Interspeech 2025
17-21 August 2025, Rotterdam, The Netherlands

Queer Waves: A German Speech Dataset Capturing Gender and Sexual Diversity from Podcasts and YouTube

Ingo Siegert[1], *Jan Marquenie*[1], *Sven Grawunder*[2]

[1]Mobile Dialog Systems, Otto von Guericke University Magdeburg, Germany
[2]Speech Science and Phonetics, Halle University, Germany

siegert@ovgu.de, jan.marquenie@ovgu.de, sven.grawunder@sprechwiss.uni-halle.de

Abstract

Developing equitable and inclusive speech technologies requires datasets that represent the full spectrum of human voices, including those of LGBTQIA+ speakers. However, capturing spontaneous, high-quality audio from marginalized gender and sexual identities presents significant ethical, logistical, and representational challenges. This paper introduces Queer Waves, a German speech corpus compiled from podcast and YouTube content featuring self-identified LGBTQIA+ speakers, with a particular focus on diverse gender identities and sexual orientations. We further address the legal and ethical considerations inherent in collecting sensitive personal data. The Queer Waves corpus comprises approximately 335 hours of speech from over 400 self-identified LGBTQIA+ speakers, spanning ages from 18 to 86 years. By expanding representation across a wide range of gender identities and orientations, Queer Waves aims to advance the development of fairer and more accurate speech technologies.
Index Terms: LGBTQIA+ Representation, Inclusive Datasets, Spontaneous Speech, Podcast Speech Data

1. Introdution

The rapid development of AI technologies has brought increased attention to the principles of fairness and inclusivity, sparking more intensive discussions about these aspects. As a result, the creation of datasets that address specific inclusivity factors has gained greater significance [1].

Despite growing recognition in the speech research community regarding the necessity of diverse datasets – covering factors like age, language, ethnicity, geographic variation, and recording conditions [2, 3, 4, 5] – one essential aspect remains notably underrepresented: gender identities (e.g., transgender, nonbinary) and sexual orientations (e.g., gay, lesbian, bisexual). Speech models are often built upon datasets that predominantly reflect binary gender perspectives.

Such lack of representation not only hampers the development of fair AI applications [6] but also significantly constrains (linguistic and phonetic) research [7, 8]. Diverse gender identities and sexual orientations contribute to the richness of human speech, offering insights into sociophonetic variation, language change, and identity expression. Excluding these voices restricts our understanding of how speech features, i.e., pitch, formant patterns, interact with social identity – insights crucial to both automatic speech processing and theoretical linguistics.

The consequences of these biases are particularly evident in applications like voice authentication [9] and anonymization systems [5], where entrenched gender assumptions can result in exclusion, misclassification, and privacy breaches. For LGBTQIA+ individuals, such outcomes can cause distress or even pose serious risks, especially in regions where gender and sexual identity are stigmatized or criminalized. Additionally, phoneticians and sociolinguists face challenges in drawing accurate conclusions about gendered speech variations when relying on datasets that overlook non-cisgender voices.

The need for an open, extensive, and high-quality database of LGBTQIA+ speakers is evident. Such a resource, recorded under optimal acoustic conditions (minimizing background noise and ensuring consistent sound quality) would not only advance inclusive AI applications but also provide a solid foundation for linguistic and sociophonetic investigations. High-quality recordings are essential for identifying authentic speech features rather than artifacts caused by technical deficiencies.

The goal of this paper is to introduce a new dataset: a German speech corpus compiled from podcast and YouTube content featuring self-identified LGBTQIA+ speakers. This dataset specifically targets spontaneous speech recorded under high-quality conditions to ensure reliable phonetic analyses and support the development of more inclusive speech technologies. By broadening the spectrum of queer voices included in both AI and linguistic studies, this work aims to promote a more accurate and inclusive understanding of human speech diversity.

