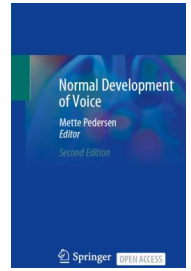
In the first book we used electroglottography with a synchronized stroboscopy for fundamental frequency to be sure that the first harmonic was measured in the octave shift in pubertal children.

- Stroboscopy shows no change in the movement pattern corresponding to the octave jump in male puberty.
- With electroglottography, there is an obvious change between the two registers in childhood and in young males.
- So you can define the full register change during puberty. Without mixing up with overtones.

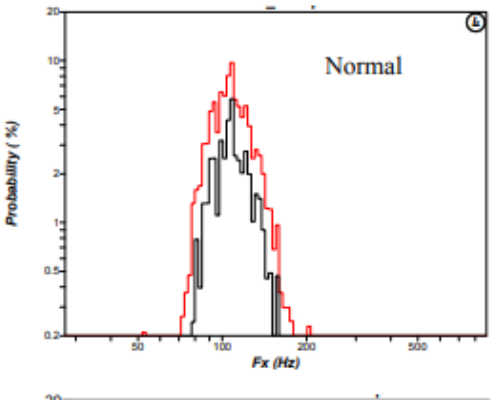


In order to secure the duty cycle, a **photocell** can be coupled to the stroboscope connecting it to the electroglottograph to define the fundamendal frequency at the level of the vocal folds

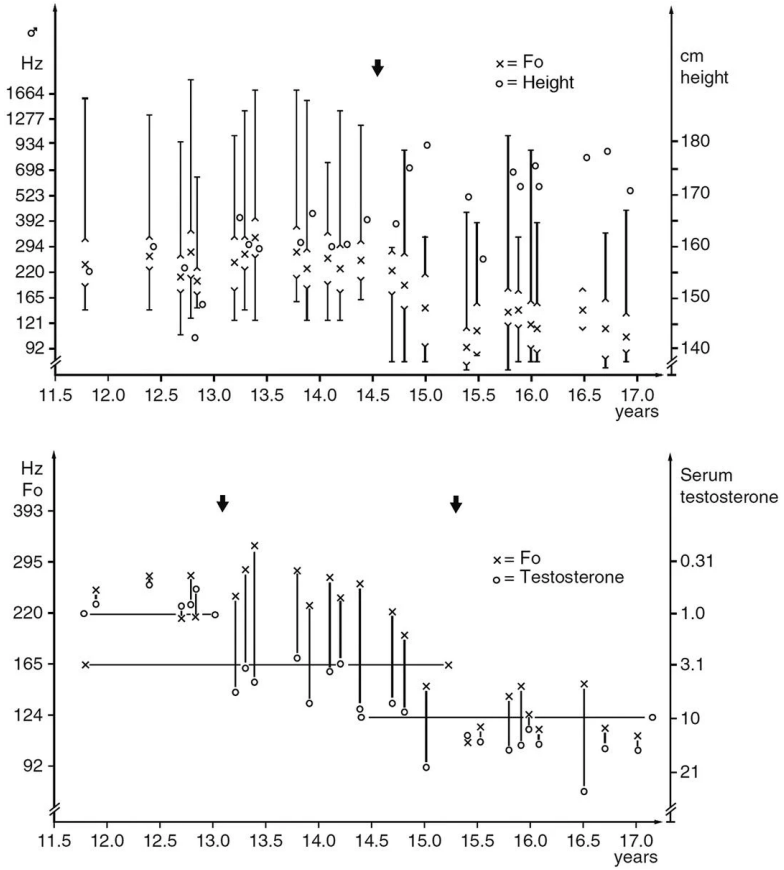
The fundamental frequency and tone range of reading of a standard text and total tonerange during puberty was measured, related to the hormonal development but that was not enough for the singing teachers.



We have used the mean of 2000 duty cycles reading of the text: the north wind and the sun in a Danish phonetically balanced translation.
The F0 always has a maximum and an area, as presented in the picture from Laryngograph LTD. In this study, 25 boys ages 8-19 were divided into class levels in a local music school.



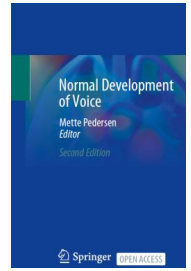
The principle of fundamental measuring frequency with electroglottography during reading



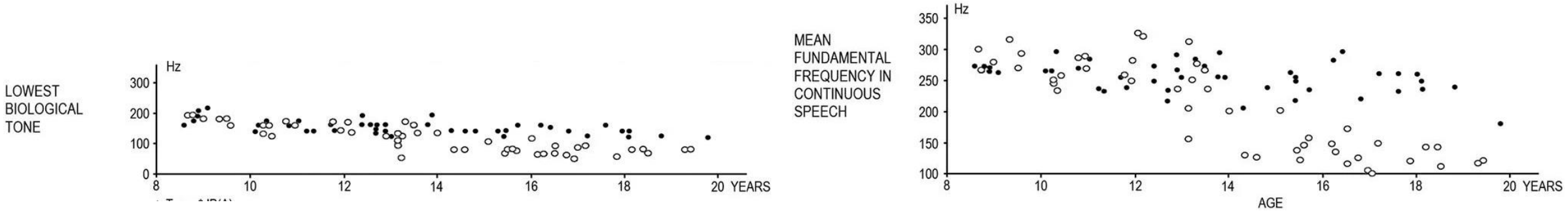
(**a**) Boys' mean fundamental frequency in continuous speech (F0), tonal range of voice in continuous speech in Hz, and total tone range in Hz compared to body height (ordinate) and age (abscissa) in 25 boys. The arrow indicates the end of the voice change. (**b**) Boys' mean fundamental frequency during continuous speech (F0) in the 25 boys compared to total serum testosterone level. The abscissa shows the age in years. The arrows indicate the beginning and end of the pubertal voice change

From the Medical Center Østergade 18, Copenhagen 1100 IT-University, Rued Langgaards Vej 7, Copenhagen 2300 Technical University of Denmark, Anker Engelunds Vej 1, Bygning 101A, Kongens Lyngby 2800 https://www.mettepedersen.org/presentations

The fundamental frequency during puberty was therefore supplemented with phonetograms, and now the singing teachers were more satisfied, especially since the use of the lowest tone in the phonetograms was helpful in the daily practice.

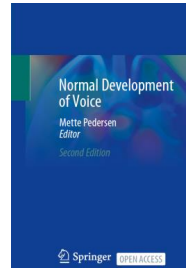


We have used the mean of 2000 duty cycles reading of the text: the north wind and the sun in a Danish phonetically balanced translation.
The F0 has always an area as presented in the picture from Laryngograph LTD. Here we present only the means Fo, ages 8-19 divided in class levels, of 48 boys and 47 girls in a local music school.

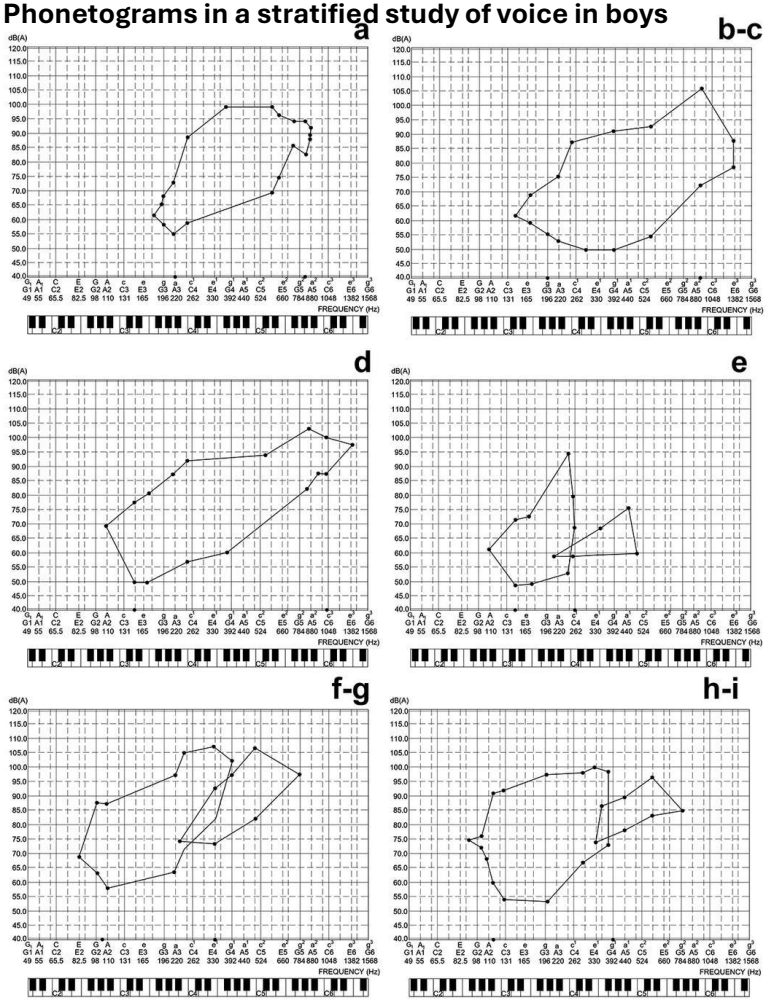


Boys' and girls' age-related comparison on the chromatic scale of the lowest tone, and the mean fundamental frequency in continuous speech (F0): filled circle: girls; open circle: boys

Phonetograms development measurement in normal childhood takes time, but the octave jump in puberty is clearly seen.

