BINUS University

Academic Career: ***Undergraduate / ~~Master / Doctoral / BINUS Online/ Professional~~*)***	**Class Program:** ***Regular/ ~~Global~~ Class*)***
☐ **Mid Exam** ☐ **Others Exam : _____** ☒ **Final Exam**	**Term : Odd / ~~Even / Compact~~*)** **Period (Only for *BINUS Online/ Master*): 1 / 2*)**
☒ **Kemanggisan** ☐ **Senayan** ☐ **Semarang** ☐ **Alam Sutera** ☐ **Bandung** ☐ **Medan** ☐ **Bekasi** ☐ **Malang** ☐ **BiOn**	**Academic Year :** **2025 / 2026**

Exam Type*	:	~~Onsite / Online~~ / Take Home	Faculty / Dept.	:	SoCS / Mathematics
Day / Date**	:	Monday / 19 January 2026	Code - Course	:	MATH6154016 - Speech and Audio Processing
Time**	:	17.00 WIB			
Exam Specification***	:	☐ Open Book ☐ Open Notes ☐ Close Book ☐ Oral Test ☐ Open E-Book	Class	:	LA05
Equipment***	:		Student ID ***	:	
☐ Examination Booklet ☐ Calculator ☐ Dictionary		☐ Laptop ☐ Drawing Paper – A3 ☐ Tablet ☐ Drawing Paper – A2 ☐ Smartphone ☐ Notes: _____sheet	Name ***	:	
			Signature ***	:	

*) *Strikethrough the unnecessary items* **) *For Online Exam, this is the due date* ***) *Only for Onsite Exam*

Please insert the test paper into the examination booklet and submit both papers after the test.*

The penalty for CHEATING is DROP OUT!

Learning Outcome for

LO1: (C2) Comprehension: explain the fundamentals of digital signal processing for audio and speech
LO2: (C3) Application: apply audio and speech processing based on deep learning method to automatic speech recognition, text to speech, and speaker recognition.
LO3: (C3) Application: build audio and speech processing code in Python programming.

Catatan:

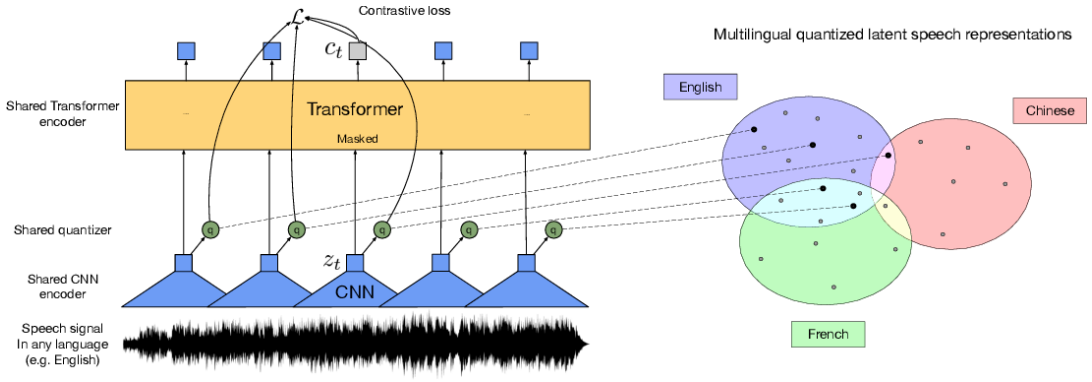
1. Jawablah soal ujian ini di dalam Python Notebook.
2. Gunakan Markdown untuk menjelaskan setiap jawaban yang diberikan. Jika diperlukan, Anda dapat menyisipkan gambar menggunakan menu Edit.
3. Anda diperbolehkan menggunakan pustaka audio yang dibutuhkan, seperti Librosa dan TorchAudio.
4. [PENTING] Kumpulkan jawaban ujian dalam format PDF dan IPYNB yang dikompres menjadi ZIP sebagai FINAL REPORT.
5. [Wajib] Buat video presentasi yang menjelaskan jawaban Anda langkah demi langkah. Unggah video tersebut ke YouTube, lalu sertakan URL-nya dalam jawaban Anda.

