forum acusticum 2026
8-12 September · Graz, Austria

SUBJECTIVE AND OBJECTIVE EMPLOYABILITY FOR DEGRADED SPEECH IN VIDEO-MEDIATED JOB INTERVIEWS

Julia Schütze[1], Thomas Biberger[2,3], Thomas P. Reber[1]

[1] *Department of Psychology, UniDistance Suisse, Switzerland*
[2] *Medical Physics and Cluster of Excellence Hearing 4 All, Carl von Ossietzky Universität Oldenburg, Germany*
[3] *Otolaryngology-Head and Neck Surgery, University Medical Center Hamburg-Eppendorf (UKE), Hamburg, Germany*

This paper is a revised version of the authors' submitted manuscript that was peer-reviewed by at least two anonymous reviewers.

Abstract

The quality of audio transmission in video conferencing systems can influence not only speech perception but also higher-level social judgments in communication contexts such as hiring decisions following job interviews. The present study investigated how degradations in speech quality affect perceived employability and whether objective audio quality models capture these effects.

The audio streams of AI-generated job interview videos were manipulated to create three levels of audio quality (high, medium, low). Participants evaluated the employability of the speakers, while the corresponding audio signals were analyzed using two objective audio quality models, ViSQOL and GPSM^{q}. Results showed a clear effect of audio quality on employability ratings, with substantially lower ratings in the low-quality condition compared to medium and high quality, and no significant difference between medium and high quality. ViSQOL reflected these differences by assigning lower scores to degraded signals, whereas GPSM^{q} showed limited sensitivity, including a floor effect in the medium-quality condition.

Correlation analysis revealed only weak associations between model outputs and employability ratings, indicating that current audio quality models capture only a limited portion of the variance in social judgments.

Overall, the findings highlight that audio quality contributes to employability judgments in video conferencing contexts, but does not fully account for them.

Keywords: *audio quality, auditory modeling, video conferencing*

***Corresponding author**: julia.schuetze@fernuni.ch.*

1. Introduction

Video conferencing systems rely on the transmission of speech signals that are often subject to degradation due to compression, bandwidth limitations, or artifacts due to signal processing for speech enhancement such as acoustic echo cancelation. Such degradations can substantially alter the acoustic properties of speech, including spectral detail and temporal envelope structure, which are known to be critical for speech perception. While the perceptual consequences of these degradations have been extensively studied in terms of speech quality and intelligibility, their impact on higher-level evaluations in realistic communication scenarios remains less well understood. Previous work has shown that variations in audio quality influence listener judgments beyond basic perceptual attributes. For example, Newman and Schwarz [1] demonstrated that identical spoken content is evaluated more positively when presented with higher audio quality. Similarly, Bild et al. [2] reported that degraded speech signals reduce the perceived credibility and reliability of speakers. In applied communication contexts such as job interviews, Walter-Terrill [3] found that speech quality affects evaluations of employability and related social attributes. Brucks et al. [4] investigated the effect of glitches in audiovisual virtual interaction and reported that these glitches damaged interpersonal judgements, as they evoke an 'uncanniness'. These findings suggest that signal degradations can propagate to higher-level judgments, emphasizing the importance of accurately characterizing perceptual audio quality. From a psychological perspective, such effects are often discussed under the concept of processing fluency, referring to the ease with which information can be processed [5] although only more recent work has begun to examine degraded audio quality as a potential source of such effects.

In addition to audio quality, employability-related judgments may also depend on vocal and prosodic

cues of the speaker. Prior work has shown that job-interview performance can be modeled from speaking style, vocal tone, prosody, fluency, and other paralinguistic features [6], [7]. This suggests that employability judgments are shaped not only by how degraded a signal is, but also by how the speaker's voice conveys confidence, competence, and interpersonal presence. A large body of work has focused on the development of objective audio quality models that aim to predict perceptual listening quality. While objective models are designed to approximate perceptual quality judgments, it remains unclear to what extent they capture those aspects of the signal that drive higher-level social evaluations. Objective speech quality models provide a means to quantify audio degradations based on signal characteristics. Reference-based approaches, such as the Virtual Speech Quality Objective Listener (ViSQOL; [8]), or the Generalized Power Spectrum Model for audio quality (GPSM^q, [9]) estimate perceptual similarity between degraded and reference signals. Despite the availability of these models, their relationship to behavioral evaluations in ecologically valid audiovisual communication scenarios is not well established. Most studies on objective quality assessment rely on controlled listening experiments, whereas applied studies on communication outcomes typically do not incorporate psychoacoustically grounded quality metrics.