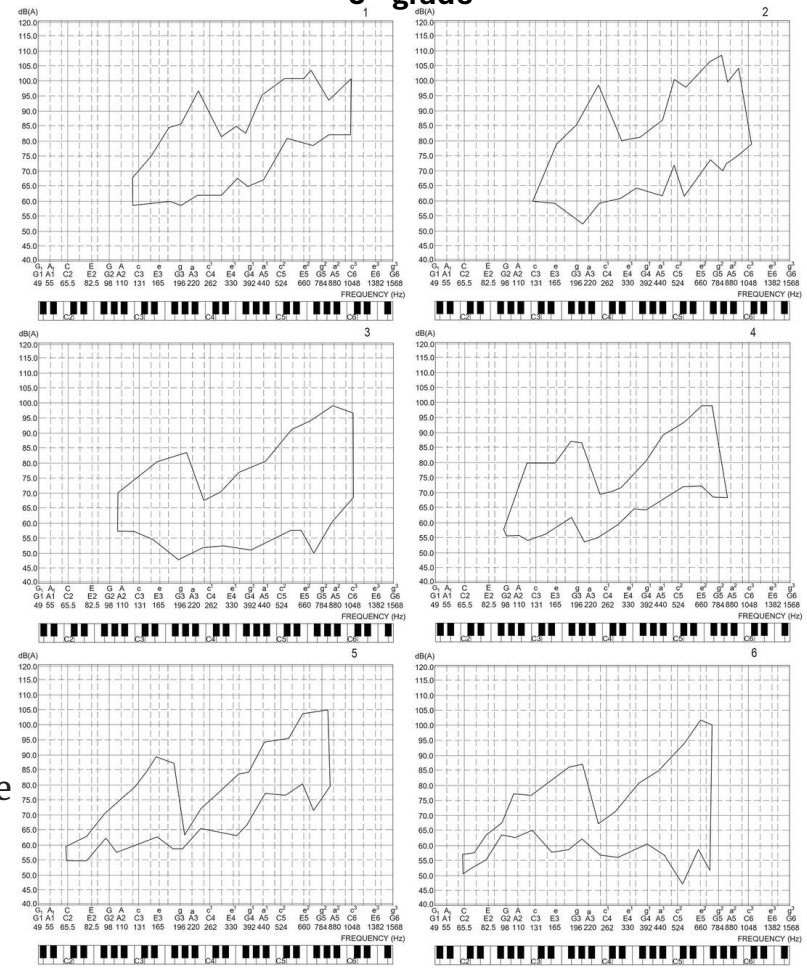


Phonetograms in a stratified study of voice in boys



Boys' biological Voice Range Profile from childhood over puberty to past puberty representing child voices and beginning adult voice ranges. (The range of the "artistically" usable singing voice is not marked on the abscissa. (**a**) 9-Year-old child at adrenarche. (**b**, **c**) Child voice (soprano) with higher intensity for upper tones. (**d**) Child voice (alto) with higher intensity for lower tones. (**e**) Voice in puberty. (**f**- **g**) Beginning adult voice (tenor). (**h**- **i**) Bass

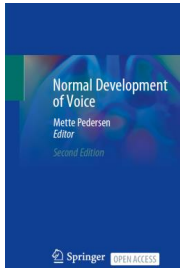
One boy in a longitudinal study with two months interval in 8th grade



6 **Voice Range Profiles of one boy measured at intervals of 2 months (age 13.7–14.6 years)** in the eighth school class. The third Voice Range Profile (December) has the biggest area and shows the smallest irregularities (c1 = C4 = 262 Hz)

From the Medical Center Østergade 18, Copenhagen 1100 IT-University, Rued Langgaards Vej 7, Copenhagen 2300 Technical University of Denmark, Anker Engelunds Vej 1, Bygning 101A, Kongens Lyngby 2800

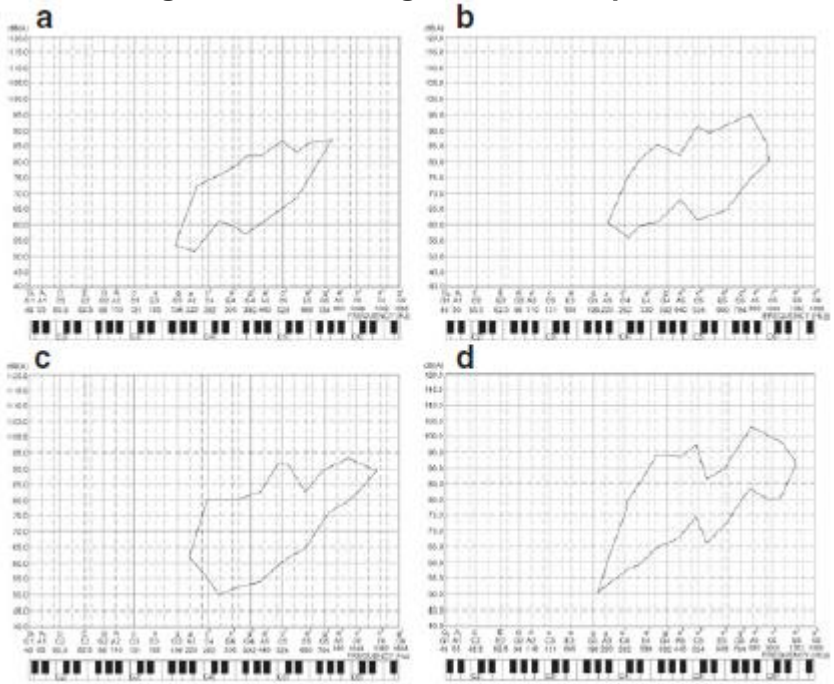
https://www.mettepedersen.org/presentations



Phonetograms development in normal childhood in girls was also satisfactory for the singing teachers because it could be shown that the phonetograms after the pubertal change were much bigger than before. A pubertal female voice is defined

.

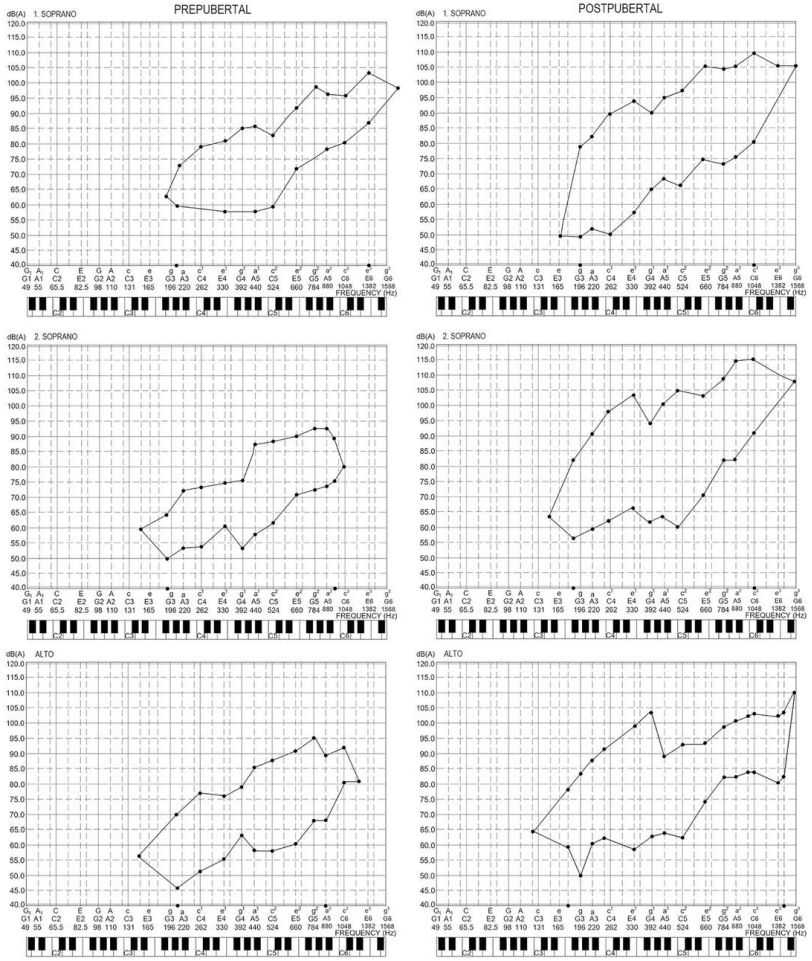
Voice Range Profiles During Voice Development



Girls' Voice Range Profiles of different ages. (**a**) 8.9 years. (**b**) 11.7 years, typical child's voice with dominating intensity in the upper part, change of register at 330–392 Hz. (**c**) 13.8 years, voice with slight register changes with greater dynamic breadth in the lower part. (**d**) 14.8 years, pubertal voice with passing reduced intensity in the middle

Girls' Voice Range Profile development during childhood. In the upper frequency range, there is a bigger intensity for the sopranos, and in the lower frequency range for the altos. Prepubertal soprano and alto and postpubertal first and second soprano and alto are shown. The singing range of the "artistic" voice is not given on the abscissa

Phonetograms in a stratified study of voice in girls



From the Medical Center Østergade 18, Copenhagen 1100 IT-University, Rued Langgaards Vej 7, Copenhagen 2300 Technical University of Denmark, Anker Engelunds Vej 1, Bygning 101A, Kongens Lyngby 2800 https://www.mettepedersen.org/presentations

Phonetograms development in normal childhood

Examples of Individual Prepubertal, Pubertal, and Postpubertal Phonetograms of voice in girls

