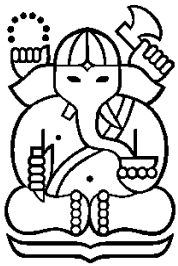
# DIARISASI PEMBICARA YANG TUMPANG TINDIH PADA PERCAKAPAN MULTI PEMBICARA BERBASIS DEEP LEARNING

**RINGKASAN DISERTASI**

Agus Komarudin
NIM: 33220005
(Program Studi Doktor Teknik Elektro dan Informatika)



Institut Teknologi Bandung
Juli 2026

# DIARISASI PEMBICARA YANG TUMPANG TINDIH PADA PERCAKAPAN MULTI PEMBICARA BERBASIS DEEP LEARNING

Disertasi ini dipertahankan pada Sidang Terbuka Sekolah Pascasarjana sebagai salah satu syarat untuk memperoleh gelar Doktor Institut Teknologi Bandung

Juli 2026

Agus Komarudin
NIM: 33220005
(Program Studi Doktor Teknik Elektro dan Informatika)

Promotor : Prof. Dr. Ir. Rinaldi Munir, M.T.
Ko-promotor : Dessi Puji Lestari, S.T., M.Eng., Ph.D
Achmad Imam K., S.T., M.Sc., Ph.D

Institut Teknologi Bandung
Juli 2026

# DIARISASI PEMBICARA YANG TUMPANG TINDIH PADA PERCAKAPAN MULTI PEMBICARA BERBASIS DEEP LEARNING

Agus Komarudin
NIM: 33220005

## 1. Latar Belakang

*Speaker diarization* adalah teknologi untuk menjawab pertanyaan "who spoke when" di dalam sebuah rekaman (Miro dkk., 2012; Park dkk., 2022). Teknologi ini membagi rekaman menjadi garis waktu aktivitas pembicara sehingga dapat diketahui kapan setiap orang mulai dan berhenti berbicara. Hasilnya dibutuhkan dalam transkripsi rapat otomatis, konferensi daring, analisis percakapan pusat layanan, dokumentasi wawancara, forensik suara, dan berbagai aplikasi yang melibatkan percakapan banyak orang.

Bagi manusia, mengikuti percakapan merupakan kegiatan yang tampak sederhana karena pendengar dapat mengenali ciri suara, arah datang suara, konteks kalimat, dan kebiasaan pergantian giliran. Namun bagi komputer, tugas tersebut jauh lebih rumit. Sistem harus menentukan bagian yang berisi ucapan, memisahkan aktivitas setiap pembicara, dan menjaga konsistensi identitas sepanjang rekaman, termasuk ketika pembicara berbicara bersamaan (Miro dkk., 2012; Park dkk., 2022).

Kondisi ketika dua orang atau lebih aktif pada waktu yang sama disebut ucapan tumpang tindih atau *overlapping speech*. Fenomena ini lazim terjadi dalam rapat dan diskusi: seseorang dapat menyela, memberi respons singkat, tertawa, atau mulai berbicara sebelum pembicara sebelumnya selesai. Pada kondisi tersebut, asumsi lama bahwa satu bagian rekaman hanya dimiliki satu pembicara tidak lagi berlaku). Jika sistem hanya memilih satu identitas, aktivitas pembicara lain akan hilang dan kesalahan diarisasi meningkat (Bullock dkk., 2020; Raj, 2021).

Pendekatan konvensional biasanya terdiri atas beberapa komponen terpisah, yaitu deteksi aktivitas suara, segmentasi, ekstraksi ciri pembicara, dan pengelompokan atau clustering. Rangkaian ini mudah dipahami, tetapi kesalahan pada satu tahap dapat merambat ke tahap berikutnya. Batas segmen yang kurang tepat akan menghasilkan ciri suara yang tercampur, lalu *clustering* dapat memberikan identitas yang salah. Kelemahan tersebut semakin nyata pada wilayah *overlap* (Miro dkk., 2012; Park dkk., 2022).

