ChatGPT

Lanskap Model Pra-Latih untuk Estimasi Kepribadian Big Five dari Audio (2025)

Ringkasan Eksekutif

Dalam riset ini kami meninjau belasan model pra-terlatih (baik berbasis Transformer maupun non-Transformer) untuk tugas estimasi kepribadian Big Five dari suara. Model self-supervised speech seperti Wav2Vec2 dan turunannya (HuBERT, WavLM, UniSpeech-SAT) menghasilkan representasi audio yang kaya dan unggul – misalnya, fitur Wav2Vec2 terbukti lebih informatif daripada fitur manual eGeMAPS untuk prediksi personality. Selain itu, model khusus paralinguistik seperti TRILLsson (Google) mencapai performa mendekati state-of-the-art dengan ukuran model jauh lebih kecil. Laporan ini menyajikan matriks perbandingan menyeluruh berbagai model (arsitektur, dimensi embedding, jumlah parameter, kecepatan inferensi, dukungan fine-tuning, dll.), skor rekomendasi untuk use-case Big Five, serta shortlist 2–4 model terbaik. Kami merekomendasikan pipeline implementasi yang mencakup ekstraksi embedding audio (16 kHz mono; normalisasi; deteksi suara/VAD), agregasi per individu (contoh: mean+std pooling), lalu pelatihan regressor multi-output (dengan opsi fine-tuning parameter-efisien, misal LoRA). Evaluasi dilakukan secara hati-hati (GroupKFold per speaker, metrik korelasi Pearson/Spearman dan error). Kami juga membahas risiko bias, perbedaan domain data, serta pentingnya pertimbangan etika dalam prediksi *apparent personality*.

A. Perbandingan Model Pra-Latih (Matriks)

Model	Arsitektur	Dim. Emb.	#Param	Input (Hz)	Prapelatihan (objektif & data)	Multi?	Lisensi	S
Wav2Vec 2.0 (Base)	CNN + Transformer encoder (12 layers)	768	≈95M	16 kHz	Mask prediction (contrastive loss) on LibriSpeech 960h (English)	False	Apache 2.0	H f w b
HuBERT (Base)	CNN + Transformer encoder (12 layers)	768	≈95M (Base), 317M (Large)	16 kHz	Predict masked cluster labels on LibriSpeech 960h (Base); Libri-Light 60k (Large)	False	Apache 2.0 (Facebook)	H f h l

Model	Arsitektur	Dim. Emb.	#Param	Input (Hz)	Prapelatihan (objektif & data)	Multi?	Lisensi	S
WavLM (Base+ / Large)	CNN + Transformer encoder (12 atau 24 layers)	768 (Base), 1024 (Large)	≈94M (Base), 317M (Large)	16 kHz	Masked prediction + denoising on Libri-Light 60k + others (Base+); +GigaSpeech, VoxPopuli (Large)	False	MIT	H r w w
Data2Vec Audio (Base)	CNN + Transformer encoder (12 layers)	768	≈95M	16 kHz	Masked latent prediction (self-distillation) on LibriSpeech 960h	False	Apache 2.0	H f c a
UniSpeech-SAT (Base+)	CNN + Transformer encoder (12 layers)	768	≈95M (Base), lebih besar ≈317M	16 kHz	Masked prediction + speaker-aware contrastive on Libri-Light 60k + GigaSpeech + VoxPopuli	False	MIT (Microsoft)	H r u s b
Whisper (Encoder)	Encoder-decoder Transformer (gunakan encoder)	512 (Base encoder output)	74M (Base enc/dec total ~95M)	16 kHz	Supervised ASR + translation (680k jam transkrip multi-bahasa)	True	Apache 2.0	H w
XLS-R (Wav2Vec2-XL)	CNN + Transformer encoder (24 layers)	1024	300M (XLS-R-300M); juga 1B & 2B	16 kHz	Masked contrastive (seperti W2V2) on 436k jam multi-bahasa (128 lang)	True	Apache 2.0	H f w x
Wav2Vec-Switch	CNN + Transformer (varian W2V2)	768	~95M	16 kHz	Contrastive dgn pasangan clean-noisy on LibriSpeech + augment noise	False	Apache 2.0 (riset)	F I