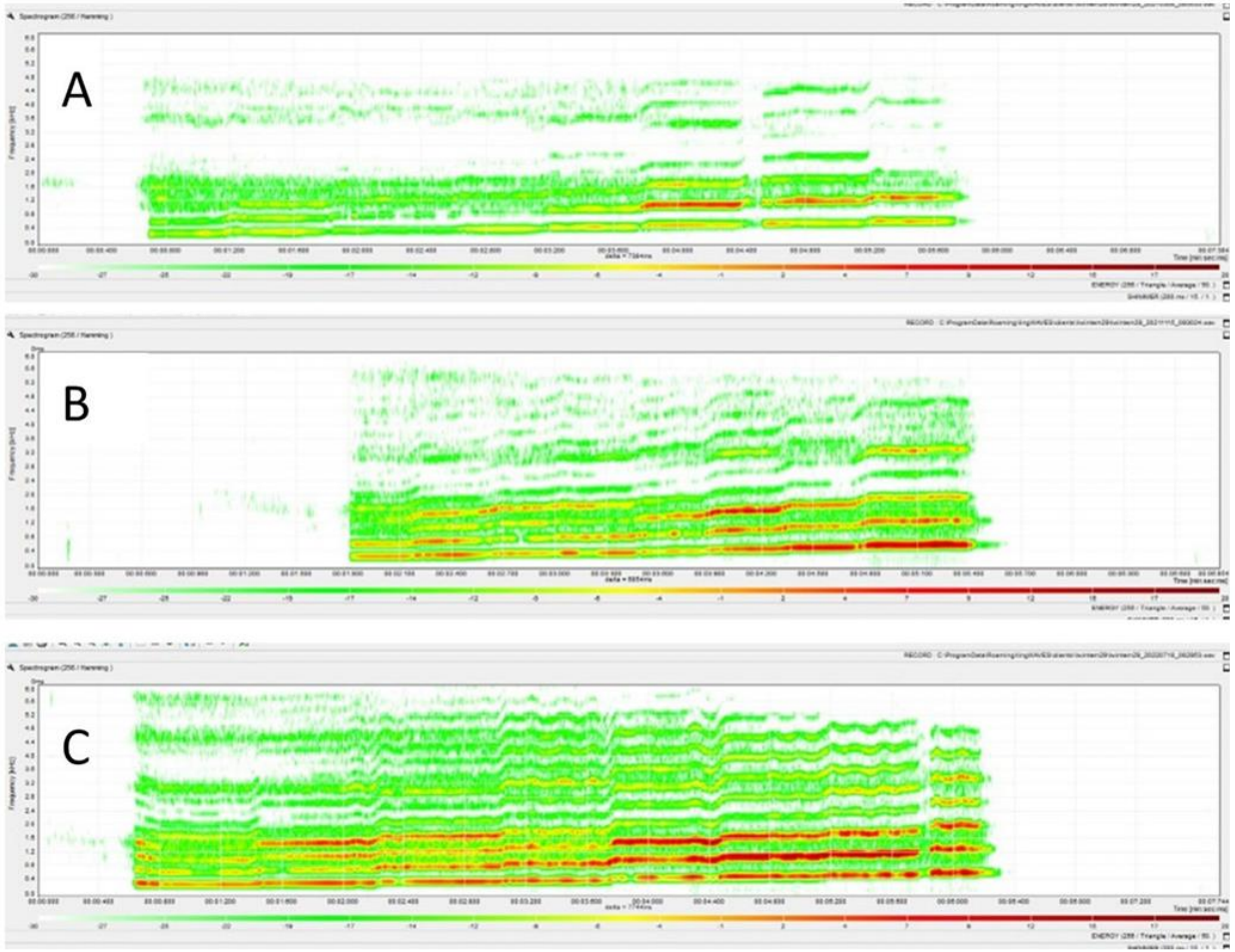
* Thesis 2026, from Griffith University of Australia

The first row shows examples of individual prepubertal phonetograms ages in months from left to right: 127, 120, and 131 (10 years old). The second row shows examples of individual pubertal phonetograms, ages in months from left to right: 166, 133, and 155. The third row shows examples of individual post-pubertal phonetograms, ages in months from left to right: 207, 193, and 210 months (16 years old). Images generated with LingWaves by WEVOSYS program.

Examples of Formant Presentations in the Spectrograms

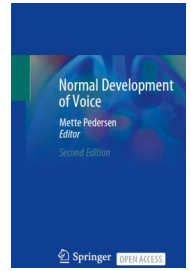
Note. Spectrogram letters correlate with their assigned formant groups. A, B, and C. Images generated in LingWaves by WEVOSYS, 2017.

* Thesis 2026, from Griffith University of Australia

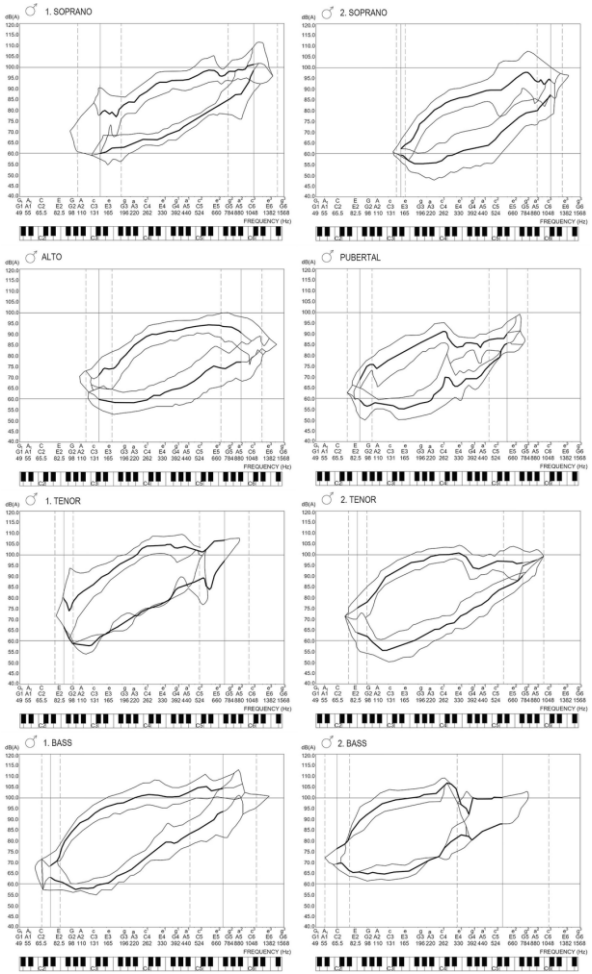


The first row shows examples of individual prepubertal spectrograms ages in months from left to right: 127, 120, and 131 (10 years). The second row shows examples of individual pubertal phonetograms, ages in months from left to right: 166, 133, and 155. The third row shows examples of individual post-pubertal spectrograms, ages in months from left to right: 207, 193, and 210 months (16 years). Images generated with LingWaves by WEVOSYS program.

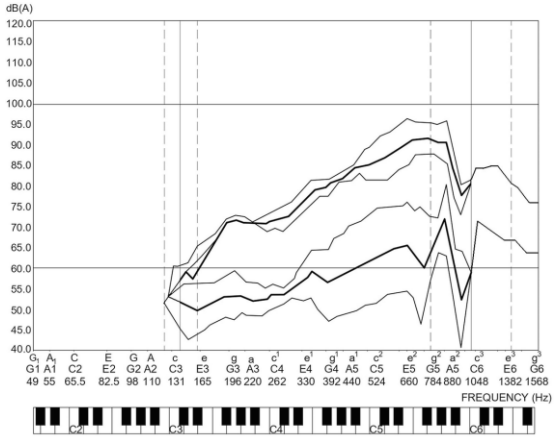
Here we show calculations of standard deviations of 5-6 phonetograms in each singing category in boys corresponding to the Thomaner Choir who have the same hormonal values.



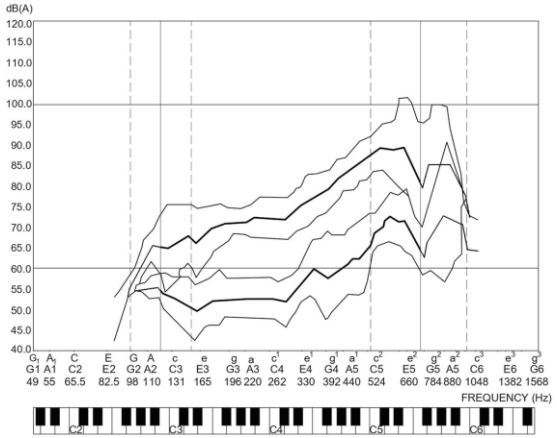
Standard deviations on all phonetograms are divided per class level with the tones **c g e** & **a** within the octaves. The lowest tones are falling and the areas are widening.



Boys' average Voice Range Profiles with standard deviation in a Danish ordinary school and high school, as a function of voice category. The abscissa is divided up into tones, and the frequency in Hz is indicated. The scale of the ordinate is dB(A)

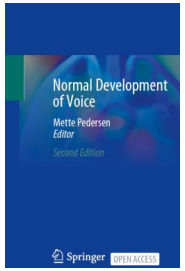


Boys' average Voice Range Profiles with standard deviation for the cohort of four sopranos and of pubertal change groups (mutants) from the Leipzig Thomanerchor school. The hormonal parameters were similar to those of the boys in the Danish school system



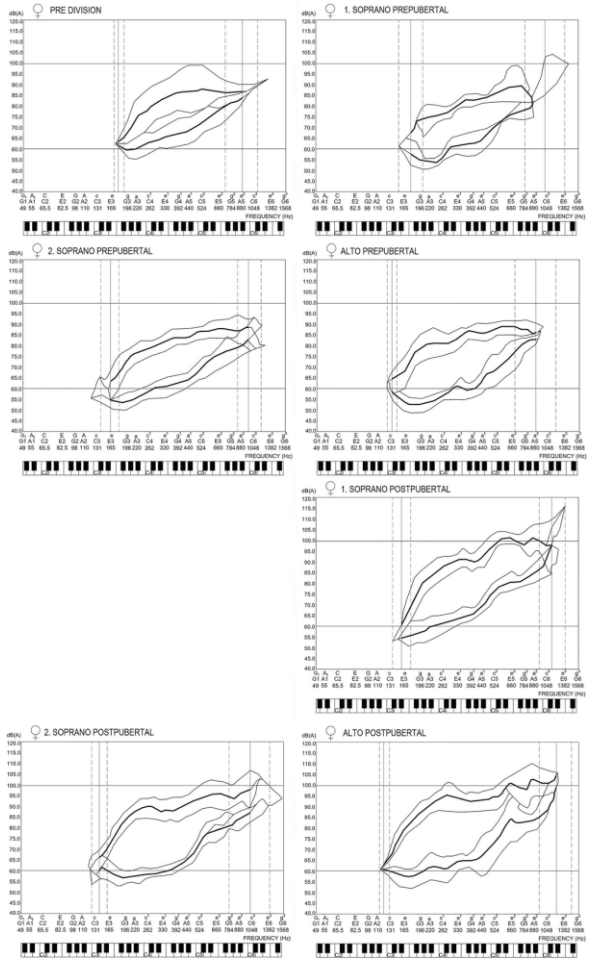
From the Medical Center Østergade 18, Copenhagen 1100 IT-University, Rued Langgaards Vej 7, Copenhagen 2300 Technical University of Denmark, Anker Engelunds Vej 1, Bygning 101A, Kongens Lyngby 2800

https://www.mettepedersen.org/presentations



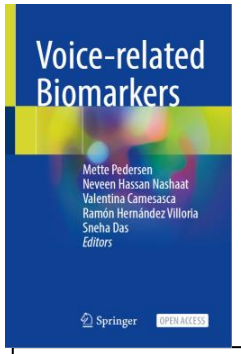
Calculations of standard deviations of 6-7 phonetograms in each category of the girls – no pubertal group was defined by the singing teachers

Standard deviations on all phonetograms are divided per class level with the tones **c g e** & **a** within the octaves. The lowest tones are falling and the areas are widening.