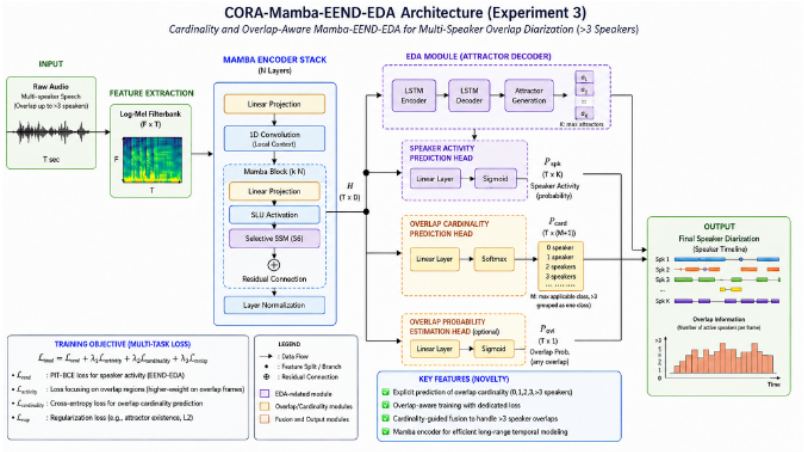
*End-to-End Neural Diarization* (EEND) dikembangkan untuk memprediksi aktivitas beberapa pembicara secara langsung dari fitur akustik. Model

menghasilkan probabilitas aktif atau tidak aktif untuk setiap pembicara pada setiap frame audio, sehingga beberapa pembicara dapat dinyatakan aktif secara bersamaan (Fujita dkk., 2019a, 2019b). *Encoder-Decoder Attractor* berbasis *EEND* (EEND-EDA) kemudian menambahkan *attractor*, yaitu representasi ringkas yang bertindak sebagai pusat identitas setiap pembicara. Mekanisme ini membuat jumlah pembicara lebih fleksibel daripada *EEND* awal (Horiguchi dkk., 2020, 2021).

Meskipun *EEND-EDA* memberikan kemajuan penting, *encoder* berbasis *self-attention* memiliki biaya komputasi yang meningkat secara kuadratik terhadap panjang urutan. Rekaman audio dapat menghasilkan ribuan hingga puluhan ribu *frame*, sehingga waktu dan memori menjadi kendala (Vaswani dkk., 2017; Gu dan Dao, 2024). Selain itu, informasi mengenai jumlah pembicara aktif dan keberadaan *overlap* masih dipelajari secara tidak langsung melalui keluaran diarisasi (Takashima dkk., 2021; Singh dkk., 2025). Pada *overlap* kompleks tiga atau empat pembicara, sinyal pembelajaran tersebut belum selalu cukup untuk membentuk representasi yang tegas.

Disertasi ini menjawab dua kebutuhan yang saling berkaitan. Pertama, dibutuhkan *encoder* yang efisien untuk memodelkan hubungan temporal pada rekaman panjang. Kedua, dibutuhkan informasi struktural yang membantu model memahami apakah suatu *frame* merupakan diam, pembicara tunggal, atau *overlap* dengan beberapa pembicara. *Mamba*, sebuah *Selective State Space Model*, dipilih karena memproses urutan dengan kompleksitas yang meningkat secara linear (Gu dan Dao, 2024). Selanjutnya, informasi *cardinality* dan *overlap* ditambahkan sebagai tugas pembelajaran tambahan (Takashima dkk., 2021; Singh dkk., 2025).

Penelitian dilaksanakan melalui tiga tahap berkelanjutan. Tahap pertama memperbaiki batas aktivitas pada *overlap* dua pembicara melalui *overlap-aware resegmentation*. Tahap kedua menghasilkan *Mamba-EEND-EDA* dengan Bidirectional Mamba sebagai encoder dan EDA sebagai pembentuk attractor. Tahap ketiga menghasilkan *CORA-Mamba-EEND-EDA* dengan *cardinality-aware learning* dan *overlap-aware learning* untuk percakapan empat pembicara. Dengan tahapan tersebut, penelitian bergerak dari perbaikan segmentasi, menuju pemodelan urutan yang efisien, lalu menuju pemahaman struktur overlap yang lebih rinci.




Gambar 1. Peta tiga tahap penelitian disertasi: *overlap-aware resegmentation*, *Mamba-EEND-EDA*, dan *CORA-Mamba-EEND-EDA*

## 2. Tujuan dan Sasaran Penelitian

Tujuan umum penelitian adalah mengembangkan sistem *speaker diarization end-to-end* yang lebih akurat dan efisien untuk percakapan multipembicara, serta lebih peka terhadap perubahan jumlah pembicara aktif dan kondisi *overlap* yang kompleks. Tujuan tersebut dijabarkan menjadi sasaran penelitian berikut.