Model	Arsitektur	Dim. Emb.	#Param	Input (Hz)	Prapelatihan (objektif & data)	Multi?	Lisensi	S
ECAPA-TDNN	TDNN + Res2Net & SE attn	192 [1]	~6M	16 kHz	Speaker ID (VoxCeleb1+2) – classific.	N/A	Apache 2.0 (SpeechBrain)	H s s e v
X-vector (DNN)	TDNN (5-layer) + stats pooling	512 [1]	~4M	8/16 kHz	Speaker ID (SRE) – classific.	N/A	Apache 2.0 (Kaldi)	K
TRILL/ TRILLsson	Conformer 600M (CAP12) distilled ke EfficientNet/ Transformer	512–2048 (variasi)	22M (small) – 80M (large) [2]	16 kHz	Self-sup. (CAP12) lalu distill [2] on YT-U 900Mh (teacher); Libri-light 53k + AudioSet 5k	True	Apache 2.0	T g r s b
YAMNet	MobileNet V1 depthwise CNN	1024	~3.8M	16 kHz	Audio event classific. (AudioSet 521 cls)	N/A	Apache 2.0	T T
PANNs (CNN14)	CNN (14-layer, VGG-like)	2048 [4]	~79M	16 kHz	Audio tagging (AudioSet 5k jam)	N/A	MIT	C Z (
VGGish	CNN (VGG-variant)	128	~62M	16 kHz	YouTube-8M audio (70M vid)	N/A	Apache 2.0 (code)	T (

Model	Arsitektur	Dim. Emb.	#Param	Input (Hz)	Prapelatihan (objektif & data)	Multi?	Lisensi	S
OpenL3 (Audio)	CNN (Music/ VGG) + proj.	512 atau 6144	~5M (512-d)	16 atau 48 kHz	Audio-Visual (YouTube) – align embed	N/A	MIT	c l (
eGeMAPS (OpenSMILE)	Hand-crafted feature set	88	0	16 kHz	N/A (expert-designed descriptors)	Yes	Apache 2.0 (toolkit)	c c e

Catatan: *Skor rekomendasi 1–5 di atas bersifat subyektif, menilai kecocokan tiap model untuk prediksi Big Five dari audio.* Model berbasis Transformer self-supervised (skor 4–5) umumnya memberikan fitur paling kaya untuk tugas ini, sedangkan model lama/khusus (skor lebih rendah) bisa dijadikan baseline atau pelengkap.

B. Shortlist Model Terbaik & Justifikasi Teknis

- **WavLM Large:** Model Transformer self-supervised berkapasitas besar (~317M parameter) dengan pretraining pada ~94 ribu jam data. WavLM menggabungkan pemodelan konten ujaran dan karakteristik pembicara dalam representasinya, sehingga sangat sesuai untuk tugas prediksi personality dari vokal. Kinerjanya unggul di berbagai task (ASR, diarization, emotion) tanpa fine-tuning penuh, menunjukkan daya generalisasi yang tinggi. Lisensi MIT (open source) memudahkan penggunaan riset maupun pengembangan lanjut.

- **UniSpeech-SAT Base+:** Model Transformer self-supervised berukuran sedang (~95M parameter) dengan pretraining *speaker-aware*, yaitu secara eksplisit memodelkan perbedaan antar-pembicara dalam latihannya. UniSpeech-SAT terbukti mencapai performa terbaik pada benchmark tugas terkait pembicara (speaker verification/diarization), menandakan kemampuannya menangkap ciri-ciri vokal yang stabil. Hal ini membuatnya sangat relevan untuk estimasi trait kepribadian dari suara. Model ini relatif ringan dibanding WavLM Large, sehingga lebih efisien, dan berlisensi MIT (bebas untuk riset).

- **TRILLsson (Large Distilled):** Model representasi paralinguistik universal hasil distilasi dari model Conformer 600M (Google CAP12). TRILLsson versi terbesar (~80M param, ~314 MB) mencapai ~96% performa model besar pada berbagai task paralinguistik, bahkan melampaui wav2vec2 pada 6 dari 7 tugas dalam NOSS benchmark [3]. Representasi TRILLsson sangat peka terhadap informasi non-semantik (intonasi, emosi, dll) yang penting untuk *apparent personality*. Keunggulannya adalah efisiensi: model lebih kecil, inferensi cepat (real-time di CPU), dan open source (Apache 2.0). Tantangan minor: implementasi aslinya di TensorFlow – perlu konversi ke PyTorch/HuggingFace untuk integrasi mudah.

- **ECAPA-TDNN:** Model embedding speaker non-Transformer berukuran sangat ringan (~6M param) namun dengan arsitektur canggih (Res2Net + attentional pooling). ECAPA-TDNN mencapai *state-of-the-art* di tugas verifikasi pembicara VoxCeleb dengan embedding berdimensi 192 yang sangat padat informasi suara. Meskipun dilatih untuk identitas pembicara, embedding

ECAPA menangkap ciri vokal stabil (timbre, gaya bicara) yang dapat berhubungan dengan trait kepribadian. Model ini cocok sebagai baseline efisien, atau dikombinasikan dengan fitur SSL untuk meningkatkan robustnes. Lisensinya Apache 2.0 (open source), bebas digunakan baik untuk riset akademik maupun aplikasi.

Lisensi: Keempat model di atas tersedia secara open-source dengan lisensi permisif (MIT/Apache), sehingga dapat digunakan tanpa kendala berarti untuk keperluan akademis/non-komersial.

