# A Database of Laryngeal High-Speed Videos with Simultaneous High-Quality Audio Recordings of Pathological and Non-Pathological Voices

**P. Aichinger, I. Roesner, M. Leonhard, D.M. Denk-Linnert, W. Bigenzahn, B. Schneider-Stickler**

Division of Phoniatrics-Logopedics, Department of Otorhinolaryngology, Medical University of Vienna, Austria
Waehringer Guertel 18-20, 1090 Vienna, Austria
E-mail: philipp.aichinger@meduniwien.ac.at

**Abstract**

Auditory voice quality judgements are used intensively for the clinical assessment of pathological voice. Voice quality concepts are fuzzily defined and poorly standardized however, which hinders scientific and clinical communication. The described database documents a wide variety of pathologies and is used to investigate auditory voice quality concepts with regard to phonation mechanisms. The database contains 375 laryngeal high-speed videos and simultaneous high-quality audio recordings of sustained phonations of 80 pathological and 40 non-pathological subjects. Interval wise annotations regarding video and audio quality, as well as voice quality ratings are provided. Video quality is annotated for the visibility of anatomical structures and artefacts such as blurring or reduced contrast. Voice quality annotations include ratings on the presence of dysphonia and diplophonia. The purpose of the database is to aid the formulation of observationally well-founded models of phonation and the development of model-based automatic detectors for distinct types of phonation, especially for clinically relevant nonmodal voice phenomena. Another application is the training of audio-based fundamental frequency extractors on video-based reference fundamental frequencies.

**Keywords:** pathological voice, dysphonia, diplophonia, laryngeal high-speed videos

## 1. Introduction

Auditory voice quality judgements are used intensively for the clinical assessment of pathologic voice. These judgements aid the indication, selection, evaluation and optimization of clinical treatment techniques. Voice quality types that are typically described include dysphonia, diplophonia and euphonia. The term dysphonia includes all types of auditory divergences from normal voice quality. Diplophonia is a concept subordinate to dysphonia, i.e. the simultaneous presence of two pitches in the voice sound. The term euphonia is the antonym to dysphonia. Voice quality concepts are fuzzily defined and poorly standardized however, which hinders scientific and clinical communication and updates of terminology are needed (Gerrat & Kreiman, 2001). These problems are tackled by using a data corpus including voice quality annotations, which aids the formulation of observationally well-founded models of voice production and perception, and consequently updates in terminology. We propose to observe voice production by means of laryngeal high-speed imaging with simultaneous high-quality audio recordings that enable acoustic and perceptual analyses. In contrast to what has often been practiced in the field of disordered voice research (Olszewski, Shen & Jiang, 2011), video and audio quality criteria that allow for reproducible and non-destructive pre-selection of voice samples are used. State-of-the-art audio equipment and acquisition strategies from the scientific field of audio engineering are deployed into the field of clinical voice research. Professional condenser microphones with symmetrical wiring and a high-quality recorder are used with a sampling rate of 48 kHz and a quantisation resolution of 24 bits. Interval wise voice quality annotations accompany subject global ratings.

## 2. Data collection

One hundred and twenty subjects of three clinical voice quality groups representing subject global ratings are included in the database. The subject groups were diplophonic dysphonic, non-diplophonic dysphonic and euphonic. Eighty dysphonic subjects were recruited among the outpatients of the Medical University of Vienna, Department of Otorhinolaryngology, Division of Phoniatrics-Logopedics. Forty of these subjects were diplophonic dysphonic and forty were non-diplophonic dysphonic. The presence of diplophonia and dysphonia was determined by medical doctors specialised on voice disorders, i.e. phoniatricians, by perceptual screening. The number of pathological subjects and their diagnoses with regard to voice quality groups are given in table 1. Additionally forty subjects that constitute the euphonic group were recruited via public announcement in Vienna. The data collection had been approved by the institutional review board of the Medical University of Vienna.

