Speech Features for Depression Detection: A Taxonomy

What is Speech? (Formal Definition)

Speech is a **time-varying acoustic signal** produced by the human vocal apparatus, consisting of:

1. **Phonation** - vocal fold vibration generating the fundamental frequency (F0)
2. **Articulation** - shaping of the vocal tract to produce distinct sounds
3. **Prosody** - suprasegmental features (rhythm, stress, intonation)

Speech can be analyzed in multiple domains:

- **Time domain** - amplitude over time (waveform)
- **Frequency domain** - spectral content (via Fourier transform)
- **Time-frequency domain** - spectrograms, MFCCs

Feature Categories

1. Prosodic Features (Suprasegmental)

1.1 Pitch (Fundamental Frequency, F0)

Definition: The rate of vocal fold vibration, perceived as voice pitch.
Measurement: In Hertz (Hz), typically 85-180 Hz (male), 165-255 Hz (female)

Features extracted:

- Mean F0
- F0 standard deviation
- F0 range (max - min)
- F0 contour (rising/falling patterns)
- Jitter (cycle-to-cycle F0 variation)

Depression association: Lower mean F0, reduced F0 variability (monotone)

1.2 Energy/Intensity

Definition: Amplitude of the speech signal, perceived as loudness.
Measurement: In decibels (dB)

Features extracted:

- Mean energy
- Energy standard deviation
- Energy range
- Shimmer (cycle-to-cycle amplitude variation)

Depression association: Lower overall energy, reduced variation

1.3 Temporal Features

Definition: Timing and rhythm of speech production.

Features extracted:

- Speech rate (syllables/phonemes per second)
- Articulation rate (excluding pauses)
- Pause duration (mean, max, total)
- Pause frequency
- Speech-to-pause ratio
- Response latency

Depression association: Slower rate, longer/more pauses, increased latency

---

2. Spectral Features

2.1 Mel-Frequency Cepstral Coefficients (MFCCs)

Definition: Coefficients representing the short-term power spectrum on a mel scale (approximating human auditory perception).

Typically extracted:

- 13 MFCCs (static)
- Delta MFCCs (first derivative)
- Delta-delta MFCCs (second derivative)
- = 39 features total

Why useful: Captures vocal tract characteristics, widely used in speech recognition

2.2 Formants

Definition: Resonant frequencies of the vocal tract, F1, F2, F3...

Features:

- F1 (related to vowel height, ~300-800 Hz)
- F2 (related to vowel frontness, ~800-2500 Hz)
- F3 (speaker characteristics, ~2000-3500 Hz)
- Formant bandwidth
- Formant transitions

Depression association: Changes in articulation precision

2.3 Other Spectral Features

- **Spectral centroid** - "center of mass" of spectrum
- **Spectral flux** - rate of change in spectrum
- **Spectral rolloff** - frequency below which X% of energy lies
- **Spectral entropy** - randomness/uniformity of spectrum
- **Harmonic-to-Noise Ratio (HNR)** - voice quality measure

---

3. Voice Quality Features

3.1 Jitter

Definition: Cycle-to-cycle variation in fundamental frequency (F0)
Types: Local jitter, RAP, PPQ5
Units: Percentage or absolute (microseconds)

3.2 Shimmer

Definition: Cycle-to-cycle variation in amplitude
Types: Local shimmer, APQ3, APQ5, APQ11
Units: Percentage or dB

3.3 Harmonic-to-Noise Ratio (HNR)

Definition: Ratio of periodic (harmonic) to aperiodic (noise) components
Units: dB
Interpretation: Higher = clearer voice; lower = breathier/hoarser

Depression association: Increased jitter/shimmer, decreased HNR

---

4. Linguistic/Content Features

4.1 Lexical Features

- Word count
- Type-token ratio (vocabulary diversity)
- Use of first-person pronouns
- Negative emotion words
- Absolute terms ("always", "never")

4.2 Semantic Features

- Sentiment scores
- Topic modeling
- LIWC (Linguistic Inquiry and Word Count) categories

Note: Requires transcript (ASR or manual)

---

5. Deep Learning Representations

5.1 Self-Supervised Representations

- **Wav2Vec 2.0** - Facebook/Meta's pre-trained model
- **HuBERT** - Hidden-Unit BERT
- **WavLM** - Microsoft's model

Advantage: Learn features directly from data, often outperform hand-crafted features

5.2 Spectrogram-based

- Log-mel spectrograms → CNN
- Raw waveform → 1D CNN

Statistical Methods for Measuring Differences

Descriptive Statistics

- Mean, median, mode
- Standard deviation, variance
- Skewness, kurtosis
- Range, percentiles

Functionals (Applied to LLDs)

Low-level descriptors (LLDs) computed frame-by-frame are summarized using functionals:

- Mean, std, min, max
- Quartiles (25th, 50th, 75th percentile)
- Linear regression coefficients
- Peaks (number, mean distance)

Statistical Testing

- **T-tests** - compare means between groups
- **Mann-Whitney U** - non-parametric comparison
- **Effect size** (Cohen's d) - magnitude of difference
- **ANOVA** - multiple group comparison

Machine Learning Evaluation

- Accuracy, Precision, Recall, F1-score
- **Concordance Correlation Coefficient (CCC)** - standard for AVEC/regression
- ROC-AUC (classification)
- Mean Absolute Error, RMSE (regression)

Common Feature Extraction Tools

Tool	Description	Language
OpenSMILE	Comprehensive feature extraction	C++
Praat	Phonetic analysis, F0/formants	Praat script
Librosa	Python audio analysis	Python
pyAudioAnalysis	Audio feature extraction	Python
SpeechBrain	Deep learning toolkit	Python
Parselmouth	Praat in Python	Python

Feature Sets / Standards

eGeMAPS (extended Geneva Minimalistic Acoustic Parameter Set)

- 88 features
- Standardized, interpretable
- Used in AVEC challenges

ComParE (Computational Paralinguistics Challenge)

- ~6000+ features
- Brute-force approach
- Used in INTERSPEECH challenges

IS09-IS13 Feature Sets

- INTERSPEECH challenge feature sets
- Various sizes (384-6373 features)

Key References

- Schuller, B., et al. (2016). "The Geneva Minimalistic Acoustic Parameter Set (GeMAPS) for Voice Research and Affective Computing." IEEE TAC.
- Cummins, N., et al. (2015). "A review of depression and suicide risk assessment using speech analysis." Speech Communication.