C. Rencana Pipeline Implementasi

Pre-processing & Ekstraksi Fitur: Semua audio dire-sampling ke 16 kHz mono dan dinormalisasi (misal, normalisasi loudness) untuk konsistensi. Disarankan menerapkan Voice Activity Detection (VAD) guna memotong segmen hening atau noise berat. Dari tiap rekaman (contoh: klip 15 detik FI-V2), ekstrak potongan-potongan pendek berdurasi 3–6 detik dengan overlap ~50% untuk menangkap variasi temporal. Setiap segmen diubah menjadi embedding menggunakan model pra-latih pilihan (contoh dengan HuggingFace):

```
import torchaudio
from transformers import Wav2Vec2Processor, Wav2Vec2Model

# Load pre-trained model and processor
processor = Wav2Vec2Processor.from_pretrained("facebook/wav2vec2-base-960h")
model = Wav2Vec2Model.from_pretrained("facebook/wav2vec2-base-960h")
model.eval()

# Load audio and preprocess
waveform, sr = torchaudio.load("input_audio.wav")
waveform = torchaudio.functional.resample(waveform, sr, 16000)
# resample to 16 kHz mono
inputs = processor(waveform.squeeze(), sampling_rate=16000,
return_tensors="pt", padding=True)

# Extract features (last hidden state) and apply mean pooling
with torch.no_grad():
    outputs = model(**inputs)
embedding = outputs.last_hidden_state.mean(dim=1).squeeze().numpy()
```

Embedding segmen dapat disimpan (misal ke format `.npy`) untuk mempercepat eksperimen. Selanjutnya, agregasikan embedding seluruh segmen per individu menjadi satu vektor representasi: metode sederhana adalah **mean+std pooling** (menghitung rata-rata dan standar deviasi dari semua embedding segmen, lalu dikonstruksi jadi vektor 2×dimensi). Alternatif lebih canggih termasuk *self-attention pooling* atau NetVLAD untuk memberi bobot lebih pada segmen informatif – hal ini bisa diuji dalam ablasi.

Training & Fine-tuning: Untuk memprediksi skor Big Five (5 dimensi kontinu) digunakan model regresi multi-output. Sebagai baseline yang kuat dan mudah, dapat dilatih regresor linear (Ridge/ElasticNet) atau ensemble (Random Forest, XGBoost) menggunakan vektor embedding per individu sebagai input features. Contoh dengan scikit-learn (Ridge Regression):

```
import numpy as np
from sklearn.linear_model import Ridge

X_train = np.load("train_embeddings.npy")   # shape (N_train, feat_dim)
y_train = np.load("train_traits.npy")       # shape (N_train, 5)
X_val = np.load("val_embeddings.npy")
y_val = np.load("val_traits.npy")

reg = Ridge(alpha=1.0)
reg.fit(X_train, y_train)
preds_val = reg.predict(X_val)
```

Sebagai alternatif, **fine-tuning end-to-end** model pra-latih dapat dicoba untuk peningkatan akurasi: misalnya menambahkan layer dense 5-output di atas embedding terakhir model dan melatih terbatas (freezing sebagian besar parameter). Karena dataset tidak terlalu besar, direkomendasikan *parameter-efficient fine-tuning (PEFT)* seperti LoRA atau adapter pada beberapa layer atas model. PEFT memungkinkan model belajar penyesuaian minor terhadap domain personality tanpa overfitting berlebihan, dengan overhead parameter kecil.

Evaluasi & Kalibrasi: Evaluasi dilakukan dengan *cross-validation* yang menjaga pemisahan speaker. Gunakan skema **GroupKFold** dengan ID pembicara sebagai grup, sehingga klip dari vlogger yang sama tidak bocor ke fold pelatihan dan validasi. Metrik utama: korelasi Pearson's r dan Spearman's ρ antara skor prediksi vs skor ground-truth untuk setiap trait, yang mencerminkan sejauh mana model menangkap peringkat relatif kepribadian. Selain itu, hitung error absolut rata-rata (MAE) dan RMSE untuk mengukur kesalahan numerik. Untuk menilai kalibrasi, dapat diplot diagram prediksi vs aktual per trait; idealnya model tidak bias sistematis (misal tidak selalu memprediksi lebih rendah/tinggi dari rata-rata).

Ablasi Eksperimen: Beberapa eksperimen disarankan untuk menguji kontribusi komponen pipeline: *(i)* **Perbandingan embedding model A vs B:** ekstraksi fitur dari dua model pra-latih berbeda (misal Wav2Vec2 vs WavLM) di dataset yang sama, latih regresor yang sama, bandingkan performa (korelasi & error). *(ii)* **Strategi pooling:** coba pooling sederhana (mean, mean+std) versus attention pooling trainable – lihat pengaruhnya terhadap korelasi, apakah informasi temporal tambahan membantu. *(iii)* **Panjang window segmen:** bandingkan durasi segmen (misal 2s vs 5s vs 10s) – durasi lebih pendek menangkap detail tapi mungkin kehilangan konteks, durasi panjang sebaliknya. *(iv)* **Fine-tuning vs frozen:** bandingkan regresor linear pada fitur fixed vs fine-tuning (atau LoRA) model – apakah ada peningkatan signifikan membiarkan model beradaptasi ke data personality. Hasil ablasi akan memberi wawasan fitur/konfigurasi mana yang paling berpengaruh.