Verified by Department,

Faisal (D5556)
12 21, 2025

1. [Audio Feature Extraction, LO 1, LO 2, LO 3, 25%]

Wav2Vec 2.0 adalah metode self-supervised learning untuk mempelajari representasi (fitur) dari suara. Model ini dilatih dari rekaman suara tanpa label dalam jumlah besar, lalu bisa di-fine-tune untuk tugas tertentu, misalnya klasifikasi atau pengenalan ucapan. Pengembangan lanjutannya adalah XLSR-Wav2Vec2 (Cross-Lingual Speech Representation). Model ini dirancang agar representasi suara yang dipelajari bisa dipakai lintas banyak bahasa. Secara konsep, cara belajarnya mirip BERT: sebagian fitur suara "disamarkan" secara acak, lalu model belajar menebak informasi yang hilang menggunakan arsitektur Transformer.



a) Penjelasan Konsep Wav2Vec 2.0

Jelaskan secara rinci cara kerja Wav2Vec 2.0, minimal mencakup:

- Tahap ekstraksi fitur dari audio (feature encoder).
- Cara model membangun representasi konteks dengan Transformer.
- Konsep *masking* dan tujuan pelatihan *self-supervised*.
- Apa yang berubah saat model dipakai untuk tugas klasifikasi ucapan (fine-tuning untuk classification).
- Berikan contoh penggunaan Wav2Vec 2.0 untuk speech classification (misalnya klasifikasi emosi, aksen, atau jenis pembicara) dan jelaskan alur kerjanya.

b) Implementasi: Klasifikasi Multikelas Aksen Bicara

Bangun model Speech Accent Multiclass Classification berbasis Wav2Vec 2.0 dengan ketentuan berikut:

Kelas (3 kelas):

- us : United States English
- australia : Australian English
- england : England English

Dataset:

- Gunakan dataset Mozilla Common Voice yang tersedia di Kaggle. https://www.kaggle.com/datasets/mozillaorg/common-voice

Target kinerja:

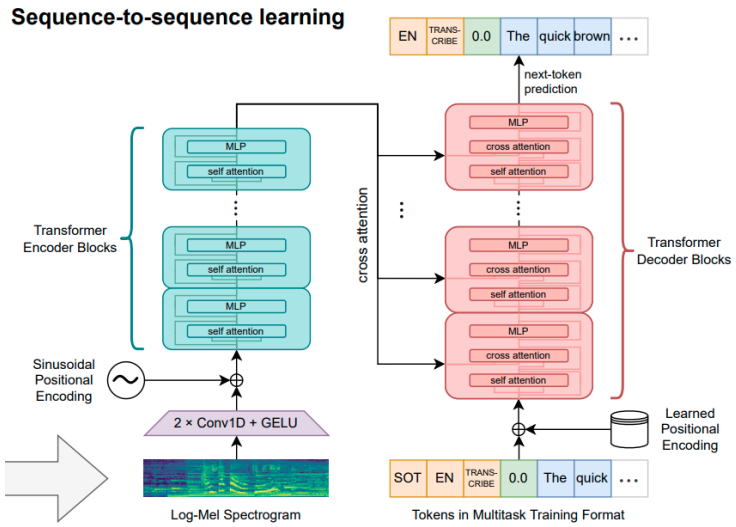
- Model yang dibuat harus mencapai minimal 90% akurasi pada data uji (test set).

2. [Automatic Speech Recognition, LO 1, LO 2, LO 3, 25%]

Salah satu *state of the art* untuk *Automatic Speech Recognition* (ASR) adalah algoritma **Whisper** dari Open-AI yang dijelaskan dalam makalah ilmiah berjudul *Robust Speech Recognition via Large-Scale Weak Supervision* (Radford et al, 2022).