2. Terminology

The term LGBTQIA+ stands for Lesbian, Gay, Bisexual, Transgender, Queer/Questioning, Intersex, Asexual, with "+" acknowledging additional identities (e.g., pansexual) [10]. Cisgender (or "cis") describes individuals whose gender identity aligns with the sex they were assigned at birth, while transgender (or "trans") serves as an umbrella term for those whose gender identity diverges from that assignment (including female-to-male (ftm), i.e. trans men, male-to-female (mtf), i.e. trans women, and nonbinary (NB) individuals). Detransition refers to the process by which an individual reverts to a gender identity that aligns with their sex assigned at birth after having previously transitioned. This process can be social, legal, or medical and is influenced by a variety of personal, social, and contextual factors. Nonbinary refers to gender identities that do not conform strictly to the traditional male/female binary, and the term queer is adopted here as an inclusive label encompassing diverse gender and sexual identities beyond heterosexual and binary expectations. The term has a dual meaning here: it serves both as an umbrella term for diverse identities and as a way individuals self-identify to describe their sexual orientation.

We acknowledge that categorizations in this context can be sensitive, with fluid boundaries and individual nuances; however, for the purpose of academic clarity and systematic analysis, some degree of classification remains essential to enable structured discussion and comparison.

10.21437/Interspeech.2025-2022

3. Literature Review

3.1. Phonetics of sexual orientation and gender identity

Apart from general differences related to sexual dimorphism in humans [11], there have been long noted sociophonetic factors that influence speech and voice features in natural language. For example, male speakers who self-identify as less masculine would tend to exhibit larger vowel spaces and higher average fundamental frequency [12]. In this way a potential relation of gender identity and speakers may actively use different articulatory strategies to index their gender identity or sexual orientation by means of articulatory or phonatory strategies, i.e. in voice pitch and other voice quality measures (including HNR, jitter, shimmer, spectral tilt) [13, 14] or more subtle segmental differences, as for sibilants like the /s/-sound [15, 16]. This has been corroborated by a bulk of research for gay speech [11]. Nonbinary persons on the other hand may not want to be categorized and attributed. Moreover, these cues can be perceived and are picked up by listeners and therefore relevant for communication in a particular community [17]. In other social contexts, speaker might not express these cues to hide their sexual/gender identity.

3.2. Known datasets

In standard speech datasets, gender identity has only been recorded for a few years, and even then, it is typically limited to the category "diverse" without further differentiation. To the best of our knowledge, sexual orientation is not recorded at all. As a result, the well-known larger datasets often do not include diverse speakers, as shown in Table 1. A similar situation can be observed in German datasets, where gender identity is equally underrepresented.

Corpus	Language	Gender (M/F)
CALLHOME	Multi	50% / 50%
CALLFRIEND	Multi	55% / 45%
TED-LIUM V3	EN	60% / 40%
Tuda-DE[18]	DE	72% / 28%
Kiel[19]	DE	55% / 45%

Table 1: *Overview of major speech corpora and their gender distribution – none of them include a category for diverse speakers.*

Recent volunteer-based datasets increasingly adopt inclusive practices, reflecting growing awareness of diversity in speech data, particularly regarding gender identity and, in some cases, sexual orientation.

Mozilla Common Voice has expanded its gender options beyond the binary categories, allowing participants to self-identify, including "other" [1]. However, representation remains limited: among 94,911 English contributors, only 3 identify as trans and 8 as non-binary. For German (19,348 contributors), only 1 trans and 2 non-binary participants are included.
The German Plapper Corpus also embraces inclusivity by collecting information on gender identity (diverse, female, male, or no answer) and sexual orientation using the Kinsey scale [12]. Despite this approach, only 8 diverse speakers are included out of 1,522 participants, contributing only a few sentences. Data on sexual orientation has not been published.