Estimasi Runtime: Pipeline di atas dapat dijalankan dengan sumber daya GPU menengah (misal NVIDIA RTX 3060 12GB). Ekstraksi embedding untuk ~10 ribu klip (15 detik per klip, total ~40 jam audio) menggunakan model Transformer base (95M param) diperkirakan memakan waktu ±30–60 menit pada 1 GPU (dengan batch processing). Pelatihan regresor linear/MLP sangat cepat (hitungan detik hingga menit). Jika melakukan fine-tuning model end-to-end, waktu training bisa meningkat (misal beberapa jam per epoch, tergantung batch size dan learning rate) – namun dengan PEFT/LoRA, bisa hanya ~1–2 jam untuk konvergen karena hanya sebagian kecil parameter dilatih.

D. Risiko, Etika, dan Validitas

- **Apparent ≠ Real Personality:** Perlu ditekankan bahwa label kepribadian dalam dataset FI-V2 adalah *apparent personality* – penilaian observasi orang lain. Model yang dilatih di sini belajar memprediksi persepsi tersebut, bukan kepribadian sejati individu. Hasil model mungkin mencerminkan stereotip atau bias persepsi para annotator. Karena itu, prediksi tidak boleh diartikan sebagai kebenaran mutlak tentang individu, melainkan *bagaimana individu terdengar* ke pendengar rata-rata.

- **Bias Annotator & Demografis:** Data kepribadian aparen rentan bias – misal, suara dengan aksen tertentu mungkin dinilai kurang *agreeable* secara stereotipik. Jika metadata demografis tersedia, perlu audit apakah model memiliki bias sistematis terhadap gender, aksen, atau kelompok etnis. Mitigasi: bisa dengan teknik debiasing atau penyeimbangan data (upsample kelompok minoritas) dalam pelatihan, serta dengan melakukan **post-hoc** adjustment pada output model.

- **Perbedaan Domain:** Dataset latih (vlogger YouTube, monolog terencana) mungkin tidak mewakili domain aplikasi sesungguhnya (contoh: percakapan spontan, pidato formal, atau suara telepon). **Domain shift** ini dapat menyebabkan penurunan akurasi saat model diaplikasikan ke data dunia nyata. Mitigasi: lakukan fine-tuning tambahan (jika ada data target domain), atau setidaknya uji model di beberapa sampel target untuk cek generalisasi. Penggunaan augmentasi data (menambahkan noise, reverb, perubahan pitch) saat training juga bisa meningkatkan robustnes model lintas kondisi akustik.

- **Overfitting & Validitas Prediksi:** Tugas prediksi Big Five dari audio memiliki sinyal yang lemah (hanya ~r=0.2–0.3 pada usaha terbaik). Ada risiko model *overfit* ke dataset sempit tanpa benar-benar menangkap pola yang bermakna (misal, bisa saja model mengenali identitas speaker dan *mapping* langsung ke rating jika ada kebocoran identitas). Untuk validitas, perlu memastikan evaluasi *speaker-independent* (seperti yang dilakukan dengan GroupKFold). Juga dianjurkan menghitung interval kepercayaan untuk metrik (misal dengan bootstrap), mengingat korelasi yang didapat cenderung rendah – ini memberi gambaran ketidakpastian model.

- **Privasi & Persetujuan:** Suara merupakan data biometrik yang mengandung informasi sensitif (emosi, kondisi kesehatan, dll). Prediksi kepribadian adalah inferensi yang invasif dan berpotensi keliru. Oleh karena itu, dalam penerapan nyata, penting memastikan subjek memberikan **persetujuan** (*informed consent*) sebelum datanya digunakan untuk analisis kepribadian. Data audio yang digunakan sebaiknya bersifat publik atau telah dianonimkan. Penyimpanan embedding suara juga perlu memperhatikan keamanan, karena bisa disalahgunakan untuk mengidentifikasi individu. Sebagai peneliti, transparansi kepada subjek mengenai sifat perkiraan (bahwa ini prediksi otomatis, bisa bias, dan bukan penilaian definitif) adalah aspek etis yang krusial.

