Literature Review: Depression Detection Through Speech

Overview

Depression detection through speech is an active area of research at the intersection of speech processing, machine learning, and clinical psychology. The field aims to identify acoustic and linguistic markers of depression that could enable automated, objective screening tools.

Key Datasets

1. DAIC-WOZ (Distress Analysis Interview Corpus - Wizard-of-Oz)

Source: USC Institute for Creative Technologies
URL: https://dcapswoz.ict.usc.edu/

Description:

- Clinical interviews designed to support diagnosis of psychological distress conditions (anxiety, depression, PTSD)
- Interviews conducted by "Ellie" - an animated virtual interviewer (controlled by human in Wizard-of-Oz setup)
- **189 sessions**, ranging 7-33 minutes (average 16 minutes)
- Includes: audio recordings, video recordings, transcripts, facial features, questionnaire responses
- Ground truth: PHQ-8 depression questionnaire scores

Key Papers:

- Gratch et al. (2014) "The Distress Analysis Interview Corpus of Human and Computer Interviews" - LREC 2014
- DeVault et al. (2014) "SimSensei kiosk: A virtual human interviewer for healthcare decision support" - AAMAS 2014

2. Extended DAIC Database

- Extension of DAIC-WOZ for AVEC 2019 challenge
- Includes recordings where virtual agent is **fully AI-driven** (no human Wizard-of-Oz)
- Tests how absence of human interviewer impacts depression assessment

3. ANDROIDS Corpus

Status: Need to verify exact GitHub location
Note: Nile mentioned exploring this - will investigate further

---

Recent Key Papers (2023-2025)

Systematic Reviews & Meta-Analyses

1. "Performance of Automatic Speech Analysis in Detecting Depression: Systematic Review and Meta-Analysis"

Authors: Maran et al. (2025)
Journal: JMIR Mental Health
DOI: 10.2196/67802

Key Findings (105 studies reviewed):

- Pooled highest accuracy: **0.81** (95% CI: 0.79-0.83)
- Pooled highest sensitivity: **0.84** (95% CI: 0.81-0.86)
- Pooled highest specificity: **0.83** (95% CI: 0.79-0.86)
- Pooled highest precision: **0.81** (95% CI: 0.77-0.84)

Lowest performance metrics:

- Accuracy: 0.66, Sensitivity: 0.63, Specificity: 0.60, Precision: 0.64

Conclusion: "ASA shows promise as a method for detecting depression, though its readiness for clinical application as a standalone tool remains limited. At present, it should be regarded as a complementary method."

2. "Speech as a biomarker for depression"

Authors: Koops, Brederoo, de Boer et al. (2023)
Journal: CNS & Neurological Disorders - Drug Targets
Cited by: 119

3. "Speech-based depression assessment: A comprehensive survey"

Authors: Leal, Ntalampiras, Sassi (2024)
Journal: IEEE Transactions
Cited by: 19

4. "Diagnostic accuracy of traditional and deep learning methods..."

Authors: Lu et al. (2025)
Journal: BMC Psychiatry
Note: Compares traditional ML vs deep learning approaches

University of Glasgow PhD Thesis (Highly Relevant!)

"Speech-based automatic depression detection via biomarkers identification and artificial intelligence approaches"

Author: Fuxiang Tao (2024)
Supervisor: Professor Alessandro Vinciarelli
Institution: University of Glasgow, School of Engineering
URL: https://theses.gla.ac.uk/84055/
DOI: 10.5525/gla.thesis.84055

Abstract Summary:

- Depression affects 300+ million people globally
- Traditional diagnosis: time-consuming, dependent on clinical experience
- Thesis shows TWO ways to benefit from automatic detection:

    1. **Identifying speech markers** (duration, pauses, correlation matrices)
    2. **Novel deep learning models** (Multi-local Attention, Cross-Data Multilevel Attention)

Key Contributions:

- Proposed speech markers: speech duration, pauses, acoustic feature correlation matrices
- Found statistically significant differences between depressed/non-depressed
- Proposed Multi-local Attention (MLA) mechanism
- Proposed Cross-Data Multilevel Attention (CDMA) model

Related Publications:

- INTERSPEECH 2023, 2020
- ICASSP 2023
- WACV 2023

Why This Matters: Same university, same topic, recent work. Could be valuable reference.