So far, we have found only a few datasets that address this gap. The Gender Spectrum Speech Corpus [20] includes recordings in English and French, consisting of read speech and interviews. The first version featured 30 English and 35 French speakers, including 8 non-binary participants, 17 cisgender participants, 3 trans men, and 1 trans woman. Additionally, this dataset collected information about participants' experiences with testosterone therapy and vocal training, providing valuable context for understanding gender-related speech variation. However, the dataset is not publicly available, and only sparse documentation exists.

The Mid-Atlantic Gender Expansive Speech (MAGES) Corpus (a subset of [21]) Corpus includes 14 transgender and nonbinary English-speaking participants. While both datasets were created with participant consent and provide valuable insights into gender-diverse speech, their small sample sizes and limited accessibility hinder comprehensive phonetic and linguistic analyses. In particular, the scarcity of data makes it difficult to study sociophonetic markers like vowel shifts or intonation patterns that distinguish queer speech varieties.

4. Dataset Collection

We have adapted and refined the methods to collect data capturing gender and sexual diversity from a previous proof of concept [22].

4.1. Selection of suitable podcasts

Podcast data as a source for scientific corpora is well-established, e.g., the MPS-Podcast Corpus [23] or 100,000 Podcasts [24]. In general, Kulkov et al. [25] provide a solid methodological framework, particularly for selecting suitable podcast sources. For our purposes, the crucial aspects in selecting podcasts are:

- the diversity of guests (with a specific focus on guests from the LGBTQIA+ spectrum),
- technical production quality, and
- the German language.

Based on a search across various podcast platforms, we selected the following 12 podcasts. These podcasts follow a common format: one or two hosts invite one or more guests per episode to discuss current topics. Additionally, we included 25 YouTube videos and channels, selected based on the search terms "transmann," "transfrau," "detrans," and "nicht/non-binär." This inclusion aimed specifically to increase the representation of trans and non-binary individuals within the dataset. Moreover, we aimed to ensure broader diversity among podcast participants, including representation from rural regions, people of color (PoC), and individuals with a migration background, which emerged as a valuable secondary outcome:

- Auf eine Tüte: Der Potcast mit Hengameh
- BBQ – Der Black Brown Queere Podcast
- Bootycall
- Böttinger Wohnung 17
- Out and About
- Queer as Berlin
- Queerkram
- Reden ist Gold
- Somewhere Over The Hay Bale
- SPUTNIK Pride
- Willkommen im Club
- Hotel Matze

4.2. Legal aspects

The collection of a podcast data corpus involves legal considerations related to data protection and copyright. Under the GDPR,

[1] https://foundation.mozilla.org/en/blog/expanding-gender-options-on-common-voice/

Article 9(2)(e) [26] permits the processing of special categories of personal data if the data subject has made the information publicly available. Additionally, processing is justified by the public interest, as research on marginalized genders' language data contributes to fairer and more inclusive AI technologies. This supports the identification and mitigation of potential discrimination, fostering societal trust in AI systems.

From a copyright perspective, §60d of the German Copyright Act (UrhG) [27] allows systematic reproduction of publicly accessible works for scientific text and data mining, provided access is restricted to a defined research group. Public distribution requires rights holders' consent. Long-term storage is permitted if necessary for future research or verification.

Based on these considerations, the presented dataset can be collected and made available for scientific purposes in compliance with legal requirements [28].

4.3. Data collection

The podcast episode descriptions were automatically downloaded using the podcasts' RSS feed link , resulting in 995 episodes. Subsequently, all episodes were excluded based on their descriptions if they did not feature any guests or were recorded in a live atmosphere (150 episodes). The remaining episodes were then analyzed in detail (using JSON descriptions, transcripts, and provided links) to extract as much information as possible about the guests (e.g., age, sexual orientation, gender identity, and regional background as a proxy for dialectal variation). At this stage, additional episodes that were not flagged during the initial screening were removed, but to a much lesser extent (40 episodes). Simultaneously, the different speakers per episode were identified. Including the hosts, this resulted in 647 unique speakers and 25 speakers from the YouTube data.