E. Lampiran


```
[
  {
    "name": "Wav2Vec 2.0 (Base)",
    "architecture": "CNN + Transformer encoder (12 layers)",
    "embedding_dim": 768,
```

```
    "num_params": "≈95M",
    "input_sampling_rate": "16 kHz",
    "pretraining_data": "LibriSpeech 960h (English)",
    "pretraining_objective": "Mask prediction (contrastive loss)",
    "multilingual": false,
    "license": "Apache 2.0",
    "model_source": "Hugging Face: facebook/wav2vec2-base-960h",
    "last_updated": "2020 (paper), HF model 2021",
    "inference_speed": "Sedang (real-time di GPU, lebih lambat di CPU)",
    "memory_footprint": "≈370 MB (fp32 model)",
    "supports_PEFT": true,
    "personality_evidence": "Digunakan untuk apparent personality
prediction",
    "pros": "Representasi suara berkualitas; terbukti efektif untuk tugas
emosi/kepribadian",
    "cons": "Perlu fine-tuning/regresi; cukup berat",
    "recommendation_score": 4
  },
  {
    "name": "HuBERT (Base)",
    "architecture": "CNN + Transformer encoder (12 layers)",
    "embedding_dim": 768,
    "num_params": "≈95M (Base), 317M (Large)",
    "input_sampling_rate": "16 kHz",
    "pretraining_data": "LibriSpeech 960h (Base); Libri-Light 60k (Large)",
    "pretraining_objective": "Predict cluster labels (masked)",
    "multilingual": false,
    "license": "Apache 2.0",
    "model_source": "Hugging Face: facebook/hubert-base-ls960",
    "last_updated": "2021",
    "inference_speed": "Sedang (mirip W2V2)",
    "memory_footprint": "≈370 MB (Base fp32)",
    "supports_PEFT": true,
    "personality_evidence": "Belum ada studi langsung; fitur SSL mirip W2V2,
kemungkinan efektif",
    "pros": "Fitur suara umum yang baik; kinerja kuat di berbagai tugas",
    "cons": "Tidak khusus untuk paralinguistik; perlu fine-tuning ke task",
    "recommendation_score": 4
  },
  {
    "name": "WavLM (Base+ / Large)",
    "architecture": "CNN + Transformer encoder (12 or 24 layers)",
    "embedding_dim": "768 (Base), 1024 (Large)",
    "num_params": "≈94M (Base), 317M (Large)",
    "input_sampling_rate": "16 kHz",
    "pretraining_data": "Libri-Light 60k + others (Base+); +10k GigaSpeech +
24k VoxPopuli (Large)",
    "pretraining_objective": "Masked prediction + denoising",
    "multilingual": false,
    "license": "MIT",
    "model_source": "Hugging Face: microsoft/wavlm-base, wavlm-large",
```

```
    "last_updated": "2021 (released 2022)",
    "inference_speed": "Sedang/Lambat (Large berat, Base sedang)",
    "memory_footprint": "≈370 MB (Base), 1.2 GB (Large)",
    "supports_PEFT": true,
    "personality_evidence": "SOTA di tugas speaker; robust untuk
paralinguistik",
    "pros": "Pretraining full-stack (konten+pembicara); unggul di banyak
tugas",
    "cons": "Varian besar cukup berat; lisensi non-komersial (riset saja)",
    "recommendation_score": 5
  },
  {
    "name": "Data2Vec Audio (Base)",
    "architecture": "CNN + Transformer encoder (12 layers)",
    "embedding_dim": 768,
    "num_params": "≈95M",
    "input_sampling_rate": "16 kHz",
    "pretraining_data": "LibriSpeech 960h",
    "pretraining_objective": "Masked latent prediction (self-distillation)",
    "multilingual": false,
    "license": "Apache 2.0",
    "model_source": "Hugging Face: facebook/data2vec-audio-base",
    "last_updated": "2022",
    "inference_speed": "Sedang (setara W2V2 Base)",
    "memory_footprint": "≈370 MB",
    "supports_PEFT": true,
    "personality_evidence": "Belum ada bukti langsung; fitur SSL umum mirip
wav2vec2",
    "pros": "Objektif terpadu lintas modalitas; fitur serbaguna yang baik",
    "cons": "Tidak khusus untuk paralinguistik; baru & kurang teruji",
    "recommendation_score": 4
  },
  {
    "name": "UniSpeech-SAT (Base+)",
    "architecture": "CNN + Transformer encoder (12 layers)",
    "embedding_dim": 768,
    "num_params": "≈95M (Base), up to 317M",
    "input_sampling_rate": "16 kHz",
    "pretraining_data": "Libri-Light 60k + 10k GigaSpeech + 24k VoxPopuli",
    "pretraining_objective": "Masked prediction + speaker-aware contrastive",
    "multilingual": false,
    "license": "MIT",
    "model_source": "Hugging Face: microsoft/unispeech-sat-base-960h",
    "last_updated": "2021",
    "inference_speed": "Sedang (sedikit > W2V2 Base)",
    "memory_footprint": "≈370 MB (Base)",
    "supports_PEFT": true,
    "personality_evidence": "Dirancang tangkap ciri pembicara, potensial
tinggi untuk personality",
    "pros":
"Unggul dalam fitur terkait pembicara; kemungkinan menangkap ciri suara",
```

```
    "cons": "Belum banyak diadopsi; terutama untuk tugas speaker ID",