Verified by Department,

Faisal (D5556)
12 21, 2025



a) Penjelasan Cara Kerja Whisper
Jelaskan secara rinci bagaimana Whisper bekerja, mencakup:

- Proses training (jenis data yang digunakan, konsep *weak supervision*, dan bagaimana model belajar dari data tersebut).
- Proses testing/inference (alur input audio sampai menjadi teks).
- Evaluation metrics yang umum dipakai untuk menilai ASR (misalnya WER/CER) dan cara membacanya.

b) Whisper dan CTC Loss
Apakah Whisper menggunakan CTC (Connectionist Temporal Classification) loss seperti pada DeepSpeech 2 (Amodei dkk., 2015)?
Jelaskan juga:

- Apa itu CTC loss dalam konteks ASR.
- Mengapa CTC penting (misalnya untuk menyelaraskan audio dan teks tanpa alignment per frame).
- Kelebihan dan keterbatasan pendekatan CTC dibanding pendekatan lain pada ASR.
- Perbandingan Whisper dengan algoritma ASR lain yang berbasis CTC, terutama dari sisi:
  - ketahanan terhadap noise/aksen,
  - kebutuhan data label,
  - kualitas hasil transkripsi,
  - dan kecepatan/biaya komputasi (secara umum).

c) Praktik: Whisper untuk Bahasa Indonesia dengan Dataset FLEURS
Berdasarkan repositori Whisper **https://github.com/openai/whisper**, lakukan demonstrasi penggunaan Whisper untuk bahasa Indonesia dengan langkah berikut:

- Gunakan notebook Multilingual_ASR.ipynb dari repositori tersebut.
- Modifikasi notebook agar memakai dataset FLEURS (Few-shot Learning Evaluation of Universal Representations of Speech) untuk bahasa Indonesia.
- Tunjukkan hasil minimal:
  - contoh beberapa audio dan hasil transkripsinya,
  - perhitungan metrik evaluasi (misalnya WER atau CER),
  - dan ringkasan hasil (berapa nilai metriknya serta interpretasinya).

3. [Speech To Text, LO 1, LO 2, LO 3, 25%]
Salah satu *state of the art* untuk *Speech to Text* (STT) adalah algoritma **FastSpeech 2** dari Microsoft yang dijelaskan dalam makalah ilmiah berjudul *FastSpeech 2: Fast and High-Quality End-to-End Text to Speech* (Ren et al, 2022).

Verified by Department, *Faisal (D5556)* *12 21, 2025*

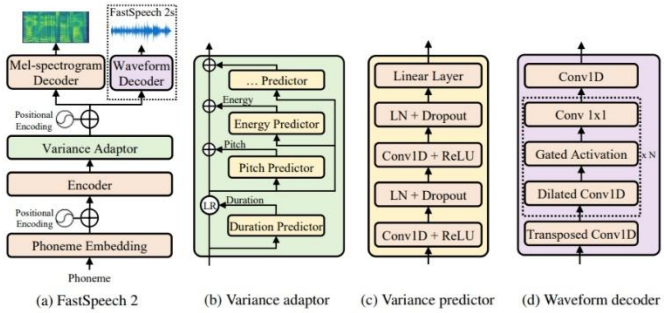


Figure 1: The overall architecture for FastSpeech 2 and 2s. LR in subfigure (b) denotes the length regulator operation proposed in FastSpeech. LN in subfigure (c) denotes layer normalization. Variance predictor represents duration/pitch/energy predictor.

a) Penjelasan Konsep FastSpeech 2
Jelaskan secara rinci cara kerja FastSpeech 2, meliputi:

- Proses training (data yang dibutuhkan, target yang dipelajari model, dan komponen penting yang dilatih).
- Proses testing (alur dari input teks sampai menjadi audio).
- Evaluation metrics yang umum dipakai untuk menilai kualitas TTS (misalnya naturalness, kualitas suara, kesesuaian durasi, dan kecepatan inferensi).
- Perbandingan FastSpeech 2 dengan metode TTS lain (contoh: Tacotron 2) dari sisi kualitas dan kecepatan.

b) Praktik: Generate Audio Bahasa Inggris dengan FastSpeech 2 (Coqui TTS)
Berdasarkan repositori Coqui TTS https://github.com/coqui-ai/TTS, lakukan demonstrasi penggunaan FastSpeech 2 dengan ketentuan:

- Gunakan model FastSpeech 2 yang tersedia di Coqui TTS.
- Buat atau modifikasi notebook agar dapat menjalankan TTS.
- Gunakan teks berikut sebagai input:

  "Text to Speech or TTS plays an important role in helping the economic sector to remind customers personally in an organized and effective manner."
- Simpan hasil audio dalam bentuk file .wav.

c) Praktik: Generate Audio Bahasa Indonesia dengan Model Pre-trained
Berdasarkan repositori model TTS Bahasa Indonesia https://github.com/Wikidepia/indonesian-tts, lakukan demonstrasi Text-to-Speech bahasa Indonesia dengan ketentuan:

- Gunakan model pre-trained Bahasa Indonesia dari repositori tersebut.
- Gunakan teks berikut sebagai input:

  "STT adalah teknologi yang memungkinkan kita untuk mengubah ucapan menjadi teks tertulis. Teknologi ini sangat berguna bagi mereka yang ingin mengetik secara lebih cepat dan efisien dengan suara."
- Simpan hasil audio dalam bentuk file .wav.

4. [Speaker Recognition, LO 1, LO 2, LO 3, 25%]

Salah satu *state of the art* untuk *Speaker Recognition* (SR) adalah algoritma **Deep Speaker** dari Baidu yang dijelaskan dalam makalah ilmiah berjudul *Deep Speaker: an End-to-End Neural Speaker Embedding System* (Li et al, 2017).

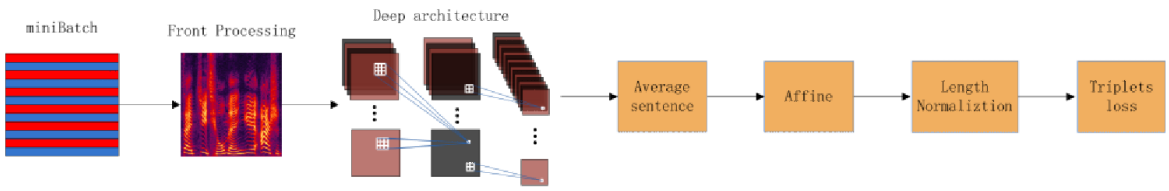


Figure 1: *Diagram of the Deep Speaker architecture*

Verified by Department,

Faisal (D5556)
12 21, 2025

a) Penjelasan Konsep Deep Speaker

Jelaskan cara kerja Deep Speaker secara rinci, mencakup:

- Proses training: bagaimana data suara dipakai untuk melatih model menghasilkan speaker embedding.
- Proses testing/inference: bagaimana model digunakan untuk membandingkan dua rekaman suara (apakah dari orang yang sama atau berbeda).
- Evaluation metrics yang umum dipakai pada speaker recognition dan arti dari metrik tersebut.
- Perbandingan Deep Speaker dengan metode lain dari sisi konsep dan performa secara umum.

b) Triplet Loss dan Hubungannya dengan Metode Similarity

Dalam Deep Speaker digunakan Triplet Loss. Jelaskan secara rinci:

- Konsep Triplet Loss (anchor, positive, negative) dan tujuan utamanya.
- Hubungan Triplet Loss dengan Contrastive Loss.
- Hubungan Triplet Loss dengan Siamese Network
- Perbedaan utama pendekatan berbasis similarity ini dibanding pendekatan klasifikasi speaker.

c) Praktik: Demo Deep Speaker dan Perbandingan Cosine Similarity

Berdasarkan repositori https://github.com/philipperemy/deep-speaker, lakukan demonstrasi penggunaan Deep Speaker dengan ketentuan:

- Gunakan model “ResCNN Softmax+Triplet trained” yang sudah dilatih menggunakan dataset LibriSpeech.
- Lakukan pengujian speaker recognition dengan cara:
    - Ambil beberapa contoh pasangan audio speaker sama dan speaker berbeda.
    - Ekstrak embedding dari masing-masing audio.
    - Hitung kemiripan menggunakan cosine similarity.
- Tampilkan hasil perbandingan secara jelas, minimal berisi:
    - pasangan audio yang diuji,
    - nilai cosine similarity,
    - kesimpulan: “speaker sama” atau “speaker berbeda” berdasarkan nilai similarity (gunakan ambang batas yang Anda tentukan, dan jelaskan alasannya).

-- Selamat mengerjakan--

Verified by Department,

Faisal (D5556)
12 21, 2025