Afterwards, for the final episodes, speaker diarization was performed using WhisperX [29], ensuring precise segmentation of individual speakers. To identify the podcast hosts across episodes, we applied Picovoice Eagle for automatic speaker identification [30]. For further refinement, we utilized YAMNet [31], a pretrained deep neural network capable of predicting 521 audio event classes based on the AudioSet-YouTube corpus [32], to distinguish speech transitions and detect introductory music. The classification scores from YAMNet were reviewed to assess the accuracy of event detection.

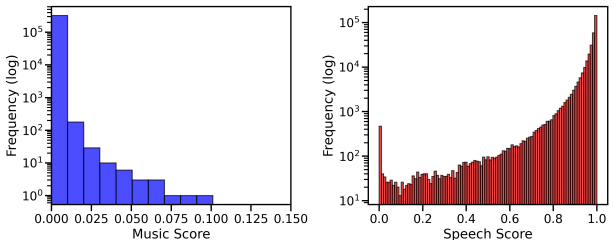


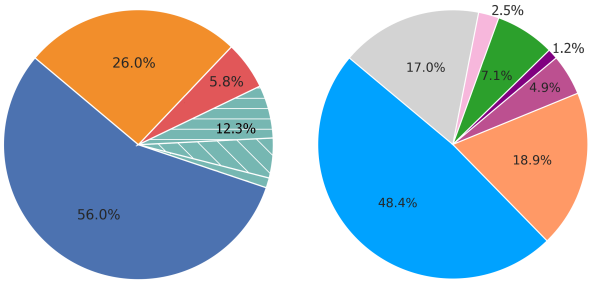
Figure 1: *Distribution of music and speech scores for podcast segments.*

To further illustrate the challenges with background music overlays, Figure 1 shows the distribution of classification scores from YAMNet for music detection. The plot demonstrates that segments with background music still exhibit relatively high speech scores, indicating that a high threshold is necessary to filter out non-speech content. In contrast, the music scores are unreliable for distinguishing speech from background music. As a final threshold, we set 0.9, resulting in the deletion of 29,704 files, all of which contained music, as confirmed by a random sample check.

5. Dataset Overview

The dataset exclusively includes individuals who identify as part of the LGBTQIA+ community. Individuals clearly identified as cisgender and heterosexual, as well as those for whom no reliable information regarding gender identity or sexual orientation was available, were excluded from the final dataset. This selection process was conducted carefully to respect privacy and maintain the integrity of the research. As a result, the final dataset comprises 407 unique speakers, encompassing a total of 335 hours of audio recordings. In addition to the hosts, who appear in every episode of their respective podcasts[2], several guests are featured multiple times, sometimes across different podcasts. Specifically, two guests appear five times, one guest four times, twelve guests three times, and forty guests twice.

5.1. Gender identification and sexual orientation



(a) *Gender identities: Female, Non-binary+, Trans (ftm), Trans(mtf), Trans/Detrans, Male,*

(b) *Sexual Orientation: Gay, Lesbian, Bi/Inter, Asexual, Queer, Pan/Poly+, No Information*

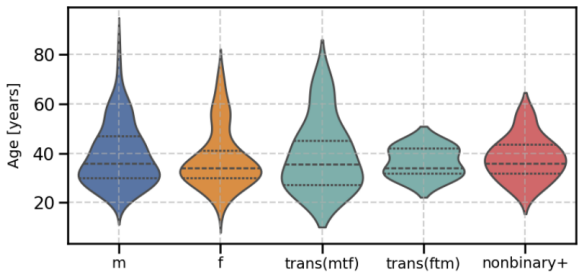
Figure 2: *Pie charts of speaker distribution regarding gender identity and sexual orientation. For the trans category, trans(ftm) and Trans(mtf) are highlighted, as far as is known.*