    "recommendation_score": 5
  },
  {
    "name": "Whisper (Encoder)",
    "architecture": "Encoder-decoder Transformer (use encoder only)",
    "embedding_dim": "512 (Base model encoder output)",
    "num_params": "74M (Base encoder; ~95M incl. decoder)",
    "input_sampling_rate": "16 kHz",
    "pretraining_data": "680k hours labeled multilingual speech",
    "pretraining_objective": "Supervised ASR + translation",
    "multilingual": true,
    "license": "Apache 2.0",
    "model_source": "Hugging Face: openai/whisper-base",
    "last_updated": "2022",
    "inference_speed": "Sedang (Base ~ realtime GPU)",
    "memory_footprint": "≈300 MB (encoder fp32)",
    "supports_PEFT": "Partial (PEFT encoder possible)",
    "personality_evidence": "Belum ada bukti; mungkin tangkap prosodi via
pretraining besar",
    "pros": "Tangguh untuk multi-bahasa & noise; data latih sangat besar",
    "cons": "Fokus pada konten kata; bisa abaikan isyarat paralinguistik",
    "recommendation_score": 3
  },
  {
    "name": "XLS-R (Wav2Vec2-XL Multi)",
    "architecture": "CNN + Transformer encoder (24 layers)",
    "embedding_dim": 1024,
    "num_params": "300M (XLS-R-300M); 1B & 2B variants",
    "input_sampling_rate": "16 kHz",
    "pretraining_data": "436k hours multilingual speech",
    "pretraining_objective": "Masked contrastive (wav2vec2)",
    "multilingual": true,
    "license": "Apache 2.0",
    "model_source": "Hugging Face: facebook/wav2vec2-xls-r-300m",
    "last_updated": "2021",
    "inference_speed": "Lambat (model besar, perlu GPU)",
    "memory_footprint": "≈1.2 GB (300M fp32)",
    "supports_PEFT": true,
    "personality_evidence": "Belum ada bukti; fitur lintas bahasa robust ke
aksen",
    "pros": "Mencakup banyak bahasa; representasi tangguh lintas domain",
    "cons": "Sangat besar; overkill utk English saja; butuh compute tinggi",
    "recommendation_score": 4
  },
  {
    "name": "Wav2Vec-Switch",
    "architecture": "CNN + Transformer (wav2vec2 variant)",
    "embedding_dim": 768,
    "num_params": "~95M",
    "input_sampling_rate": "16 kHz",
```

```
    "pretraining_data": "LibriSpeech + noisy aug pairs",
    "pretraining_objective": "Contrastive w/ clean-noisy pairs",
    "multilingual": false,
    "license": "Apache 2.0 (research code)",
    "model_source": "ICASSP 2022 paper",
    "last_updated": "2022",
    "inference_speed": "Sedang (mirip W2V2 Base)",
    "memory_footprint": "≈370 MB",
    "supports_PEFT": "Likely",
    "personality_evidence":
"Tidak ada; fitur robust noise bisa bantu di data noisy",
    "pros": "Ketahanan noise lebih baik; arsitektur sama dgn W2V2",
    "cons": "Belum ada model siap pakai; manfaat kecil kecuali banyak noise",
    "recommendation_score": 3
  },
  {
    "name": "ECAPA-TDNN",
    "architecture": "TDNN (Res2Net & SE attention)",
    "embedding_dim": "192",
    "num_params": "~6M",
    "input_sampling_rate": "16 kHz",
    "pretraining_data": "VoxCeleb1+2 (speaker ID)",
    "pretraining_objective": "Classification (speaker)",
    "multilingual": "N/A",
    "license": "Apache 2.0",
    "model_source": "HF: speechbrain/spkrec-ecapa-voxceleb",
    "last_updated": "2020",
    "inference_speed": "Cepat (CPU-friendly)",
    "memory_footprint": "≈24 MB",
    "supports_PEFT": "No",
    "personality_evidence": "Belum ada; tangkap timbre suara stabil, mungkin
terkait trait",
    "pros": "Ringan; embedding pembicara terkini",
    "cons": "Dioptimalkan utk speaker ID; tdk dilatih utk trait/prosodi",
    "recommendation_score": 3
  },
  {
    "name": "X-vector (DNN)",
    "architecture": "TDNN (5-layer) + stats pooling",
    "embedding_dim": "512",
    "num_params": "~4M",
    "input_sampling_rate": "8 kHz or 16 kHz",
    "pretraining_data": "Speaker ID (SRE)",
    "pretraining_objective": "Classification (speaker)",
    "multilingual": "N/A",
    "license": "Apache 2.0 (Kaldi)",
    "model_source": "Kaldi (n/a on HF)",
    "last_updated": "2018",
    "inference_speed": "Cepat",
    "memory_footprint": "≈16 MB",
    "supports_PEFT": "No",
```

```
    "personality_evidence": "Pernah jd baseline di studi paralinguistik, tdk
spesifik Big5",
    "pros": "Ringan & sederhana; baseline pembicara tangguh",
    "cons": "Kalah oleh model terbaru; kapasitas terbatas",
    "recommendation_score": 2
  },
  {
    "name": "TRILL/TRILLsson",
    "architecture": "Conformer (600M) distilled to EffNet/Transformer",
    "embedding_dim": "512 to 2048",
    "num_params": "22M (small) to 80M (large)",