Girls' average Voice Range Profile and ranges with standard deviation for the lowest and highest tones from a Danish ordinary and high school with choirs, as a function of voice category. The abscissa is divided up into tones, and the frequency in Hz is indicated. The scale of the ordinate is dB(A). One group could not be securely defined during puberty

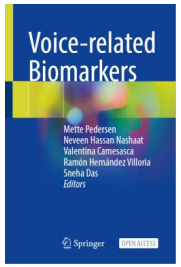
The study shows the importance of basic science and measurements in laryngology, illustrated by the finding of a pubertal girl's voice



Reference: guidelines of the European Laryngological Society and Union of the European Phoniatricians. Eur Arch Otorhinolaryngol. 2023 Dec;280(12):5459-5473. doi: 10.1007/s00405-023-08211-6. Epub 2023 Sep 14. PMID: 37707614.

In the second book, Voice-related biomarkers are discussed by phoniatrians with one AI expert. A suggested Voice test, similar to a Hearing test based on manual voice-related biomarkers, is possible, and can in the future be performed quickly with AI.

- **Voice Handicap Index:** Subjective complaints (self-reported)
- **GRBAS tests (or Cape V):** Perceptive evaluation (by more or less qualified specialists)
- **Acoustical measures: (**minimum**)** F0, Jitter, Shimmer, Harmonics to Noise
- **Maximum Phonation time: (**Airflow, - as a minimum measure**)**



Voice Handicap Index questionnaire as a Voice-related Biomarker illustrates patient complaints

- **Functional**
  How does this affect daily functioning?
- **Physical**
  Sensations (roughness, vocal fatigue, etc.)
- **Emotional**
  Frustration, concern, Irritation

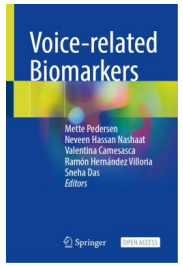
0-30 - None or slight effect on the voice
31-60 - Moderate effect
61-120 - Serious effect

Patient complaints can be answered by the patient in the waiting room – **the score can be used in the AI result**

Feature	VHI	SVHI	SVHI-10
Full name	Voice Handicap Index	Singing Voice Handicap Index	Singing Voice Handicap Index – 10
Primary target	Speaking voice disorders	Singing voice disorders	Singing voice disorders (screening)
Population	General voice patients	Singers (professional & amateur)	Singers
Number of items	30	36	10
Domains / focus	Functional, Physical, Emotional	Singing-specific function, range, endurance, emotional impact	Core singing limitations
Response scale	0–4 Likert (Never → Always)	0–4 Likert (Never → Always)	0–4 Likert (Never → Always)
Total score range	0–120	0–144	0–40
Sensitivity to singing issues	Low–moderate	**High**	**Moderate–high**
Sensitivity to subtle deficits	Limited	**High**	Moderate
Typical administration time	~5 minutes	~6–8 minutes	~2 minutes
Clinical use	General dysphonia assessment	Gold standard for singers	Rapid screening / follow-up
Pre–post treatment tracking	Yes	**Yes (preferred for singers)**	Yes
Use in professional voice clinics	Standard tool	**Primary tool**	Supplementary
Copyrighted	Yes	Yes	Yes

Reference: Jacobson BH, Johnson A, Grywalski C, et al. The Voice Handicap Index (VHI): Development and Validation. *American Journal of Speech-Language Pathology*. 1997;6(3):66–70. DOI: 10.1044/1058-0360.0603.66

From the Medical Center Østergade 18, Copenhagen 1100 IT-University, Rued Langgaards Vej 7, Copenhagen 2300 Technical University of Denmark, Anker Engelunds Vej 1, Bygning 101A, Kongens Lyngby 2800 https://www.mettepedersen.org/presentations



Voice Handicap Index questionnaire

Reference: Jacobson BH, Johnson A, Grywalski C, et al. The Voice Handicap Index (VHI): Development and Validation. *American Journal of Speech-Language Pathology*. 1997;6(3):66–70. DOI: 10.1044/1058-0360.0603.66

Functional (F) subscale

1. My voice makes it difficult for people to hear me.
2. People have difficulty understanding me in a noisy room.
3. My voice difficulties restrict personal and social life.
4. I feel left out of conversations because of my voice.
5. My voice problem causes me to lose income.
6. I have trouble using the telephone because of my voice.
7. I avoid groups of people because of my voice.
8. People ask me, “What’s wrong with your voice?”
9. My voice problem causes me to avoid social outings.
10. My voice problem limits my participation in social activities.

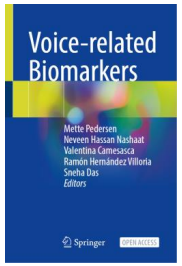
Physical (P) subscale

1. I run out of air when I talk.
2. The sound of my voice varies throughout the day.
3. People ask, “What’s wrong with your voice?” *(note: appears in functional domain in some clinical adaptations; retained here in original item list)*
4. My voice sounds creaky and dry.
5. I feel as though I have to strain to produce voice.
6. The clarity of my voice is unpredictable.
7. My voice problem upsets me.
8. My voice makes it difficult for people to hear me. *(parallel perception item)*
9. I feel discomfort when I speak.
10. My voice sounds rough or hoarse.

Emotional (E) subscale

1. I feel tense when talking to others because of my voice.
2. I am annoyed by my voice problem.
3. I feel embarrassed when people ask me to repeat myself.
4. My voice makes me feel handicapped.
5. I feel frustrated with my voice problem.
6. I feel depressed because of my voice problem.
7. I feel ashamed of my voice problem.
8. My voice problem makes me feel incompetent.
9. I am less outgoing because of my voice problem.
10. My voice problem affects my self-esteem.

From the Medical Center Østergade 18, Copenhagen 1100 IT-University, Rued Langgaards Vej 7, Copenhagen 2300 Technical University of Denmark, Anker Engelunds Vej 1, Bygning 101A, Kongens Lyngby 2800 https://www.mettepedersen.org/presentations

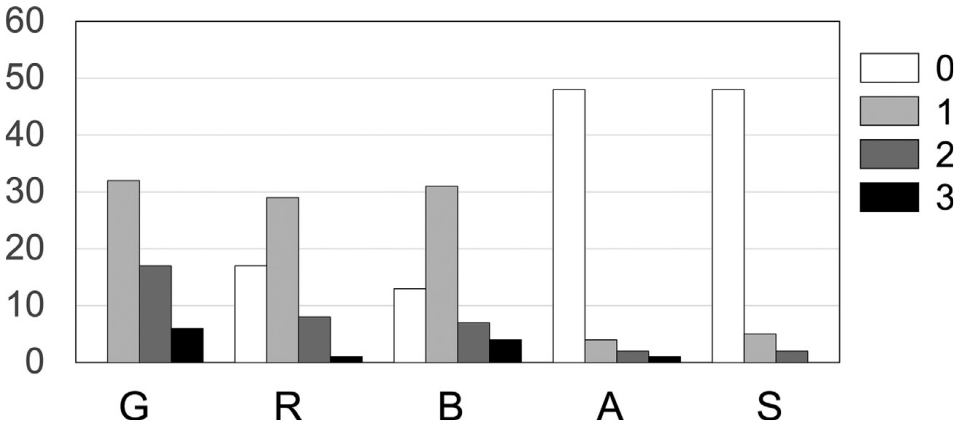


GRBAS tests as a Voice-related biomarker(further developed in Cape-V in the US).

G - **G**rade
R - **R**oughness
B - **B**reathiness
A - **A**stenia
S - **S**train

Evaluation 0-3 :
0: Light,
1: Light/not normal.
2: clear Technical Voice disturbance.
3: marked pathological voice

An automatic evaluation can be done with AI



Distribution of the median values of ratings for test data by the **three evaluators. Significance of agreement minimum** >0.667.