A laryngeal high-speed camera and a portable audio recorder were used for data collection. The camera was a HRES ENDOCAM 5562 (Richard Wolf GmbH) and the audio recorder was a TASCAM DR-100. The frame rate of the laryngeal high-speed camera was set to 4 kHz. The CMOS camera sensor had three colour channels, i.e. red, green and blue. Its spatial resolution was 256x256 (interpolated from red: 64x128, green: 128x128, blue 64x128). The light intensity values were quantized with a resolution of 8 bits. The videos are 2.048 seconds long each. 132 videos were taken from the euphonic group, 123 from the diplophonic dysphonic group and 120 from the non-diplophonic dysphonic group.

| | Diplophonic | Non-diplophonic |
|---|---|---|
| **Laryngitis (acute/chronic)** | 2 | 8 |
| **Sulcus** | 2 | 2 |
| **Nodules** | 1 | 3 |
| **Polyp** | 3 | 2 |
| **Oedema** | 6 | 5 |
| **Cyst** | 4 | 2 |
| **Scar** | 1 | 2 |
| **Paresis** | 13 | 2 |
| **Dysfunction** | 5 | 12 |
| **Benign tumour** | 1 | 0 |
| **Bamboo nodes** | 1 | 0 |
| **Neurological** | 0 | 1 |
| **Unknown** | 1 | 1 |

Table 1: Number of pathological subjects with respect to diagnoses and voice quality groups.

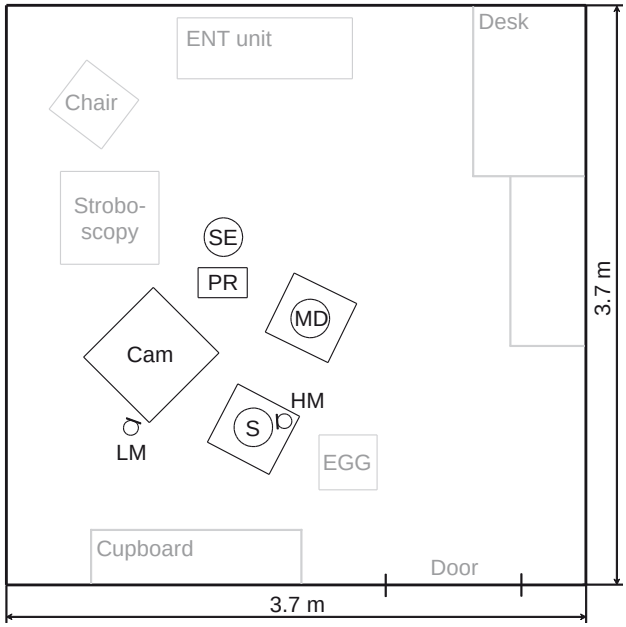


Figure 1: Footprint of the recording room. Ear, nose and throat (ENT) unit, sound engineer (SE), portable audio recorder (PR) (TASCAM DR-100), laryngeal high-speed camera (Cam) (HRES ENDOCAM 5562, Richard Wolf GmbH), medical doctor (MD), headworn microphone (HM) (AKG HC 577 L), subject (S), lavalier microphone (LM) (AKG CK 77 WR-L), electroglottograph (EGG).

Two microphones were used for recording audio signals. A headworn microphone AKG HC 577 L was used to record the subjects' voices. The loudest noise source in the room was the cooling fan of the camera's light source. A lavalier microphone AKG CK 77 WR-L was put next to the fan to record its noise. Both microphones were used with the original cap (no presence boost) and with windscreens AKG W77 MP. The microphones were connected to phantom power adapters AKG MPA V L (linear response setting) and the portable audio recorder (headworn: left channel, lavalier: right channel). The sampling rate was 48 kHz and the quantization resolution was 24 bits. The subjects' recording sessions mostly consisted of one to six videos and were entirely microphone recorded. The audio recordings were carried out by the author PA. The sound quality was continuously monitored during the recordings with AKG K 271 MK II headphones. The audio files are saved in the uncompressed PCM/WAV file format.