    "input_sampling_rate": "16 kHz",
    "pretraining_data": "YT-U 900Mh (teacher); 58k public
(LibriLight+AudioSet)",
    "pretraining_objective": "Self-supervised (teacher) then distillation",
    "multilingual": true,
    "license": "Apache 2.0",
    "model_source": "TFHub: google/nonsemantic-speech-benchmark",
    "last_updated": "2022",
    "inference_speed": "Cepat (model kecil realtime CPU)",
    "memory_footprint": "22MB to 314MB",
    "supports_PEFT": "No",
    "personality_evidence": "Fitur paralinguistik unggul; kalahkan wav2vec2
di emosi",
    "pros": "Sangat baik untuk tugas non-semantik; tersedia model kecil",
    "cons": "Format TF (perlu konversi ke PyTorch); tdk untuk konten
leksikal",
    "recommendation_score": 5
  },
  {
    "name": "YAMNet",
    "architecture": "MobileNet V1 depthwise CNN",
    "embedding_dim": "1024",
    "num_params": "~3.8M",
    "input_sampling_rate": "16 kHz",
    "pretraining_data": "AudioSet (~2M clips, 521 classes)",
    "pretraining_objective": "Audio event classification",
    "multilingual": "N/A",
    "license": "Apache 2.0",
    "model_source": "TFHub / TF Lite",
    "last_updated": "2019",
    "inference_speed": "Sangat cepat (on-device)",
    "memory_footprint": "≈15 MB",
    "supports_PEFT": "No",
    "personality_evidence": "Tidak ada; tangkap audio umum (sedikit info
pembicara)",
    "pros": "Ringan; pengenalan bunyi umum",
    "cons": "Tidak dioptimalkan utk ciri suara; embedding kasar",
    "recommendation_score": 2
  },
  {
```

```
    "name": "PANNs (CNN14)",
    "architecture": "CNN (14-layer, VGG-like)",
    "embedding_dim": "2048",
    "num_params": "~79M",
    "input_sampling_rate": "16 kHz",
    "pretraining_data": "AudioSet 5000h",
    "pretraining_objective": "Audio tagging (multi-label)",
    "multilingual": "N/A",
    "license": "MIT",
    "model_source": "GitHub/Zenodo",
    "last_updated": "2020",
    "inference_speed": "Sedang (CNN14 berat di CPU)",
    "memory_footprint": "≈300 MB",
    "supports_PEFT": "No",
    "personality_evidence": "Tidak ada; fitur audio umum, sedikit info
speech",
    "pros": "Embedding audio generik berkualitas; terbukti di banyak tugas
audio",
    "cons": "Besar & lambat, kurang spesifik suara; perlu reduksi dimensi",
    "recommendation_score": 3
  },
  {
    "name": "VGGish",
    "architecture": "CNN (VGG-like)",
    "embedding_dim": 128,
    "num_params": "~62M",
    "input_sampling_rate": "16 kHz",
    "pretraining_data": "YouTube-8M (audio subset)",
    "pretraining_objective": "Audio classification (embedding for
YouTube-8M)",
    "multilingual": "N/A",
    "license": "Apache 2.0",
    "model_source": "TensorFlow (Google)",
    "last_updated": "2017",
    "inference_speed": "Sedang",
    "memory_footprint": "≈250 MB",
    "supports_PEFT": "No",
    "personality_evidence": "Dipakai di multimodal personality research
(baseline)",
    "pros": "Embedding terkenal; ringkas (128-dim)",
    "cons": "Arsitektur usang; detail embedding terbatas",
    "recommendation_score": 2
  },
  {
    "name": "OpenL3 (Audio)",
    "architecture": "CNN (music/VGG) + projection",
    "embedding_dim": "512 or 6144",
    "num_params": "~5M (512-d model)",
    "input_sampling_rate": "16 kHz or 48 kHz",
    "pretraining_data": "Audio+Video (YouTube) multimodal",
    "pretraining_objective": "Cross-modal alignment",
```

```
    "multilingual": "N/A",
    "license": "MIT",
    "model_source": "openl3 PyPI/Repo",
    "last_updated": "2019",
    "inference_speed": "Sedang",
    "memory_footprint": "≈20 MB (512-d)",
    "supports_PEFT": "No",
    "personality_evidence": "Tidak ada; tangkap semantik audio high-level",
    "pros": "Manfaatkan pembelajaran audio-visual; embed kecil/besar
tersedia",
    "cons": "Embedding 6144-dim kurang praktis; tdk dilatih khusus utk
voice",
    "recommendation_score": 3
  },
  {
    "name": "eGeMAPS (OpenSMILE)",
    "architecture": "Hand-crafted features",
    "embedding_dim": 88,
    "num_params": 0,
    "input_sampling_rate": "16 kHz",
    "pretraining_data": "N/A",
    "pretraining_objective": "N/A",
    "multilingual": true,
    "license": "Apache 2.0 (toolkit)",
    "model_source": "openSMILE config",
    "last_updated": "2016",
    "inference_speed": "Sangat cepat",
    "memory_footprint": "N/A",
    "supports_PEFT": "N/A",
    "personality_evidence": "Baseline umum apparent personality",
    "pros": "Mudah diinterpretasi; biaya komputasi rendah",
    "cons": "Kapasitas representasi terbatas; akurasi rendah",
    "recommendation_score": 2
  }
]
```