When it comes to gender identity, the dataset covers individuals identifying as *male*, *female*, *trans/detrans*, i.e., trans or detransitioned (including trans male-to-female (mtf), trans female-to-male (ftm), trans, and detransitioned (detrans)), and *non-binary+* (encompassing nonbinary, genderfluid, genderfree, gender nonconformity), covering a range of gender expressions, see Figure 2a. Concerning sexual orientation, the dataset includes speakers who identify as *gay*, *lesbian*, *bi/inter*, i.e. bisexual or intersexual, *asexual*, *queer*, *pan/poly+*, i.e., pansexual or polyamorous. Those for whom no information is available are denoted as *na*. The distribution regarding sexual orientation is depicted in Figure 2b.

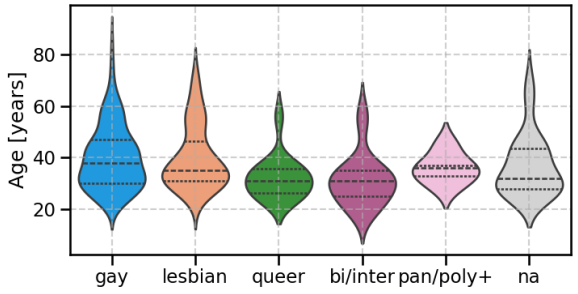
5.2. Age distribution

We were able to determine the birth year for 315 speakers. The results indicate that the overall dataset is diverse in terms of age distribution, with ages ranging from 18 to 86 years. Among the gender identity groups (Figure 3a), female speakers (f) had a mean age of 37.99 years, ranging from 18 to 72 years, while male speakers (m) had a mean age of 39.39 years, with ages spanning from 20 to 86 years. Nonbinary and gender-expansive

[2]There are two podcasts where hosts change intermittently



(a) *Gender identities*



(b) *Sexual Orientation*

Figure 3: *Age Distribution for included LGBTQIA+ speakers. Note: not for all speakers age information was given.*

individuals (nonbinary+) had a mean age of 38.11 years, ranging from 25 to 55 years. Participants identifying as transgender (FTM or MTF) had a mean age of 38.27 years, with ages ranging from 26 to 70 years. Specifically, the trans (FTM) group had a mean age of 36.43 years (ranging from 30 to 43 years), while the trans (MTF) group had a mean age of 39.56 years (ranging from 26 to 70 years). For the two detrans participants, no age information was available.

Regarding sexual orientation (Figure 3b), bisexual or intersex (bi/inter) participants had a mean age of 31.37 years, ranging from 18 to 58 years. Gay participants had a mean age of 40.19 years, spanning from 21 to 86 years, while lesbian participants had a mean age of 40.02 years, with ages between 23 and 72 years. Participants who did not disclose their orientation (na) had a mean age of 36.84 years, ranging from 25 to 70 years. Those identifying as pansexual or polyamorous (pan/poly+) had a mean age of 36.00 years, with ages between 29 and 45 years. Lastly, queer participants had a mean age of 32.21 years, ranging from 24 to 56 years. Details regarding these broader clusters for gender identity or sexual orientation are given in Figure 3.

5.3. Speaker duration overview

The majority of speakers have between 15 and 45 minutes of recorded material, with a minimum recording duration of 5 minutes, ensuring that even shorter contributions are linguistically valuable, see Figure 4. This distribution indicates that most participants are guests or short-term contributors. Only 19 speakers exceed 2 hours, representing the podcast hosts, with two podcasts featuring rotating or dual hosts. This distribution highlights a clear distinction between hosts and guests in terms of speaking time. The minimum recording length of 5 minutes, combined with the diversity of speaker durations, ensures that there is sufficient speech material for all categories of gender identity and sexual orientation, enabling meaningful comparative analyses. The hosts' longer recordings make them particularly valuable for longitudinal analysis, such as tracking speaking patterns or vocal changes over time, while the guest's recordings contribute a diverse range of voices, accents, and speaking styles, supporting comparative studies and facilitating speaker variability analyses across different identity categories.