1. Mengembangkan *Mamba-EEND-EDA* dengan mengganti *encoder* berbasis *self-attention* pada *EEND-EDA* menggunakan *Bidirectional Mamba*, sementara *EDA* tetap dipertahankan untuk membentuk *speaker attractor*.
2. Menghasilkan representasi temporal yang mendukung prediksi aktivitas beberapa pembicara secara bersamaan pada percakapan empat pembicara, dengan jumlah pembicara aktif yang berubah pada setiap *frame*.
3. Menerapkan *cardinality*, yaitu jumlah pembicara aktif pada setiap *frame*, sebagai sinyal pembelajaran tambahan agar model dapat membedakan diam, pembicara tunggal, dan *overlap* multipembicara.
4. Menerapkan informasi *overlap* sebagai sinyal pembelajaran tambahan untuk meningkatkan kepekaan model terhadap *frame* yang mengandung lebih dari satu pembicara aktif.
5. Menganalisis kontribusi *Baseline Mamba-EEND-EDA*, *Cardinality Only*, *Overlap Only*, dan *Full CORA* terhadap *Diarization Error Rate* serta komponen *Speaker Miss*, *False Alarm*, dan *Speaker Confusion*.
6. Menganalisis peningkatan efisiensi dari penggunaan *Mamba* melalui waktu inferensi, *Real-Time Factor*, penggunaan memori *GPU*, dan *throughput*.

Sasaran akhir penelitian bukan hanya memperoleh nilai kesalahan yang lebih rendah, tetapi juga menghasilkan pemahaman ilmiah tentang jenis informasi tambahan yang paling berguna untuk membimbing representasi temporal. Dengan demikian, hasil penelitian diharapkan bermanfaat bagi pengembangan sistem transkripsi percakapan, dokumentasi rapat, analisis interaksi, dan aplikasi suara lain yang membutuhkan pemisahan aktivitas banyak pembicara.

## 3. Metode Penelitian

Penelitian menggunakan pendekatan eksperimen kuantitatif yang dilaksanakan dalam tiga tahap. Setiap tahap menjawab persoalan pada tingkat kompleksitas yang lebih tinggi dan hasilnya menjadi dasar bagi tahap berikutnya. Seluruh model dinilai dengan metrik diarisasi yang sama agar perubahan kinerja dapat dibandingkan secara objektif.

### 3.1 Tahap Penelitian

Tahapan penelitian disusun sebagai berikut:

1. Tahap pertama (SK1) menggunakan *overlap-aware resegmentation* pada data *AMI-Headset* untuk memperbaiki batas aktivitas pembicara dalam percakapan dua pembicara. Tahap ini membangun dasar bahwa wilayah *overlap* perlu ditangani secara eksplisit.
2. Tahap kedua (SK2) mengembangkan *Mamba-EEND-EDA*. *Encoder* berbasis perhatian pada *EEND-EDA* diganti dengan *Bidirectional Mamba*, sedangkan *EDA* dipertahankan untuk membentuk *attractor* pembicara. Data campuran dibentuk dari *LibriSpeech* dan diperkaya dengan derau *MUSAN* serta *Room Impulse Response*.
3. Tahap ketiga (SK3) mengembangkan *CORA-Mamba-EEND-EDA* untuk *overlap kompleks* empat pembicara. Dua jalur tambahan, *cardinality head* dan *overlap head*, digunakan selama pelatihan untuk memberi modul tambahan kepada *encoder*.

### 3.2 Pembentukan Data dan Representasi Audio

Pada tahap akhir, eksperimen menggunakan 5.728 *dataset* audio campuran simulasi empat pembicara untuk pelatihan dan 500 rekaman untuk evaluasi. Penggunaan campuran simulasi memungkinkan jumlah pembicara, pola aktivitas, dan *overlap* dikendalikan sehingga perbandingan antar konfigurasi berlangsung pada kondisi yang sama.

*Audio mono* 16 kHz dibagi menjadi beberapa *frame* pendek. Dari setiap *frame* dihitung 23 fitur *log-Mel filterbank* yang menggambarkan distribusi energi pada skala frekuensi yang mendekati persepsi pendengaran manusia. Untuk menyediakan konteks lokal, tujuh *frame* sebelum, *frame* saat ini, dan tujuh *frame* sesudah

ditumpuk menjadi 15 frame konteks. Hasilnya adalah vektor 23 × 15 atau 345 dimensi.
Urutan kemudian mengalami subsampling faktor 10 untuk mengurangi panjang komputasi. Vektor 345 dimensi diproyeksikan menjadi 256 dimensi sebelum memasuki *encoder*. Pengaturan tersebut mempertahankan informasi spektral dan konteks temporal, tetapi mengurangi jumlah langkah yang harus diproses model.