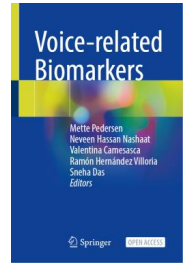
TABLE 1.
Inter-Rater Agreement of the Ratings by TensorFlow, Core ML, and Human

	G	R	B	A	S
Krippen-dorff's alpha	0.6982	0.2232	0.5913	0.3027	0.2763

Reference: Kojima T, Fujimura S, Hasebe K, Okanoue Y, Shuya O, Yuki R, Shoji K, Hori R, Kishimoto Y, Omori K. Objective assessment of pathological voice using artificial intelligence based on the GRBAS scale. *Journal of Voice*. 2021. doi:10.1016/j.jvoice.2021.11.021

Acoustical Measures of Voice-related Biomarkers

(Method dependence on: software, recording quality, noise, vocal variant (sustained vowel, speaking, singing))



F0 (Fundamental frequency speaking)

Group	Mean F0 (Hz)	Typical range (Hz)
Adult females	~200–210 Hz	**165–255 Hz**
Adult males	~110–120 Hz	**85–180 Hz**

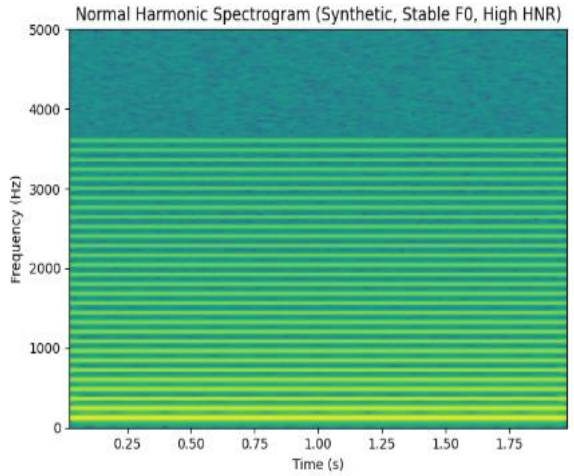
Age-related changes in adults
Females
Young adulthood (20–40 yrs): F0 typically ~200–220 Hz
Middle age (40–60 yrs) :Slight decrease in F0
Older age (60+ yrs): Often a further lowering of F0
Males
Young adulthood (20–40 yrs):
F0 typically ~110–120 Hz
Middle age (40–60 yrs):
Gradual increase in F0
Older age (60+ yrs):
Noticeable increase in F0

Normal values in the speaking area (a chosen frequency area is possible)

Jitter (*Frequency variation)*
Definition, normal**:** <1%
At best: 0,5-0,6%
Roughness: ⬆

Shimmer (Amplitude variation)
Definition, normal: ≤3-5%
At best: <3%
Strain: ⬆

Normal Harmonic Spectrogram (Synthetic, Stable F0, High HNR)

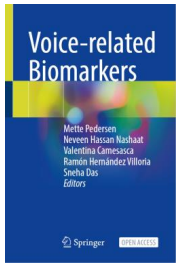


Harmonics to Noise
Definition, normal: ≥ 20 dB
Clear harmonic bands, very little noise between them. Parallel stable lines, clear structure
Minor deviation: 15-20 dB

An automatic evaluation can be done with AI software

Maryn Y, Roy N, De Bodt M, Van Cauwenberge P, Corthals P. Acoustic measurement of overall voice quality: a meta-analysis. J Acoust Soc Am. 2009 Nov;126(5):2619-34. doi: 10.1121/1.3224706. PMID: 19894840.

From the Medical Center Østergade 18, Copenhagen 1100 IT-University, Rued Langgaards Vej 7, Copenhagen 2300 Technical University of Denmark, Anker Engelunds Vej 1, Bygning 101A, Kongens Lyngby 2800 https://www.mettepedersen.org/presentations



Maximum Phonation Time

- Definition: The longest duration a person can sustain a vowel (typically /a/) on a single exhalation with stable voice quality. This measure reflects respiratory control, vocal fold closure (glottal efficiency), and neuromuscular coordination.

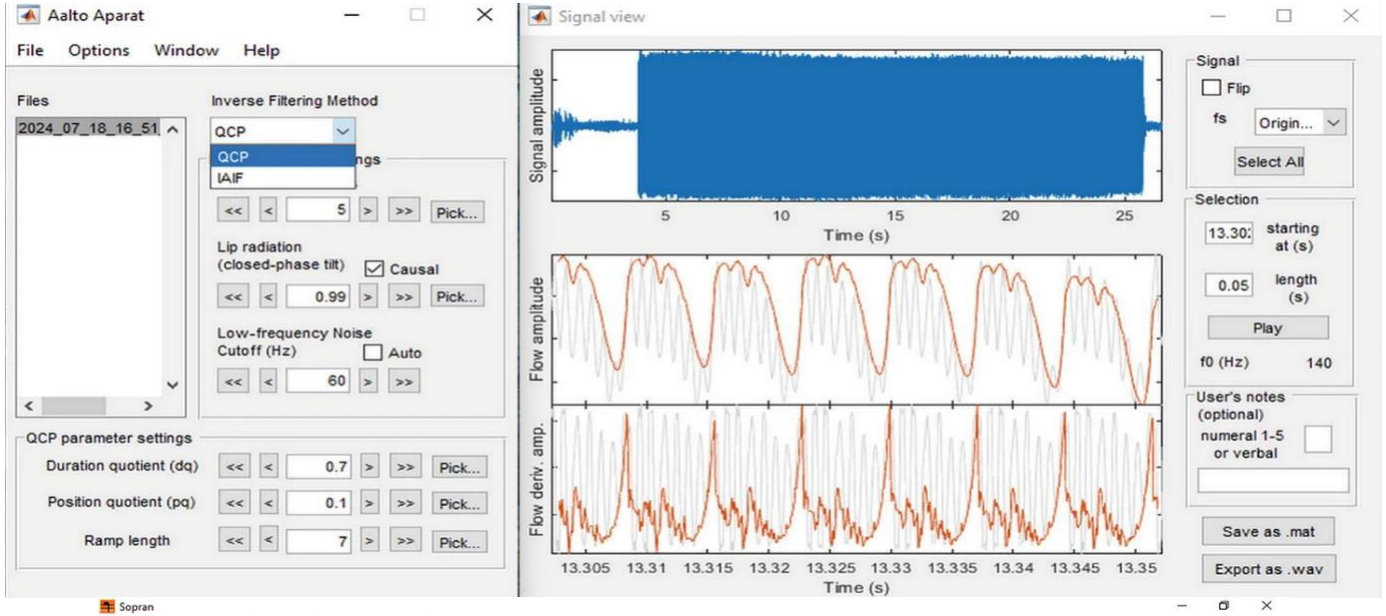
- Normal: Adult females ≥ 15 -25 sec, Adult males ≥ 25-35 sec
- Trained singers: Lower register (25-40 sec)

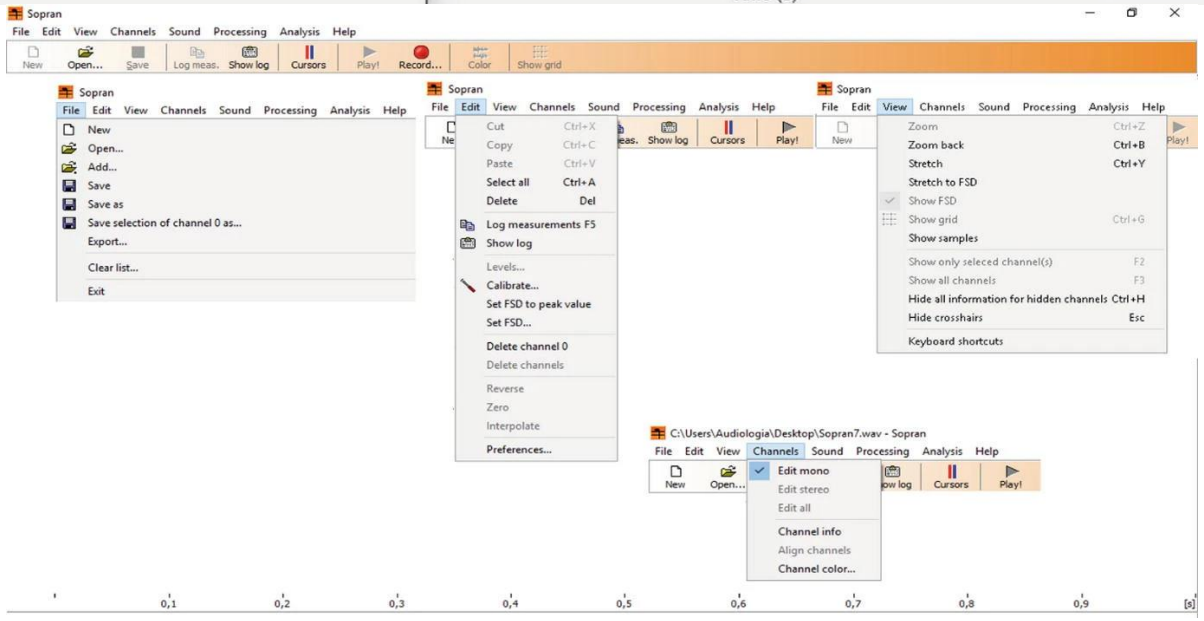
  Upper register (≥ 30 sec)
- At best: Coordinated to Harmonics to Noise ratio, and Jitter, Shimmer

- **It can be recorded in a soundproof room where audiograms are made anyway and analysed for GRBAS test, and Acoustical measures with relevant benchmark algorithms. The VHI score can be added – and an AI result can be calculated!**

Akinoğlu B, Çoban S, Shehu SU, Yilmaz AE. Investigating the Relationship Between Maximum Phonation Time of Different Vowel Sounds and Respiratory Function. J Voice. 2025 Apr 30:S0892-1997(25)00162-6. doi: 10.1016/j.jvoice.2025.04.012. Epub ahead of print. PMID: 40312191.