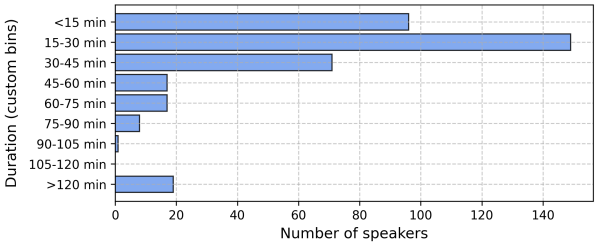


Figure 4: *Distribution of Speaker Durations.*

6. Data availability statement

The statistical summaries and dataset overviews will be made publicly available on GitHub, see `https://ingosiegert.github.io/QueerWavesWebsite/`. Speaker metadata and download-skripts and timing information will be shared only upon request and solely for research purposes.

7. Conclusion

In conclusion, this paper presents Queer Waves, a German speech corpus aimed at fostering more inclusive and equitable speech technologies by amplifying the voices of LGBTQIA+ speakers. With approximately 335 hours of audio from over 400 self-identified LGBTQIA+ speakers, Queer Waves addresses the underrepresentation of LGBTQIA+ speakers in existing datasets.

A key reason for overrepresenting LGBTQIA+ speakers in this corpus lies in addressing structural biases in public datasets. As shown for popular speech datasets, LGBTQIA+ speakers are often underrepresented, leading to visibility gaps and increased risks of misclassification or discriminatory outcomes in speech technologies [33]. By deliberately enriching datasets with LGBTQIA+ voices, Queer Waves provides a valuable resource for augmenting existing models, reducing visibility bias, and ensuring fairer, more robust speech recognition systems.

Moreover, this targeted collection closes an essential research gap, as self-disclosure of identity is seldom captured in large-scale datasets. The result is a dataset that not only supports the development of fairer models but also offers significant value for scientific inquiry into underrepresented voices. Queer Waves thus stands as both a corrective and a complement to existing datasets, contributing to greater fairness, visibility, and protection against discrimination in automated speech technologies.

8. Acknowledgements

We thank Mareille Leonhard and ChiaraLindloff for their support in preparing and reviewing the data. Their help with speaker data and transcription checks contributed to the reliability of the corpus.

9. References

[1] N. Markl, "Mind the data gap(s): Investigating power in speech and language datasets," in *Proceedings of the Second Workshop on Language Technology for Equality,*

Diversity and Inclusion. Dublin, Ireland: Association for Computational Linguistics, May 2022, pp. 1–12. [Online]. Available: https://aclanthology.org/2022.ltedi-1.1/

[2] I.-E. Veliche, Z. Huang, V. Ayyat Kochaniyan, F. Peng, O. Kalinli, and M. L. Seltzer, "Towards measuring fairness in speech recognition: Fair-speech dataset," in *Interspeech 2024*, 2024, pp. 1385–1389.

[3] W.-S. Chien and C.-C. Lee, "An investigation of group versus individual fairness in perceptually fair speech emotion recognition," in *Interspeech 2024*, 2024, pp. 3205–3209.

[4] M. Schubert, D. Duran, and I. Siegert, "Challenges of German speech recognition: A study on multi-ethnolectal speech among adolescents," *Interspeech 2024*, pp. 3045–3049, 2024.

[5] A. Leschanowsky and S. Das, "Examining the interplay between privacy and fairness for speech processing: A review and perspective," in *4th Symposium on Security and Privacy in Speech Communication*, 2024, pp. 1–11.

[6] J. Smith, A. Patel, and D. Wong, "Promoting fairness and diversity in speech datasets for mental health and neurological disorders research," *Journal of Artificial Intelligence Research*, vol. 82, pp. 937–972, 2025.