### 3.3 Arsitektur *Mamba-EEND-EDA*

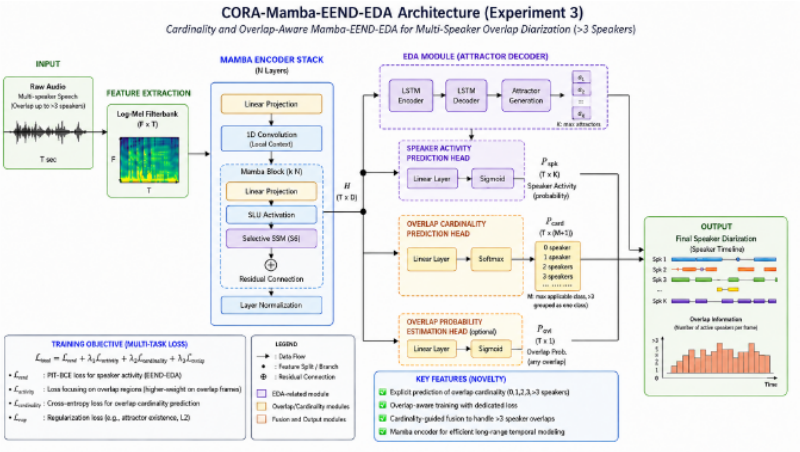
*Bidirectional Mamba* terdiri atas jalur maju dan jalur mundur. Jalur maju membaca rekaman dari awal ke akhir, sedangkan jalur mundur membaca urutan yang dibalik. Kedua representasi digabungkan sehingga setiap frame memanfaatkan informasi sebelum dan sesudahnya. Pendekatan dua arah sesuai untuk sistem *offline* karena keseluruhan rekaman telah tersedia saat inferensi.

Di dalam blok *Mamba* terdapat proyeksi internal, konvolusi satu dimensi untuk konteks lokal, *Selective State Space Model* untuk membawa informasi sepanjang urutan, mekanisme *gating*, *residual connection*, dan normalisasi. Berbeda dari *self-attention* yang membandingkan semua pasangan posisi, *Mamba* memperbarui keadaan secara berurutan sehingga kompleksitasnya meningkat secara linear terhadap panjang rekaman.

Keluaran *encoder* diteruskan ke *Encoder-Decoder Attractor*. *LSTM encoder* merangkum urutan *embedding*, sedangkan *LSTM decoder* membentuk *attractor* satu persatu. Setiap *attractor* mewakili karakteristik seorang pembicara. Probabilitas aktivitas diperoleh dari hubungan antara *embedding frame* dan *attractor*, lalu *sigmoid* mengubahnya menjadi nilai antara nol dan satu.

### 3.4 Arsitektur CORA-Mamba-EEND-EDA

*CORA* merupakan singkatan dari *Cardinality-Aware dan Overlap-Aware*. Arsitektur ini mempertahankan keluaran utama *speaker activity* dan menambahkan dua tugas tambahan selama pelatihan. Kedua tugas tidak menggantikan keputusan diarisasi, tetapi memberikan gradien tambahan kepada *encoder* bersama.




Gambar 2. Arsitektur *CORA-Mamba-EEND-EDA*: ekstraksi fitur, *Bidirectional Mamba*, *EDA*, *speaker activity head*, *cardinality head*, dan *overlap head*

*Cardinality head* memprediksi jumlah pembicara aktif pada setiap *frame* dalam lima kelas: 0, 1, 2, 3, atau 4. Kelas nol merepresentasikan diam, kelas satu pembicara tunggal, sedangkan kelas dua sampai empat menunjukkan *overlap* dengan tingkat yang berbeda. Karena memuat jumlah yang tepat, *cardinality* memberikan informasi lebih rinci daripada keputusan *overlap biner*.

*Overlap head* menghasilkan satu probabilitas per *frame*. Target bernilai nol ketika tidak lebih dari satu pembicara aktif dan bernilai satu ketika dua atau lebih pembicara aktif. Tugas ini memberi sinyal langsung mengenai batas antara kondisi tunggal dan tumpang tindih.

Fungsi objektif menggabungkan *diarization loss, attractor loss, cardinality loss,* dan *overlap loss*. *Permutation Invariant Training* digunakan pada *diarization loss* agar kesalahan tidak bergantung pada urutan label pembicara. Empat skenario diuji: *Baseline* tanpa tugas tambahan, *Cardinality Only, Overlap Only*, dan *Full CORA* dengan kedua tugas.