Figure 1 shows the footprint of the recording room, which is a standard room for ENT examinations. The figure shows the subject and the medical doctor who were sitting on chairs, the sound engineer, the high-speed camera, the lavalier microphone, the headworn microphone and the portable recorder. The background noise was 48.6 dB(A) and 55.5 dB(C), measured with a PCE-322A sound level meter (IEC 61672-1 class 2), with temporal integration set to "slow". The tip of the tongue was lightly held by the medical doctor when she inserted the endoscope into the mouth of the subject way back to the pharynx. The larynx was illuminated and filmed by the endoscopic camera and the medical doctor previewed the camera pictures on a computer screen. She adjusted the position of the camera for optimal sight of the vocal folds. Once an optimal position was achieved the medical doctor instructed the subject to phonate an [i], which positioned the epiglottis so as to allow direct sight of the vocal folds. The produced sounds were schwa-like, due to the endoscope and the lowered position of the tongue.

## 3. Data annotations

The collected data were annotated for video quality, voice quality, and audio quality.

### 3.1 Video quality

The laryngeal high-speed videos were annotated with regard to video quality. The visibility of relevant structures and the presence of artefacts were evaluated. The videos were annotated with the Anvil tool (Kipp, 2007), which allows for placing time interval tiers considering time-variant video quality.

The visibility of the vocal fold edges and vocal fold vibration is important for facilitating analyses like, for instance, spectral video analysis, glottal area waveform extraction or videokymography. Parts of the vocal fold edges can be hidden behind the epiglottis, the epiglottis' tubercle, the false vocal folds, the aryepiglottic folds or the arytenoid structures. The criteria used for judging the visibility of the glottal gap were the visibility of the anterior commissure and the visibility of the processus vocales. The processus vocales act as nodes in the vocal fold oscillation modes, and thus are adequate for orientation in the anatomy when oscillation patterns are observed and interpreted.

The presence of blurring and/or reduced contrast, false vocal fold artefacts, mucus that blurs the vocal fold edges and a remote mucus yarn may impede video analysis. Blurring occurred if the camera was not focused on the glottal gap, or if there were fluid artefacts on the endoscope. A false vocal fold artefact occurred when the

| Criteria | (Labels) classes |
|---|---|
| **Visibility** | |
| Vocal fold edges | (0) Not visible<br>(1) Visible |
| Vocal fold vibration | (0) Not visible<br>(1) Visible |
| Anterior commissure | (0) Not visible<br>(1) Visible |
| Processus vocales | (0) None visible<br>(1) One visible<br>(2) Both visible |
| **Artefacts** | |
| Blurring | (0) Absent<br>(1) Mild<br>(2) Severe |
| Reduced contrast | (0) Absent<br>(1) Mild<br>(2) Severe |
| Mucus on the vocal folds | (0) Not visible<br>(1) Visible<br>(2) Blurring the vocal folds edges<br>(3) Connecting the vocal folds |
| Extraglottal mucus yarn artefact | (0) Absent<br>(1) Present<br>(2) Blurring the vocal folds edges |
| False vocal fold artefact | (0) Absent<br>(1) Hiding the vocal fold edges |
| Aryepiglottic folds artefact | (0) Absent<br>(1) Hiding the vocal fold edges |

Table 2: Overview of the video quality annotation criteria.

vocal fold edges were partly hidden behind the false vocal folds. Reduced contrast occurred for several reasons. It was not always possible to illuminate the larynx sufficiently due to inter-individual anatomical differences, for instance, the epiglottis was not fully lifted and in subjects with lowered larynx the light reaching the vocal folds was weaker. Blurring and reduced contrast are labelled as absent, mild or severe. Table 2 lists the used criteria, the labels and the classes.