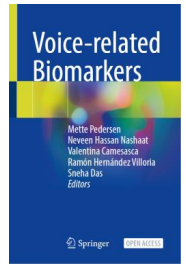
Inverse filtering apps are fine as supplements in the clinic

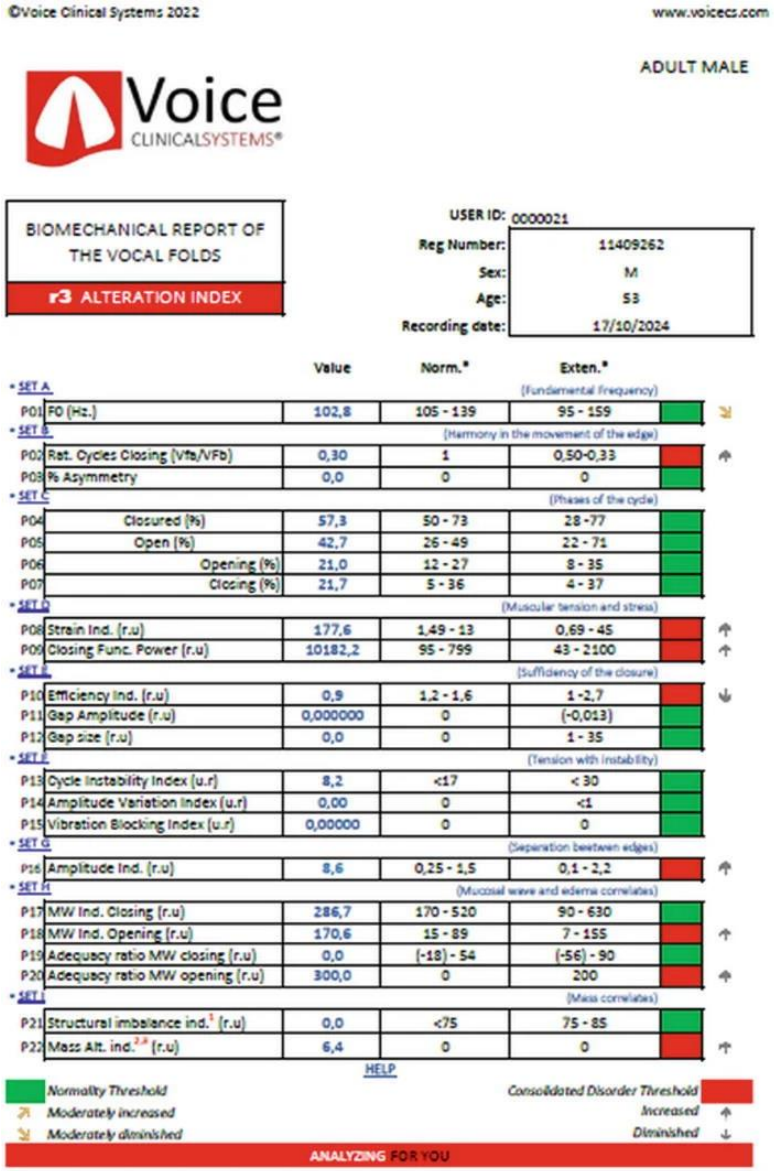




Aalto Aparat. Main window and signal view window. An audio signal.wav file has been opened from a directory. The small box contains a drop-down excerpt of the available GIF analysis methods. By default, QCP is selected. Screenshot of Aalto Aparat. © Paavo Alku et al., Aalto University

Sopran. The main window and the first four menus are displayed. Screenshot of Sopran, a freeware sound analysis tool developed by Tolvan Data. © Tolvan Data 2025. Used with permission





©Voice Clinical Systems 2022 www.voicecs.com

ADULT MALE

Voice CLINICALSYSTEMS®

BIOMECHANICAL REPORT OF THE VOCAL FOLDS

r3 ALTERATION INDEX

USER ID: 0000021

Reg Number:	11409262
Sex:	M
Age:	53
Recording date:	17/10/2024

		Value	Norm.*	Exten.*	
• SET A				(Fundamental Frequency)	
P01	F0 (Hz.)	102,8	105 - 139	95 - 159	↗↘
• SET B				(Harmony in the movement of the edge)	
P02	Rat. Cycles Closing (Vfa/VFb)	0,30	1	0,50-0,33	↑
P03	% Asymmetry	0,0	0	0	
• SET C				(Phases of the cycle)	
P04	Closured (%)	57,3	50 - 73	28 - 77	
P05	Open (%)	42,7	26 - 49	22 - 71	
P06	Opening (%)	21,0	12 - 27	8 - 35	
P07	Closing (%)	21,7	5 - 36	4 - 37	
• SET D				(Muscular tension and stress)	
P08	Strain Ind. (r.u)	177,6	1,49 - 13	0,69 - 45	↑
P09	Closing Func. Power (r.u)	10182,2	95 - 799	43 - 2100	↑
• SET E				(Sufficiency of the closure)	
P10	Efficiency Ind. (r.u)	0,9	1,2 - 1,6	1 -2,7	↓
P11	Gap Amplitude (r.u)	0,000000	0	(-0,013)	
P12	Gap size (r.u)	0,0	0	1 - 35	
• SET F				(Tension with instability)	
P13	Cycle Instability Index (u.r)	8,2	<17	< 30	
P14	Amplitude Variation Index (u.r)	0,00	0	<1	
P15	Vibration Blocking Index (u.r)	0,00000	0	0	
• SET G				(Separation between edges)	
P16	Amplitude Ind. (r.u)	8,6	0,25 - 1,5	0,1 - 2,2	↑
• SET H				(Mucosal wave and edema correlates)	
P17	MW Ind. Closing (r.u)	286,7	170 - 520	90 - 630	
P18	MW Ind. Opening (r.u)	170,6	15 - 89	7 - 155	↑
P19	Adequacy ratio MW closing (r.u)	0,0	(-18) - 54	(-56) - 90	
P20	Adequacy ratio MW opening (r.u)	300,0	0	200	↑
• SET I				(Mass correlates)	
P21	Structural imbalance ind.[1] (r.u)	0,0	<75	75 - 85	
P22	Mass Alt. ind.[2,3] (r.u)	6,4	0	0	↑

HELP

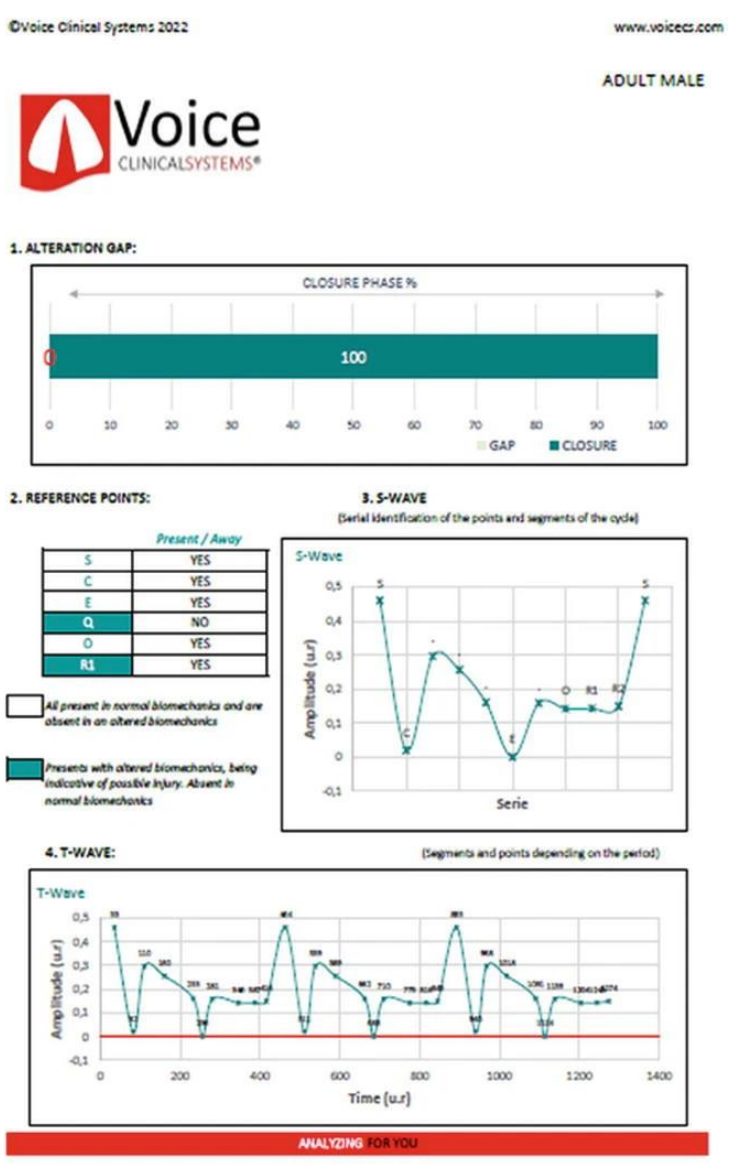
Normality Threshold — Consolidated Disorder Threshold

↗ Moderately increased — Increased ↑

↘ Moderately diminished — Diminished ↓

ANALYZING FOR YOU

4 of 9



©Voice Clinical Systems 2022 www.voicecs.com

ADULT MALE

Voice CLINICALSYSTEMS®

1. ALTERATION GAP:

2. REFERENCE POINTS:

	Present / Away
S	YES
C	YES
E	YES
Q	NO
O	YES
R1	YES

All present in normal biomechanics and are absent in an altered biomechanics

Presents with altered biomechanics, being indicative of possible injury. Absent in normal biomechanics

3. S-WAVE

(Serial identification of the points and segments of the cycle)

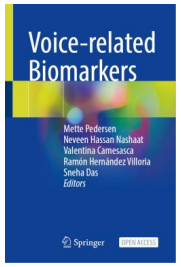
4. T-WAVE:

(Segments and points depending on the period)