### 3.5 Konfigurasi Eksperimen dan Evaluasi

Metrik utama adalah *Diarization Error Rate* (DER). DER merupakan jumlah tiga jenis kesalahan yang dinormalisasi terhadap total durasi ucapan referensi. *Speaker Miss* adalah ucapan yang seharusnya ada tetapi tidak terdeteksi. *False Alarm* adalah aktivitas yang diprediksi ketika referensi tidak menunjukkan aktivitas. *Speaker Confusion* adalah durasi ketika ucapan terdeteksi tetapi diberikan kepada identitas pembicara yang salah. Analisis ketiga komponen penting karena dua model dengan *DER* hampir sama dapat memiliki perilaku kesalahan yang berbeda.

**Tabel 1. Konfigurasi utama eksperimen tahap akhir**

| Komponen | Konfigurasi |
|---|---|
| **Data pelatihan** | 5.728 campuran simulasi empat pembicara |
| **Data evaluasi** | 500 rekaman |
| **Audio** | Mono, 16 kHz |
| **Fitur** | 23 log-Mel × 15 frame konteks = 345 dimensi |
| **Subsampling** | Faktor 10 |
| **Encoder** | Bidirectional Mamba, 2 lapisan, dimensi 256 |
| **EDA** | LSTM encoder-decoder attractor, dimensi 256 |
| **Cardinality** | 5 kelas: 0–4 pembicara aktif |
| **Overlap** | Keluaran biner non-overlap/overlap |

## 4. Hasil dan Pembahasan

Hasil penelitian dibahas sesuai tiga tahap agar terlihat hubungan antara perbaikan batas aktivitas, efisiensi pemodelan urutan, dan modul tambahan pada *overlap* kompleks. Nilai kuantitatif utama tersedia pada SK2 dan SK3, sedangkan SK1 berfungsi sebagai dasar konseptual dan metodologis penanganan *overlap*.

### 4.1 Hasil Tahap Pertama: Overlap-Aware Resegmentation

Tahap pertama menunjukkan bahwa kesalahan diarisasi pada wilayah *overlap* tidak cukup ditangani hanya dengan *clustering*. Segmentasi awal dapat menghilangkan pembicara kedua atau menempatkan batas pergantian pembicara secara kurang tepat. Dengan menambahkan informasi *overlap* pada *resegmentation*, batas aktivitas dapat diperbaiki dan sistem memperoleh representasi garis waktu yang lebih sesuai dengan percakapan sebenarnya.

Temuan tahap pertama memberikan landasan bagi tahap berikutnya: *overlap* harus dipandang sebagai kondisi multilabel, bukan sebagai gangguan yang dibuang. Namun, pendekatan resegmentation masih bergantung pada rangkaian komponen konvensional dan belum menyelesaikan persoalan pemodelan rekaman panjang secara *end-to-end*.

### 4.2 Hasil Tahap Kedua: *Mamba-EEND-EDA*

Penggantian *encoder* berbasis *self-attention* dengan *Bidirectional Mamba* meningkatkan akurasi dan efisiensi. *DER baseline EEND-EDA* sebesar 18,5% turun menjadi 14,7%, atau membaik relatif sekitar 20,5%. *False Alarm* turun dari 4,3% menjadi 2,9%, *Speaker Miss* dari 8,4% menjadi 6,5%, dan *Speaker Confusion* dari 5,8% menjadi 5,3%.

Peningkatan efisiensi lebih menonjol. Waktu inferensi turun dari 2.340 ms menjadi 520 ms atau berkurang 77,8%. *Real-Time Factor* turun dari 0,47 menjadi 0,10. Penggunaan memori GPU berkurang dari 1.240 MB menjadi 680 MB atau hemat

45,2%, sedangkan *throughput* meningkat dari 685 menjadi 2.910 sampel per detik atau naik 324,8%.