The present study aims to bridge this gap by combining controlled audio degradation with instrumental audio quality measures and behavioral evaluation. This paper reports a subset of results from a larger study comprising multiple experimental manipulations. To maintain focus, only the audio-related conditions are considered here. Speech signals from simulated video interview scenarios were systematically degraded across multiple quality levels and analyzed using ViSQOL and GPSM^q. In parallel, listeners evaluated the corresponding stimuli with respect to employability. This approach allows us to examine whether objective auditory quality metrics capture signal degradations that are relevant not only for perceptual quality assessment but also for higher-level evaluations in video-mediated communication.

2. Methods

2.1. Participants

In total, 299 participants (157 female, 141 male, 1 divers, mean age = 40.4 years, SD = 14.0 years) took part in the experiment. Although the experiment was conducted online, the experimenters were physically present in the same room as the participants. Participants had self-reported normal-hearing and had sufficient knowledge of German. All participants gave informed consent by checking off a tick box in an online form before they proceeded to the experiment. Participants were recruited by the students in a class in autumn 2025 at UniDistance Suisse. Participation was voluntary and without compensation.

2.2. Stimulus Material and Audio Manipulation

The stimulus material consisted of avatar videos, in which the avatars presented themselves as if they were in a job interview for a senior sales manager position.

2.2.1. Generating job interview videos

A total of 27 avatars (14 male, 13 female avatars) were generated using the AI video generator software HeyGen (Avatar IV, 2025) with a resolution of 1080x1920 pixels and 25 frames per second (see section 'high quality' in Table 1 for details). The videos had an average duration of 34 seconds. We translated the original Text by Walter-Terril [3] to German, and created ten slightly different statements with a highly similar content. In text a candidate presents themselves and explains why they are a good fit for the company. Because the stimulus set included different AI-generated voices, the avatars likely differed in baseline prosodic characteristics such as pitch range, speech rate, intensity dynamics, articulation clarity, and voice quality.

2.2.2. Experimental Design

The experiment implemented a 3x3 design with the factors audio-quality (high, medium, low) and video quality (high, medium, low). For the scope of the current work, we only analyze effects of the factor audio quality and set video quality constant to high quality. Data with low and medium video quality are not considered. All different levels of audio quality were varied within the same participant: Each participant saw videos in all three combinations of the two factors. The assignment of videos and order was randomized, to not have a confounding of specific avatar characteristics with the effects of the experimental manipulation. The dependent variables in the experiment were the rating of employability of the avatar on a scale ranging from 0 (very unlikely) to 100 (very likely).

2.2.3. Audio Manipulation

The audio quality of the videos was manipulated in three different levels (high, medium, low). To create these different levels of audio manipulation, the bitrate was reduced for the medium- and low-quality presentations (see Table 1). For the low-quality videos, audio signals were further degraded by using a bitcrusher and band-pass filter in GarageBand. A band-pass type equalization was applied by combining a high-pass and a low-pass filter. Frequencies below approximately 200–300 Hz were strongly attenuated, while frequencies above approximately 6–8 kHz were also substantially reduced. The passband was centered in the mid-frequency range (approximately 800 Hz to 3 kHz), which contains the most relevant spectral information for speech intelligibility. The sampling rate of the audio signals was 48 kHz. The different videos had a duration between 24 and 44 seconds and an average duration of 34 seconds.

The video quality in these videos was high, as the resolution was 1080x1920 pixels, with more than 25

Table 1: Audio bitrate for the three different audio manipulation categories low, medium and high quality.

Quality	Low	Medium	High
Audio bitrate (kbps)	37	281	2326

frames per second, 24 bits per pixel and 'RGB' video format.

2.2.4. Measurement Procedure

The experiment was conducted within a university course in autumn 2025 at UniDistance Suisse, with students acting as experimenters under supervision. As a result, audio presentation relied on either the participants' or the experimenters' headphones, with playback levels individually adjusted to a comfortable listening level. The experiment was set up using lab.js [10] and included a short demographic questionnaire. Participants watched each avatar video individually, before they rated the likeliness of employability of that very avatar on a scale of 0 – 100.

2.3. Auditory Quality Modelling

Two audio quality models were applied to the videos in order to obtain additional objective ratings.

2.3.1. ViSQOL

Virtual Speech Quality Objective Listener (ViSQOL; [8]) was used to give an objective measure across the two manipulated levels of audio quality for each of the 27 used videos. ViSQOL is a signal-based tool, which uses a reference and a manipulated signal, to analyze the manipulated audio on the basis the Mean Opinion Score (MOS) estimate. The 5-point MOS-scale (excellent, good, fair, poor, bad) has been used to quantify perceived quality of media [11]. The ViSQOL was applied using MATLAB R2025b (The MathWorks). ViSQOL compares auditory-inspired time–frequency representations of reference and degraded signals using a patch-based similarity measure, which is subsequently mapped to a MOS-Listening Quality Objective score, thereby capturing spectro-temporal distortions relevant for human auditory perception.