4. SEWA Corpus

- Cross-cultural emotion recognition dataset
- German, Hungarian, Chinese cultures
- Audio-visual recordings "in-the-wild" (webcams, home/workplace)

Benchmark Challenges

AVEC (Audio/Visual Emotion Challenge) Series

The AVEC challenge series has been instrumental in advancing depression detection research:

AVEC 2019: "State-of-Mind, Detecting Depression with AI, and Cross-cultural Affect"

Venue: ACM Multimedia 2019, Nice, France

Three Sub-Challenges:

1. **State-of-Mind Sub-Challenge (SoMS)**
    - Predict self-reported mood (10-point Likert scale) from audio-visual recordings
    - Evaluates continuous adaptation of human state-of-mind
2. **Detecting Depression with AI Sub-Challenge (DDS)** ☆ Most relevant
    - Predict PHQ-8 depression severity scores
    - Used DAIC-WOZ corpus + new AI-driven interviews
    - Performance metric: Concordance Correlation Coefficient (CCC)
3. **Cross-cultural Emotion Sub-Challenge (CES)**
    - Transfer learning across cultures (German/Hungarian → Chinese)

Research Contributions Sought:

- Multimodal affect sensing (audio, video, physiological)
- Transfer learning
- Semi-supervised/unsupervised learning
- Personalized recognition
- Context in emotion recognition

Previous AVEC Challenges

- AVEC 2016: Depression Sub-Challenge (DSC) - predecessor to 2019 DDS
- AVEC 2018: Cross-cultural emotion recognition

Speech Features for Depression Detection

Acoustic Features (Low-Level Descriptors)

1. **Prosodic Features**
    - Pitch (F0) - fundamental frequency
    - Pitch variability
    - Speaking rate
    - Pause patterns (duration, frequency)
    - Energy/intensity contours
2. **Spectral Features**
    - Mel-Frequency Cepstral Coefficients (MFCCs)
    - Formant frequencies (F1, F2, F3)
    - Spectral flux
    - Spectral centroid
3. **Voice Quality Features**
    - Jitter (pitch perturbation)
    - Shimmer (amplitude perturbation)
    - Harmonics-to-Noise Ratio (HNR)
4. **Temporal Features**
    - Speech rate
    - Articulation rate
    - Pause-to-speech ratio
    - Response latency

Common Observations in Depressed Speech

- Reduced pitch variability (monotone)
- Slower speech rate
- Longer pauses
- Reduced energy/volume
- Lower F0 (pitch)
- Different formant patterns

Tools & Libraries

SpeechBrain

URL: https://speechbrain.github.io/
GitHub: https://github.com/speechbrain/speechbrain

- Open-source PyTorch toolkit for conversational AI
- 200+ training recipes on 40+ datasets
- Supports: speech recognition, speaker recognition, speech enhancement, emotion recognition
- Pre-trained models on HuggingFace
- Good for feature extraction and model development

Other Relevant Tools

- OpenSMILE - audio feature extraction
- Praat - phonetic analysis
- Librosa - Python audio analysis
- Wav2Vec 2.0 / HuBERT - self-supervised speech representations

---

Research Gaps (Potential Angles)

1. **Generalization across datasets** - models often don't transfer well
2. **Cross-cultural validity** - most research on Western populations
3. **Real-world deployment** - lab conditions vs. real clinical settings
4. **Longitudinal monitoring** - tracking depression over time
5. **Multimodal fusion** - combining audio with text/video
6. **Explainability** - which features drive predictions?
7. **Privacy-preserving methods** - on-device processing

Key Questions to Address in Dissertation

From Advisor:

1. Is there anything commercially available?
2. What are current approaches?
3. How many people in UK are affected?
4. Is there a need? What is the need for AI tools?
5. What is speech formally? How do you measure differences?
6. What statistical methods exist?

Additional Research Questions:

- What features are most predictive of depression?
- How do different ML architectures compare?
- What are the ethical considerations?
- How would this work in clinical practice?

Next Steps

1. [ ] Find and review more recent papers (2020-2025)
2. [ ] Identify exact commercial products (if any)
3. [ ] Deep dive into DAIC-WOZ dataset structure
4. [ ] Review AVEC challenge winning approaches
5. [ ] Explore UK-specific research/applications

References (BibTeX to compile)

```
@inproceedings{gratch2014distress,
  title={The distress analysis interview corpus of human and computer interviews},
  author={Gratch, Jonathan and Artstein, Ron and Lucas, Gale M and Stratou, Giota and Scherer, Stefan and Nazarian, Angela and Wood, Rachel and Boberg, Jill and DeVault, David and Marsella, Stacy and others},
  booktitle={Proceedings of LREC},
  pages={3123--3128},
  year={2014}
}

@inproceedings{devault2014simsensei,
  title={SimSensei kiosk: A virtual human interviewer for healthcare decision support},
  author={DeVault, David and Artstein, Ron and Benn, Grace and others},
  booktitle={Proceedings of AAMAS},
  year={2014}
}

@inproceedings{ringeval2019avec,
  title={AVEC 2019 workshop and challenge: state-of-mind, detecting depression with AI, and cross-cultural affect recognition},
  author={Ringeval, Fabien and Schuller, Björn and Valstar, Michel and others},
  booktitle={Proceedings of AVEC},
  pages={3--12},
  year={2019}
}
```