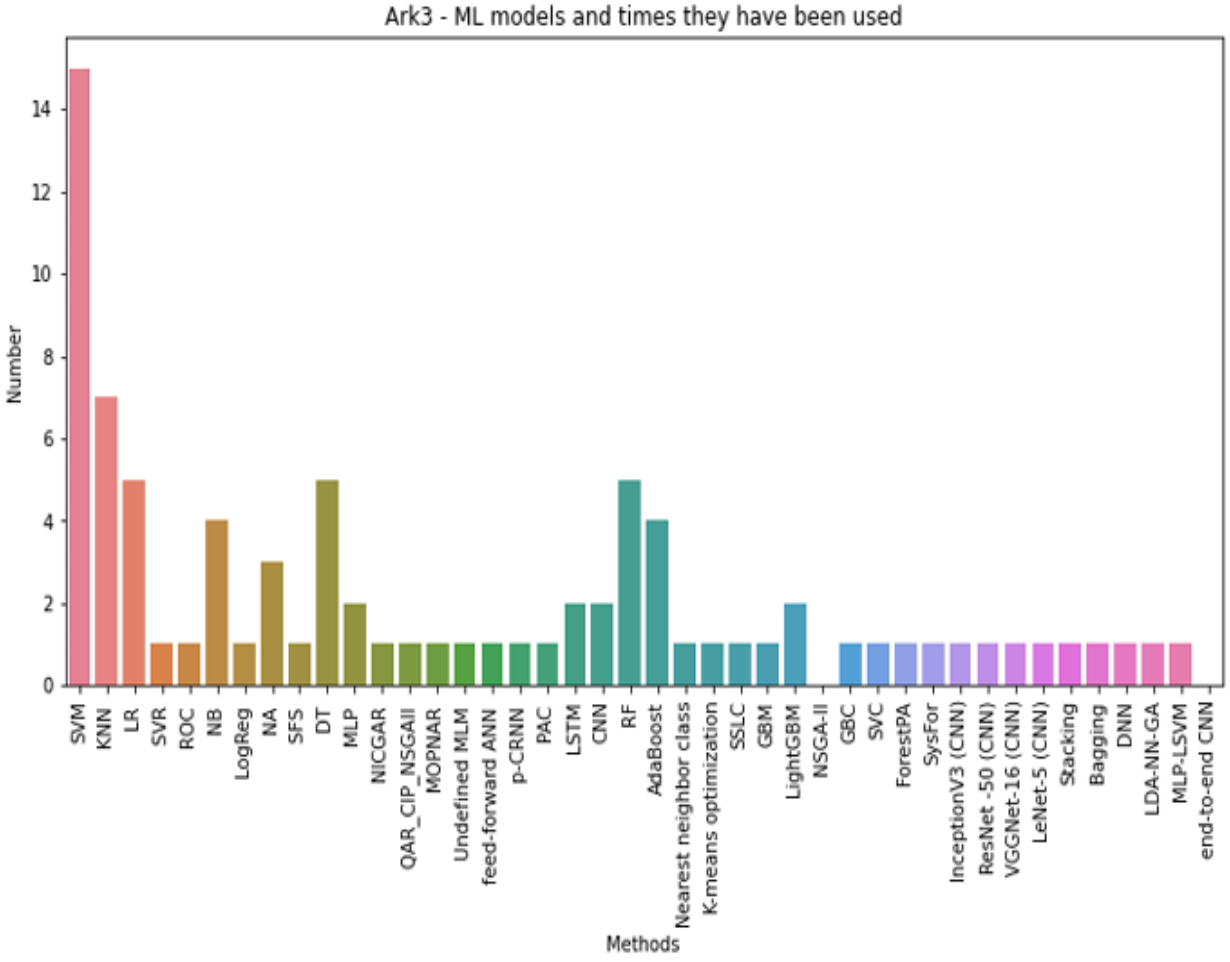
ANALYZING FOR YOU

7 of 9

Clinical test report for biomechanical analysis of the vocal folds. Online Lab VCS. Two-page example of a complete clinical test report for biomechanical analysis of the vocal folds. Screenshot from Online Lab, ©Voice Clinical Systems. Used with permission



The usage of Machine Learning models in voice analysis is growing, till now with no benchmarks.
Here is a presentation of the huge amount of algorithms used in 18 recent articles on voice analysis in Parkinson´s disease.



Reference:
Voice-related Biomarkers – ed. Springer Publishers 2026
and for further reading also Pedersen M. (2025**). Artificial Intelligence for Screening Voice Disorders: Aspects of Risk Factors : Research Article. American Journal of Medical and Clinical Research & Reviews**, 4(2), 1–8. https://doi.org/10.58372/2835-6276.1254

AI in Parkinson's Disease with methodological limitations in the 18 recent Voice-AI papers

- Articles do not provide data
- lack of demographic diversity,
- insufficient dataset size, and
- incomplete metric reporting.

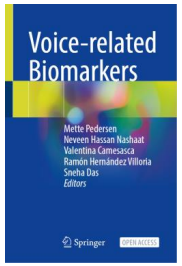
These insufficiencies limit real-world relevance.

These are the categories we looked for in 18 articles on voice in Parkinsons Disease during a recent 5 years period and what problems we identified, and how many articles have this as well how many that hasn't, including the reason for the missing data.

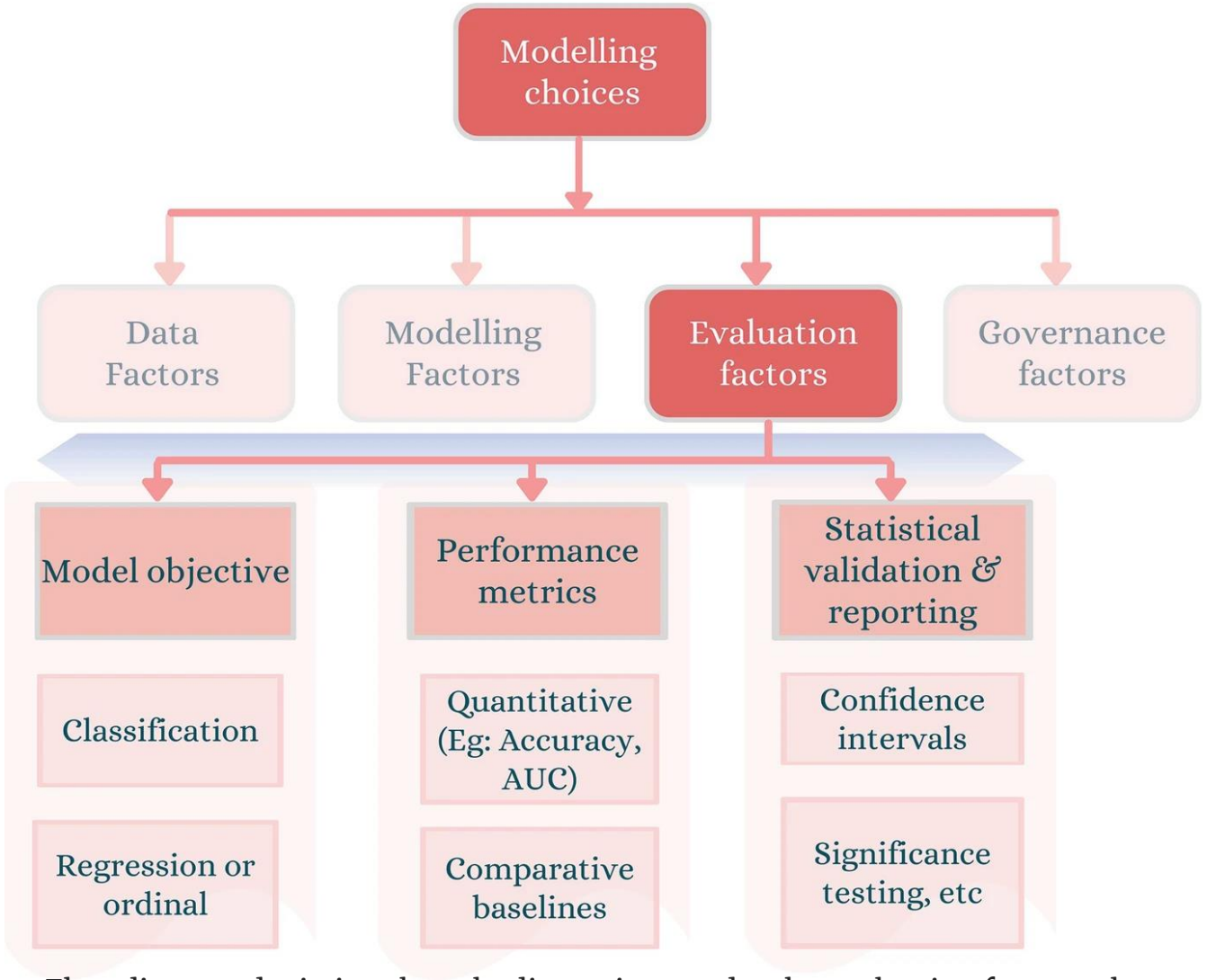
Acoustical Datasets

Category	Problems Identified	Articles That Provide Data	Articles That Do Not Provide Data	Reason for Missing Data
Usability	Insufficient dataset size	6 articles	12-13 articles	Articles focus on model performance
Precision	Variability in measurement protocols	5 articles	13-14 articles	Many articles assume standardized datasets
Content	Limited diversity in vocal tasks	7 articles	11-12 articles	Studies focus on specific phonemes
Population	Underrepresentation of demographic groups	4-5 articles	13-15 articles	Articles do not address demographic diversity
Disorder Characteristics	Limited characterization of specific vocal imparimetns	6 articles	12-13 articles	Some studies focus purely on classification

Reference: Pedersen M. (2025**). Artificial Intelligence for Screening Voice Disorders: Aspects of Risk Factors : Research Article. American Journal of Medical and Clinical Research & Reviews**, 4(2), 1–8. https://doi.org/10.58372/2835-6276.1254



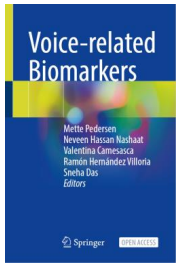
A flow diagram depicting the sub-dimensions under model choices of software algorithms



Flow diagram depicting the sub-dimensions under the evaluation factors, that influence model choices

Regression means the model predicts a continuous numerical value

Diagram illustrating the flow through a tree structure of criteria



Flow diagram depicting the sub-elements of governance factors for software algorithms