Seed Queries (Penelusuran):
- "apparent personality from speech dataset Big Five"
- "First Impressions V2 audio personality prediction"
- "wav2vec2 paralinguistic regression Big Five"
- "ECAPA-TDNN speaker embedding personality"
- "TRILL TRILLsson Google paralinguistic"
- "Big Five OCEAN speech dataset features"
- "HuBERT UniSpeech speaker-aware SSL"

Log Verifikasi (Klaim & Sumber):
- *Wav2Vec2 vs eGeMAPS untuk personality*: Barchi et al. (2023) menunjukkan fitur wav2vec 2.0 paling berguna untuk prediksi Big Five, mengungguli fitur eGeMAPS (akses 19 Sep 2025).
- *Parameter Wav2Vec2 Base/Large*: Menurut Kerpicci et al. (2023), Wav2Vec2-Base memiliki ~95 juta parameter, Large ~317 juta (diakses 19 Sep 2025).
- *Embedding ECAPA vs X-vector*: Li et al. (2023) menyebut dimensi embedding ECAPA-TDNN = 192 dan x-

vector = 512 [1] (akses 19 Sep 2025).
- *Model TRILLsson & kinerjanya*: Shor et al. (2022) merilis TRILLsson (22 MB s.d. 314 MB) dengan akurasi 90–96% dari model 600M, mengungguli wav2vec2 pada 6 dari 7 task paralinguistik [3] (akses 19 Sep 2025).
- *UniSpeech-SAT untuk ciri pembicara*: Chen et al. (2022) melaporkan UniSpeech-SAT (speaker-aware SSL) paling unggul di tugas speaker dibanding model SSL lain (akses 19 Sep 2025).
- *Dataset ChaLearn First Impressions*: Ponce-Lopez et al. (2016) memperkenalkan dataset 10.000 video vlog (15 detik) dengan label Big Five apparent personality (akses 19 Sep 2025).

Daftar Pustaka:

1. V. Ponce-Lopez et al. *"ChaLearn LAP 2016: First Round Challenge on First Impressions - Dataset and Results."* ECCV Workshops, 2016.
2. R. Barchi et al. *"Apparent personality prediction from speech using expert features and wav2vec 2.0."* Proc. SMM Workshop @ Interspeech 2023.
3. A. Baevski et al. *"wav2vec 2.0: A framework for self-supervised learning of speech representations."* NeurIPS 2020.
4. W.-N. Hsu et al. *"HuBERT: Self-Supervised Speech Representation Learning by Masked Prediction of Hidden Units."* Trans. ACL, 2021.
5. S. Chen et al. *"WavLM: Large-Scale Self-Supervised Pre-Training for Full Stack Speech Processing."* arXiv:2110.13900, Oct 2021.
6. J. Shor et al. *"TRILLsson: Distilled Universal Paralinguistic Speech Representations."* arXiv: 2203.00236 / Proc. ICASSP 2022 [3].
7. B. Desplanques, J. Thienpondt, K. Demuynck. *"ECAPA-TDNN: Emphasized Channel Attention, Propagation and Aggregation in TDNN-Based Speaker Verification."* Proc. Interspeech 2020.
8. M. Kerpicci et al. *"Application of Knowledge Distillation to Multi-task Speech Representation Learning."* arXiv:2210.16611, v2 May 2023.
9. F. Eyben et al. *"The Geneva Minimalistic Acoustic Parameter Set (GeMAPS) for Voice Research."* Proc. Interspeech 2016.

---

[1] Phonetic-aware Speaker Embedding for far-field speaker verification
https://arxiv.org/html/2311.15627

[2] [3] [2203.00236] TRILLsson: Distilled Universal Paralinguistic Speech Representations
https://ar5iv.labs.arxiv.org/html/2203.00236

[4] GitHub - qiuqiangkong/audioset_tagging_cnn
https://github.com/qiuqiangkong/audioset_tagging_cnn