[7] K. Sarveswaran, A. Vaidya, B. K. Bal, S. Shams, and S. Thapa, "Proceedings of the first workshop on challenges in processing south asian languages (chipsal 2025)," in *Proceedings of the First Workshop on Challenges in Processing South Asian Languages (CHiPSAL 2025)*, 2025.

[8] A. Gesser-Edelsburg, R. Hijazi, E. Eliyahu, and A. Tal, "Bridging the divide: An empirical investigation of artificial intelligence and generative artificial intelligence integration across genders, disciplines and academic roles." *European Journal of Open, Distance & E-Learning*, vol. 26, no. 1, 2024.

[9] G. Fenu, H. Lafhouli, and M. Marras, "Exploring algorithmic fairness in deep speaker verification," in *Computational Science and Its Applications – ICCSA 2020*. Cham: Springer International Publishing, 2020, pp. 77–93.

[10] A. L. Koenig, "Glossary of terms," in *Creative Arts Therapies and the LGBTQ Community: Theory and Practice*, B. MacWilliam, B. T. Harris, D. G. Trottier, and K. Long, Eds. Jessica Kingsley Publishers, Jun. 2019.

[11] B. Munson and M. Babel, "The phonetics of sex and gender," in *The Routledge Handbook of Phonetics*. Routledge, 2019, pp. 499–525.

[12] M. Weirich, D. Duran, and S. Jannedy, "Gender and age based f0-variation in the German Plapper Corpus," in *Proceedings of Interspeech 2024*, 2024, pp. 1565–1569.

[13] L. Zimman, "Gender as stylistic bricolage: Transmasculine voices and the relationship between fundamental frequency and/s," *Language in Society*, vol. 46, no. 3, pp. 339–370, 2017.

[14] B. LeAnn and P.-L. Claire, "Bright Voice Quality and Fundamental Frequency Variation in Non-binary Speakers," *Journal of Voice*, vol. 39, no. 1, pp. 282.e1–282.e17, Jan. 2025.

[15] S. Fuchs and M. Toda, "Do differences in male versus female /s/ reflect biological or sociophonetic factors?" in *Turbulent Sounds*, S. Fuchs, M. Toda, and M. Zygis, Eds. DE GRUYTER MOUTON, Apr. 2010, pp. 281–302.

[16] N. Houle, M. P. Lerario, and S. V. Levi, "Spectral analysis of strident fricatives in cisgender and transfeminine speakers," vol. 154, no. 5, pp. 3089–3100. [Online]. Available: https://pubs.aip.org/jasa/article/154/5/3089/2921294/Spectral-analysis-of-strident-fricatives-in

[17] M. Hope and J. Lilley, "Gender expansive listeners utilize a non-binary, multidimensional conception of gender to inform voice gender perception," *Brain and Language*, vol. 224, p. 105049, Jan. 2022.

[18] S. Radeck-Arneth, B. Milde, A. Lange, E. Gouvêa, S. Radomski, M. Mühlhäuser, and C. Biemann, "Open Source German Distant Speech Recognition: Corpus and Acoustic Model," in *Text, Speech, and Dialogue*. Pilsen, Czech Republic: Springer International Publishing, 2015, pp. 480–488.

[19] J. Peters, F. Kleber, and N. Scheffer, "The Kiel Corpus of Spoken German: Documentation and Structure," Institute for Scandinavian, Frisian, and General Linguistics, University of Kiel, Tech. Rep., 2017, accessed: 2025-02-13. [Online]. Available: https://www.isfas.uni-kiel.de/de/linguistik/forschung/kiel-corpus/docs/Info_KielCorp_2017.pdf

[20] J. S. German, M. Candea, L. Brown, T. Mahrt, and O. Reid-Collins, "Gender spectrum speech corpus," 2022, ORTOLANG (Open Resources and TOols for LANGuage) –www.ortolang.fr. [Online]. Available: https://hdl.handle.net/11403/gender_spectrum_speech/v2.1

[21] H. T. Bunnell, J. Lilley, and K. McGrath, "The ModelTalker Project: A Web-Based Voice Banking Pipeline for ALS/MND Patients," in *Interspeech 2017*, 2017, pp. 4032–4033.