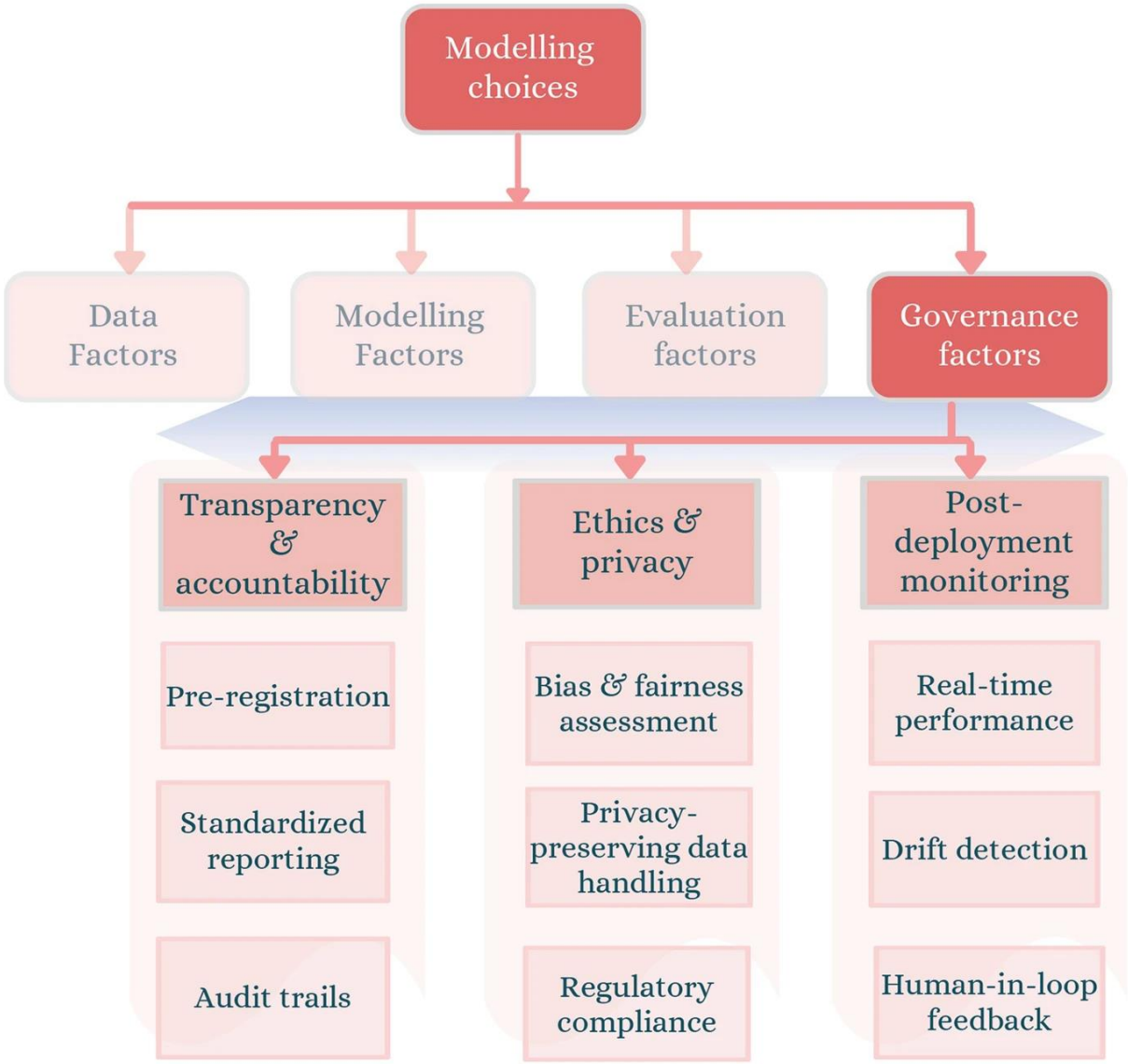
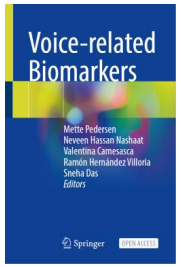


Diagram illustrating the flow through a tree structure of criteria

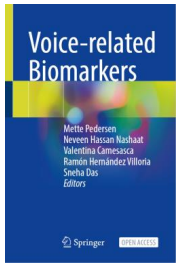
Flow diagram depicting the sub-elements of governance factors



Partitioning strategies for voice AI evaluation. Each source of variability can compromise generalization if not explicitly controlled.

Table 9.1.
The recommended strategies help ensure that evaluation outcomes reflect true model robustness and generaization rather than dataset artifacts

Source of variability	Risk if uncontrolled	Impact on evaluation	Recommended strategy
Speaker identity	Memorization of speaker traits	Inflated accuracy due to identity recognition	*Enforce speaker-level independence; exclude any voice samples from the same speaker across partitions*
Recording session	Leakage of temporal or contextual cues	Apparent performance gain from repeated recording conditions	*Partition at the session level for longitudinal data; keep sessions or days strictly separate*
Device or channel	Model learns recording setup instead of pathology	Poor cross-device generalization	*Stratify by device type or perform cross-device validation; document channel conditions*
Task or elicitation type	Task covaries with diagnosis (e.g., patients read different text)	Model captures task structure rather than disease signal	*Balance or randomize tasks across diagnostic groups; treat task as a stratification variable*
Cohort membership	Dataset-specific artifacts mistaken for generalizable features	Limited external validity; cohort bias	*Conduct intra- and inter-cohort evaluations; use chronological or site-based splits to test transferability*
Demographic or linguistic subgroup	Bias against underrepresented accents, genders, or age groups	Unfair or non-generalizable performance	*Maintain distributional fidelity and fairness via stratified sampling across demographics*
Temporal drift	Dataset drift across time	Model obsolescence post-deployment	*Apply chronological partitioning; periodically re-benchmark on new data*
Analytical leakage (double dipping)	Test data influence feature selection or preprocessing	Overestimated performance; circular validation	*Maintain strict data and pre-processing isolation; document normalization and tuning procedures*



Demands for Voice-related Biomarkers & a Foundation Model to be used in the clinic

Machine Learning evidence

The objective is: Bridging the gap between high-performance 'black-box' models and biomarkers

1. The foundation Model Shift is: Moving from handcrafted features, VHI, GRBAS test acoustical features (F0, jitter, shimmer,harmonis to noise ratio) and MPT to self-supervised learning (SSL) embeddings.
2. The challengeis are: Navigating the 'reproducibility crisis' in AI-driven speech pathology

Use methodological modes to actively guard against failures

a) Data integrity & leakage

1. Patient-level information leakage
2. Post-hoc threshold tuning on the test set

b) Bias and confounding

1. Measuring bias
2. Confounding by language, smoking, age, sex, or device type

c) Opacity in pipeline

1. Inadequate reporting of preprocessing, segmentation, or voice activity detection

And have a minimal checklist for a clinically credible study

a) Context and Acquisition

1. Intended use + clinical reference standard explicitly stated
2. Data Acquisition protocol described sufficiently for replication

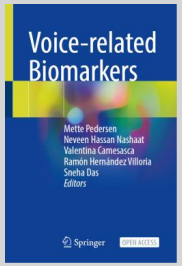
b) Rigorous Evaluation

1. Patient level splits; leakage checks; external validation with site-stratified reporting
2. Performance: discrimination + calibration + clinical decision curves where relevant
3. Subgroup analysis and fairness consideration
4. Error analysis with representative failure cases; uncertainty estimates

c) Interpretation & Governance

1. Dataset governance: consent, access model, and privacy-preserving release strategy

From the Medical Center Østergade 18, Copenhagen 1100 IT-University, Rued Langgaards Vej 7, Copenhagen 2300 Technical University of Denmark, Anker Engelunds Vej 1, Bygning 101A, Kongens Lyngby 2800 https://www.mettepedersen.org/presentations



Roadmap and recommended worked -through suggested project –based on the book

Phase 1:
Harmonization

- **Multicenter Dataset**: Curate and harmonize a high-fidelity dataset following European Union of Phoniatrics (UEP) standards.
- **Unified Tasks**: Sustained phonation, and eventual ½connected speech (standardized passages).
- **Rich Metadata**: Linked Patient-Reported Outcome Measures (PROMs,VHI) and eventual expert stroboscopy labels (ground truth).
- **Standardization**:Dataset Standard for voic-related e biomarkers to guide international collaboration and data exchange.

Phase 2:
The Foundation Model Benchmark

- **Compare pretrained speech models** (wav2vec2/HuBERT/WavLM/HeAR) to determine which self-supervised learning (SSL) embeddings are most clinically useful for phoniatric classification and prediction across datasets.

Phase 3:
Multimodal models

- **Multimodal Integration**: Move beyond 'voice-only' models. Integrate audio embeddings with endoscopic video features and PROMs,VHI.
- **Uncertainty Reporting**: Implement Bayesian or conformal prediction methods to report uncertainty (e.g: 'Model confidence: 65% - recommend human stroboscopy").

Phase 4:
Robustness & Fairness Evaluation

- **Subgroup Robustness**: Rigorous calibration analysis across sex, language/dialect, and recording device (e.g: studio mic vs. smartphone).
- **Error Taxonomy**: Publish a formal taxonomy of failure modes (e.g: technical noise vs. physiological mimics).
- **Fairness Calibration**: Ensure probability estimates are clinically reliable across diverse demographic groups.
- **Prospective validation in real-world clinical practice**, following SPIRIT-AI and CONSORT-AI standards to ensure rigorous study design, transparent reporting, clear description of model inputs and outputs, and robust evaluation of clinical utility and safety.

From the Medical Center Østergade 18, Copenhagen 1100 IT-University, Rued Langgaards Vej 7, Copenhagen 2300 Technical University of Denmark, Anker Engelunds Vej 1, Bygning 101A, Kongens Lyngby 2800 https://www.mettepedersen.org/presentations

Thank you for listening!