2.3.2. GPSM^q

The GPSM^q [9] extends the GPSM [12], which predicts basic aspects of spectral and temporal masking and was shown to account for the results of various psychoacoustic and speech intelligibility experiments, to audio quality predictions. The GPSM^q combines power- and envelope-domain signal-to-noise ratios (SNR_{DC} and SNR_{AC}) to the Objective Perceptual Measure (OPM) to capture perceptually relevant distortions introduced by signal processing algorithms or sound reproduction systems. Besides OPM, which includes lower and upper perceptual limits (see Equation 6 in [9]), GPSM^q also provides the output measure OPM_raw without perceptual limits.

2.3.3. Correlation Analysis

Correlations were computed at the audio-stimulus level, using employability ratings averaged per stimulus across participants for the low- and medium-quality conditions (N = 54; high quality excluded as reference).

3. Results

3.1. Ratings of employability

The means of employability ratings as a function of audio quality (low, medium, high) are depicted in Figure 1. Paired-sample t-tests were conducted to compare the three audio quality conditions. A Bonferroni correction was applied to account for multiple comparisons. The results revealed significant differences between low and medium ($p < 0.001$) and between low and high ($p < 0.001$). In contrast, no significant difference was found between medium and high ($p = 0.813$). Employability ratings are lowest in the low-quality condition, increases substantially in the medium-quality condition and remains at a comparable level in the high-quality condition. The largest difference is observed between the low and medium quality conditions, whereas the difference between medium and high quality appears minimal. Post hoc comparisons indicate that the increase from low to medium audio quality is statistically significant, while no significant difference is observed between medium and high quality.

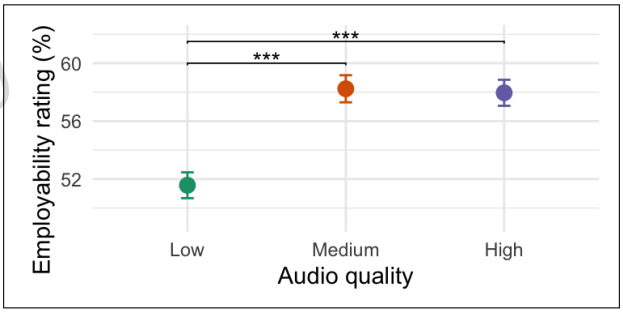

Figure 1: Participants ratings of employability for the three different levels of audio quality. Error bars represent standard errors. Statistically significant differences between conditions are indicated through an asterisk and a dotted line between configurations ($*** : p < 0.001$).

3.2. Audio quality models

Both audio quality models, ViSQOL and GPSM^q are reference-based models, whereas the high quality audio material was used as a reference. Accordingly, audio quality was predicted for the low and medium audio material. For both models, audio quality predictions were obtained by averaging predictions for the left and right ear channels.

3.2.1. ViSQOL

Using a reference audio quality of 5 (unnoticeable distortion), medium-quality audio achieved a mean ViSQOL MOS score of 4.68 (SD = 0.02), whereas low-

quality audio yielded a substantially lower mean score of 2.13 (SD = 0.24), indicating poor audio quality.

3.2.2. GPSM^q

For the medium-quality condition, GPSM^q yielded an OPM of 0.00 (SD = 0.00), with a mean OPM_raw of -4.86 (SD = 1.94). For the low-quality condition, GPSM^q yielded a mean OPM of 17.10 (SD = 0.55) and an OPM_raw of 13.59 (SD = 1.30).

3.2.3. Correlation Analysis

Prediction performance is characterized by the Spearman rank correlation coefficient between subjective and objective ratings. The Spearman rank correlation analysis revealed only a weak association between ViSQOL scores and human ratings ($\rho = 0.26$). In comparison, GPSM^q showed a similar prediction performance as indicated by a Spearman rank correlation of $\rho = 0.24$.

4. Discussion

Our results of declining social judgements with lower audio quality are in line with earlier reported results [1], [3], [4].