[22] J. Marquenie, M. Leonhardt, S. Grawunder, and I. Siegert, "Gender spectrum data from podcasts – a proof of concept," in *Studientexte zur Sprachkommunikation: Elektronische Sprachsignalverarbeitung 2025*, S. Grawunder, Ed. TUDpress, Dresden, Mar. 2025, pp. 239–246. [Online]. Available: https://www.essv.de/pdf/2025_239_246.pdf

[23] R. Lotfian and C. Busso, "Building naturalistic emotionally balanced speech corpus by retrieving emotional speech from existing podcast recordings," *IEEE Transactions on Affective Computing*, vol. 10, no. 4, pp. 471–483, October-December 2019.

[24] A. Clifton, S. Reddy, Y. Yu, A. Pappu, R. Rezapour, H. Bonab, M. Eskevich, G. Jones, J. Karlgren, B. Carterette, and R. Jones, "100,000 Podcasts: A Spoken English Document Corpus," in *Proc. of the 28th COLING*, Barcelona, Spain (Online), Dec. 2020, pp. 5903–5917. [Online]. Available: https://aclanthology.org/2020.coling-main.519/

[25] I. Kulkov, J. Kulkova, R. Rohrbeck, and L. Menvielle, "Leveraging podcasts as academic resources: A seven-step methodological guide," *International Journal of Qualitative Methods*, vol. 23, 2024.

[26] (2016) Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC (General Data Protection Regulation). Official Journal of the European Union, L 119, 4 May 2016, p. 1–88. Accessed on 2 February 2025. [Online]. Available: http://data.europa.eu/eli/reg/2016/679/oj

[27] T. Margoni and M. Kretschmer, "A Deeper Look into the EU Text and Data Mining Exceptions: Harmonisation, Data Ownership, and the Future of Technology," *GRUR International*, vol. 71, no. 8, pp. 685–701, August 2022. [Online]. Available: https://doi.org/10.1093/grurint/ikac054

[28] A. Jobin, M. Ienca, and E. Vayena, "The global landscape of AI ethics guidelines," *Nature Machine Intelligence*, vol. 1, pp. 389–399, 2019. [Online]. Available: https://doi.org/10.1038/s42256-019-0088-2

[29] M. Bain, J. Huh, T. Han, and A. Zisserman, "WhisperX: Time-Accurate Speech Transcription of Long-Form Audio," *Interspeech 2023*, 2023.

[30] P. Inc.q, "Eagle speaker recognition verify and identify speakers during natural conversations," 2024, accessed: 2024-12-13. [Online]. Available: https://picovoice.ai/platform/eagle/

[31] S. Hershey, S. Chaudhuri, D. P. W. Ellis, J. F. Gemmeke, A. Jansen, C. Moore, M. Plakal, D. Platt, R. A. Saurous, B. Seybold, M. Slaney, R. Weiss, and K. Wilson, "Cnn architectures for large-scale audio classification," in *IEEE ICASSP 2017*, 2017. [Online]. Available: https://arxiv.org/abs/1609.09430

[32] J. F. Gemmeke, D. P. W. Ellis, D. Freedman, A. Jansen, W. Lawrence, R. C. Moore, M. Plakal, and M. Ritter, "Audio set: An ontology and human-labeled dataset for audio events," in *Proc. IEEE ICASSP 2017*, New Orleans, LA, 2017.

[33] D. McAra-Hunter, "How ai hype impacts the lgbtq+ community," *AI Ethics*, vol. 4, no. 771–790, 2024. [Online]. Available: https://doi.org/10.1007/s43681-024-00423-8