### 3.2 Voice quality

Author PA has annotated all phonated video synchronous audio recordings for voice quality in agreement with author IR. In contrast to clinical group allocation in terms of subject global voice quality, the annotation aims at assessing voice quality interval wise. The used voice quality labels and classes are (1) euphonic, (2) diplophonic dysphonic and (3) non-diplophonic dysphonic. Diplophonia was defined as the simultaneous presence of two different pitches or a distinct impression of beating. Simple pitch breaks were not considered to be diplophonic. When the perceptual determination of the presence of diplophonia was doubtful, the audio waveforms and spectrograms were visually inspected. The spectral criterion was the presence of two separate harmonic series. The waveform criterion was the presence of metacycles, i.e. diplophonic beating. The audio signal only was available to the annotator.

| Criteria | (Labels) classes |
|---|---|
| **Auditory assessment** | |
| Phonation | (0) Absent<br>(1) Present |
| Group | (0) Absent<br>(1) Euphonic<br>(2) Diplophonic dysphonic<br>(3) Non-diploph. dysphonic |
| **Audio artefacts** | |
| Examiner's voice | (0) Absent<br>(1) Present |
| Other artefacts | (1) Present<br>(0) Absent |
| **Video relation** | |
| sync | (Video ID) - |
| Video | (Video ID) Present<br>(0) Absent |

Table 3: Overview of the audio annotation criteria.

### 3.3 Audio quality

Audio quality was evaluated with respect to the presence of artefacts and synchronizability. Parts of some recordings contain the medical doctor's voice giving instructions to the subject and only the remaining parts can be used for audio analyses. Other artefacts are background noise or stem from microphone contact. Artefacts were successively minimized as they were recognized during data collection, but were not fully avoidable. The audio files were synchronized to the video by visually matching their waveforms to the waveforms of the audio that was recorded with the camera's inbuilt microphone. Table 3 lists the used criteria, the labels and the classes.

## 4. Discussion & Conclusion

Added values and limitations of the database are listed and an outlook is given.

One added value is that the creation of the described database stimulated interdisciplinary communication between medical doctors and engineers. The obtained expert knowledge regarding voice quality and data quality is represented in the database by the annotations, which enable non-destructive selection of subcorpora. Other added values are twofold. First, voice quality is assessed interval wise instead of subject globally. Second, the audio quality exceeds that of most other work in the field of clinical voice research.

However, two limitations exist and should be taken into account when the data are analysed and conclusions are drawn. First, representativeness of the data for spoken language maybe limited. In addition, the recording and playback conditions were not perfect, because experimental conditions in a hospital are similar to field conditions rather than laboratory conditions.

Future work may focus on addressing the limitations of the database. To increase representativeness and ecological validity, studies on the prevalence of different voice quality types in spoken language of the general population are needed, and less obstructive methods of observation should be found. Perceptual effects in the clinical environment need further investigation with calibrated recording and playback setups that account for different room acoustics. It is also important to investigate what data quality is sufficient for what kind of analysis. Moreover, efficient data handling tools should be established.

A state-of-the-art database of laryngeal high-speed videos with synchronous high-quality audio recordings was created successfully. The database has proven to be highly valuable for answering clinically relevant research questions. Research results based on the described database have already been published and more are on the way (Aichinger, 2015, Aichinger et al., 2013; 2013a; 2015; 2015a; 2015b; 2015c; accepted; in Press; Schenk et al., 2014; 2015).

## 5. Acknowledgements

The authors would like to thank Jean Schoentgen for comments on the text and Richard Wolf GmbH for providing the camera.

## 6. Duplication remark

Parts of the text were duplicated from the first author's PhD thesis (Aichinger, 2015).

## 7. References

Aichinger, P., Schneider-Stickler, B., Bigenzahn, W., Fuchs, A.K., Geiger, B., Hagmüller, M. and Kubin, G. (2013). Double pitch marks in diplophonic voice, In *Proceedings of the 38th IEEE International Conference on Acoustics, Speech, and Signal Processing*, Vancouver, Canada, pp. 7437--7441.