The degraded job interview audio sample from Walter-Terrill et al. [3] has yielded a ViSQOL MOS score of 2.37, which is in the range of our low-audio quality ViSQOL MOS scores of 2.13 (SD = 0.24). The employability ratings in [3] were generally higher with 76.04 (SD = 18.98) for the clear speech sample and 68.50 (SD = 24.07) for the distorted sample. In our study the employability rating difference between high and low-audio quality is 6.4 which compares well with the difference of 7.54 reported by Walter-Terrill. We did not observe significant differences in employability ratings between medium and high audio quality in the presented videos. This is reflected in the results of the audio quality models:

While ViSQOL provides a continuous estimate of perceptual quality, it does not directly model just-noticeable differences (JND). Prior work [13], [14] suggests that JNDs in audio quality typically correspond to MOS differences on the order of 0.2–0.5. The results of GPSM^q for medium quality yielded an OPM of 0, indicating no perceptual difference between the medium and the high audio-quality signal, which agrees to subjective ratings. However, for the low-quality audio, both models yield scores indicating a strong degradation. This is in line with the reported significant difference in employability ratings.

While the correlation analysis shows that audio quality models contribute to explaining employment ratings, their explanatory power is limited, and most of the variance in social judgments cannot be accounted for by audio quality alone.

One possible explanation for the weak correlations is that employability ratings were also influenced by speaker-related vocal and prosodic cues, such as intonation, speech rate, articulation clarity, or voice quality [6], [7]. Since ViSQOL and GPSM^q are not designed to model these socially relevant cues, future work could examine how audio degradation interacts with prosodic characteristics in shaping employability judgments [15], [16].

Although the participants of this study have self-reported normal hearing, it cannot be ruled out that some of the older participants may in fact have hearing impairements. Such hearing impairments may influence the perception of signal distortions [17], for example through reduced sensitivity to nonlinear distortions as hearing loss increases. In line with this, recent work on videoconferencing has demonstrated that individuals with hearing loss show reduced speech perception and increased listening effort, even when supported by hearing aids [18]. These findings suggest that degraded or suboptimal audio signals impose additional perceptual and cognitive demands on affected listeners. Both models employed here are designed to predict audio quality for normal-hearing listeners and are not suitable for predicting audio quality in hearing-impaired populations.

An additional limitation concerns the reference signal: both ViSQOL and GPSM^q were validated against natural speech, whereas the high-quality reference used here was itself AI-generated and may contain subtle synthetic artifacts. This could partly contribute to the models' limited predictive power, independent of the applied audio degradations.

5. Conclusion

The results of the employability ratings demonstrate that degraded audio quality substantially reduces employability ratings, whereas improvements beyond a moderate quality level do not further enhance evaluations. The results indicate that audio quality models account for only a limited portion of the variance in employability ratings. While audio quality degradations contribute to social judgments, a substantial share of the variance remains unexplained by these models. This suggests that employability judgments are influenced by a broader set of factors beyond audio quality alone. In particular, aspects such as vocal characteristics (e.g., speech rate, intonation, or clarity), perceived confidence, and other paralinguistic cues, as well as potential visual or contextual information in videoconferencing scenarios, are likely to play a significant role. These findings highlight that, although audio quality is a contributing factor, social judgments in communication settings are inherently multidimensional and cannot be fully captured by technical quality metrics alone.

6. Acknowledgments

The authors would like to thank the students of "M08: Methoden III: Experimentelle Übungen" at Uni Distance Suisse in autumn 2025 as well as the participants of the study.

7. Funding

TB was supported by the Deutsche Forschungsgemeinschaft (grant number 352015383 – SFB1330 A2).

References

[1] E. J. Newman and N. Schwarz, "Good Sound, Good Research: How Audio Quality Influences Perceptions of the Research and Researcher," *Science Communication*, vol. 40, no. 2, pp. 246–257, 2018. DOI: 10.1177/1075547018759345.

[2] E. Bild, A. Redman, E. J. Newman, B. R. Muir, D. Tait, and N. Schwarz, "Sound and credibility in the virtual court: Low audio quality leads to less favorable evaluations of witnesses and lower weighting of evidence," *Law and Human Behavior*, vol. 45, no. 5, pp. 481–495, Oct. 2021, ISSN: 1573-661X, 0147-7307. DOI: 10.1037/lhb0000466.

[3] R. Walter-Terrill, J. D. K. Ongchoco, and B. J. Scholl, "Superficial auditory (dis)fluency biases higher-level social judgment," *Proceedings of the National Academy of Sciences*, vol. 122, no. 13, e2415254122, Apr. 2025, ISSN: 0027-8424, 1091-6490. DOI: 10.1073/pnas.2415254122.

[4] M. S. Brucks, J. R. Rifkin, and J. S. Johnson, "Video-call glitches trigger uncanniness and harm consequential life outcomes," *Nature*, vol. 650, no. 8101, pp. 1–8, Dec. 2025, ISSN: 1476-4687. DOI: 10.1038/s41586-025-09823-0.

[5] R. Reber, N. Schwarz, and P. Winkielman, "Processing Fluency and Aesthetic Pleasure: Is Beauty in the Perceiver's Processing Experience?" *Personality and Social Psychology Review*, vol. 8, no. 4, pp. 364–382, Nov. 2004, ISSN: 1088-8683. DOI: 10.1207/s15327957pspr0804_3.

[6] V. Soman and A. Madan, "Social signaling: Predicting the outcome of job interviews from vocal tone and prosody," in *IEEE International Conference on Acoustics, Speech, and Signal Processing*, 2009.

[7] I. Naim, M. I. Tanveer, D. Gildea, Mohammed, and Hoque, *Automated Analysis and Prediction of Job Interview Performance*, arXiv:1504.03425 [cs.HC], Apr. 2015. DOI: 10.48550/arXiv.1504.03425.

[8] A. Hines, J. Skoglund, A. C. Kokaram, and N. Harte, "ViSQOL: An objective speech quality model," *EURASIP Journal on Audio, Speech, and Music Processing*, vol. 2015, no. 1, p. 13, Dec. 2015, ISSN: 1687-4722. DOI: 10.1186/s13636-015-0054-9.

[9] T. Biberger, J.-H. Fleßner, R. Huber, and S. Ewert, "An Objective Audio Quality Measure Based on Power and Envelope Power Cues," *Journal of the Audio Engineering Society*, vol. 66, no. 7/8, pp. 578–593, Aug. 2018, ISSN: 1549-4950. DOI: 10.17743/jaes.2018.0031.

[10] F. Henninger, Y. Shevchenko, U. K. Mertens, P. J. Kieslich, and B. E. Hilbig, "Lab.js: A free, open, online study builder," *Behavior Research Methods*, vol. 54, no. 2, pp. 556–573, Apr. 2022, ISSN: 1554-3528. DOI: 10.3758/s13428-019-01283-5.

[11] R. C. Streijl, S. Winkler, and D. S. Hands, "Mean opinion score (MOS) revisited: Methods and applications, limitations and alternatives," *Multimedia Systems*, vol. 22, no. 2, pp. 213–227, Mar. 2016, ISSN: 0942-4962, 1432-1882. DOI: 10.1007/s00530-014-0446-1.

[12] T. Biberger and S. D. Ewert, "Envelope and intensity based prediction of psychoacoustic masking and speech intelligibility," *The Journal of the Acoustical Society of America*, vol. 140, no. 2, pp. 1023–1038, Aug. 2016, ISSN: 0001-4966, 1520-8524. DOI: 10.1121/1.4960574.

[13] M. Chinen, F. S. C. Lim, J. Skoglund, N. Gureev, F. O'Gorman, and A. Hines, *ViSQOL v3: An Open Source Production Ready Objective Speech and Audio Metric*, arXiv:2004.09584 [eess], Apr. 2020. DOI: 10.48550/arXiv.2004.09584.

[14] J. Zhu, H. Amirpour, W. Zhou, and P. L. Callet, *Is there a relationship between Mean Opinion Score (MOS) and Just Noticeable Difference (JND)?* arXiv:2602.17010 [eess], Feb. 2026. DOI: 10.48550/arXiv.2602.17010.

[15] I. Siegert and O. Niebuhr, "Case Report: Women, Be Aware that Your Vocal Charisma can Dwindle in Remote Meetings," *Frontiers in Communication*, vol. 5, p. 611 555, Jan. 2021, ISSN: 2297-900X. DOI: 10.3389/fcomm.2020.611555.

[16] O. Niebuhr and I. Siegert, "*High on emotion?* how audio codecs interfere with the perceived charisma and emotional states of men and women," in *Studientexte zur Sprachkommunikation: Elektronische Sprachsignalverarbeitung 2022*, O. Niebuhr, M. S. Lundmark, and H. Weston, Eds., TUDpress, Dresden, Mar. 2022, pp. 243–252, ISBN: 978-3-95908-548-9.

[17] T. Biberger and S. D. Ewert, "Audio Quality Perception of Hearing-Impaired Listeners in Complex Acoustic Environments," *Trends in Hearing*, vol. 29, Sep. 2025, Art. no. 23312165251374938, ISSN: 2331-2165. DOI: 10.1177/23312165251374938.

[18] V. W. Zhang, J. Tsiolkas, A. Sebastian, A. Wong, and P. Kitterick, "Understanding and enhancing the online video call experience of individuals with hearing loss," *International Journal of Audiology*, pp. 1–14, May 2025, ISSN: 1499-2027, 1708-8186. DOI: 10.1080/14992027.2025.2505541.