Aichinger, P., Roesner, I., Schneider-Stickler, B., Bigenzahn, W., Feichter, F., Fuchs, A.K., Hagmüller, M., Kubin, G. (2013a). Spectral analysis of laryngeal high-speed videos: case studies on diplophonic and euphonic phonation. In *Proceedings of the 8th International Workshop on Models and Analysis of Vocal Emissions for Biomedical Applications*, Firenze, Italy, pp. 81--84.

Aichinger, P. (2015). Diplophonic Voice – Definitions, models, and detection. PhD thesis, Graz University of Technology, Austria.

Aichinger, P., Hagmüller, M., Roesner, I., Bigenzahn, W., Schneider-Stickler, B., Schoentgen, J. and Pernkopf, F. (2015a). Measurement of fundamental frequencies in diplophonic voices. In *Proceedings of the 9th International Workshop on Models and Analysis of Vocal Emissions for Biomedical Applications*, Firenze, Italy, pp. 21--24.

Aichinger, P., Hagmüller, M., Roesner, I., Bigenzahn, W., Schneider-Stickler, B. and Schoentgen, J. (2015b). Differentiating diplophonia from other types of severe dysphonia by acoustic analysis. In *Pan European Voice Conference Abstract Book*, Firenze, Italy, p. 34.

Aichinger, P., Schneider-Stickler, B., Bigenzahn, W., Hagmüller, M., Sontacchi, A. and Schoentgen, J. (2015c). Assessment and psychoacoustic modelling of auditory streams in diplophonic voice. In *Proceedings of the 9th International Workshop on Models and Analysis of Vocal Emissions for Biomedical Applications*, pp. 135--138.

Aichinger, P., Roesner, I., Leonhard, M., Schneider-Stickler, B., Denk-Linnert, D.M., Bigenzahn, W., Fuchs, A.K., Hagmüller, M. and G. Kubin. (accepted). Towards objective voice assessment: the diplophonia diagram. *Journal of Voice*.

Aichinger, P., Roesner, I., Leonhard, M., Schneider-Stickler, B., Denk-Linnert, D.M., Bigenzahn, W., Fuchs, A.K., Hagmüller, M. and G. Kubin. (in Press). Comparison of an audio-based and a video-based approach for detecting diplophonia. *Biomedical Signal Processing and Control*.

Gerratt, B.R. and Kreiman J. (2001). Toward a taxonomy of nonmodal phonation. *Journal of Phonetics*, 29(4), pp. 365--381.

Kipp, M. (2007). Anvil: The video annotation research tool. [Online]. Available: www.anvil-software.org [Accessed: April 7, 2014]

Lohscheller, J., Toy, H., Rosanowski, F., Eysholdt, U. and Döllinger, M. (2007). Clinically evaluated procedure for the reconstruction of vocal fold vibrations from endoscopic digital high-speed videos. *Medical Image Analysis*, 11(4), pp. 400--413.

Olszewski, A., Shen, L. and Jiang, J. (2011). Objective methods of sample selection in acoustic analysis of voice. *The Annals of Otology, Rhinology, and Laryngology*, 120(3), pp. 155--161.

Schenk, F., Urschler, M., Aigner, C., Roesner, I., Aichinger, P. and Bischof, H. (2014). Automatic glottis segmentation from laryngeal high-speed videos using 3D active contours. In *Proceedings of 18th Conference on Medical Image Understanding and Analysis*, Egham, UK, pp. 111--116.

Schenk, F., Aichinger, P., Roesner, I. and Urschler, M. (2015). Automatic high-speed video glottis segmentation using salient regions and 3D geodesic active contours. *Annals of the British Machine Vision Association*, 2015(1), pp. 1--15.