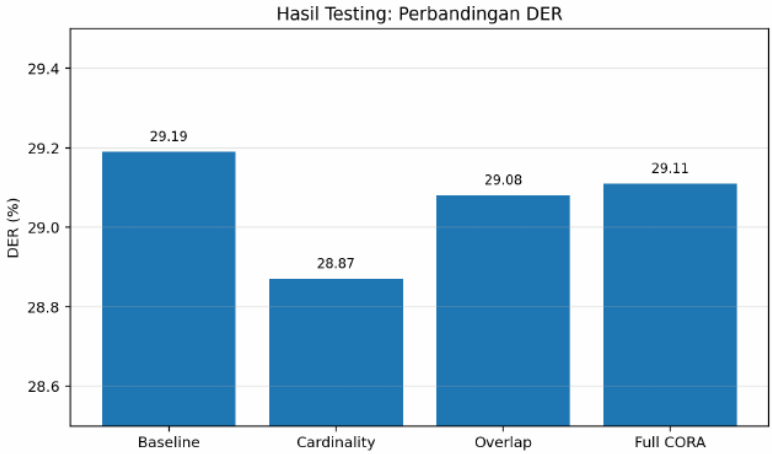
Hasil tersebut mendukung hipotesis bahwa *Mamba* sesuai untuk urutan audio panjang. Model mempertahankan mekanisme *attractor EDA*, tetapi menghindari biaya perbandingan semua pasangan *frame* pada *self-attention*. Dengan demikian, *Mamba-EEND-EDA* menjadi fondasi yang lebih efisien bagi eksperimen empat pembicara.

### 4.3 Hasil Tahap Ketiga: *CORA-Mamba-EEND-EDA*

Empat konfigurasi dibandingkan pada 500 rekaman empat pembicara. Seluruh konfigurasi dengan modul tambahan menghasilkan DER yang lebih rendah daripada *baseline*, tetapi besar perbaikannya berbeda.

**Tabel 2. Hasil pengujian akhir pada percakapan empat pembicara**

| Konfigurasi | DER (%) | Perubahan dari baseline |
|---|---|---|
| **Baseline Mamba-EEND-EDA** | 29,19 | – |
| **Cardinality Only** | 28,87 | −0,32 poin |
| **Overlap Only** | 29,08 | −0,11 poin |
| **Full CORA** | 29,11 | −0,08 poin |




Gambar 3. Perbandingan DER pada pengujian akhir; Cardinality Only memberikan nilai terendah.

### 4.4 Analisis Cardinality Only

*Cardinality Only* menghasilkan DER 28,87%, turun 0,32 poin persentase dari *baseline*. Nilai ini merupakan hasil terbaik pada eksperimen SK3. Informasi jumlah pembicara aktif membantu encoder membedakan tiga struktur dasar percakapan: diam, pembicara tunggal, dan overlap multipembicara. Lebih jauh, kelas dua, tiga,

dan empat memberikan tingkat kepadatan aktivitas yang tidak tersedia pada target overlap biner.

Pada percakapan empat pembicara, persoalan utama bukan hanya mengetahui bahwa overlap terjadi, tetapi juga memahami berapa banyak aktivitas yang harus dipertahankan. *Cardinality* memberikan batas struktural terhadap banyaknya pembicara aktif tanpa memaksa identitas tertentu. Oleh karena itu, sinyal ini dapat memperbaiki organisasi *embedding* temporal dan membantu keluaran aktivitas utama.

### 4.5 Analisis Overlap Only

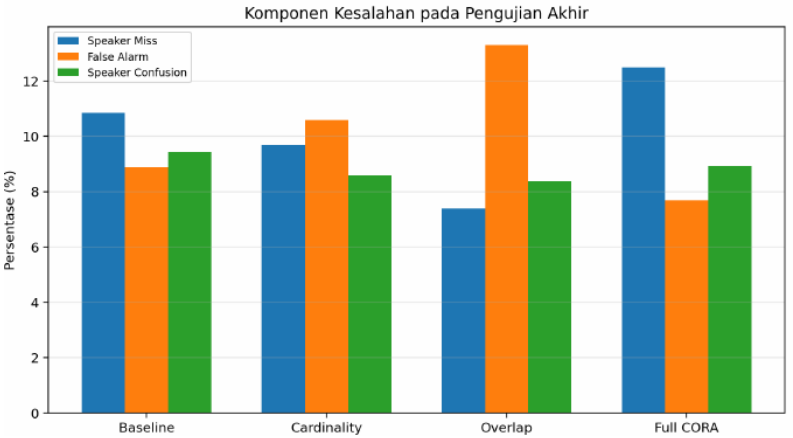
*Overlap Only* menghasilkan *DER* 29,08%. Tugas ini meningkatkan sensitivitas terhadap wilayah tumpang tindih dan menghasilkan *Speaker Miss* terendah di antara konfigurasi SK3. Artinya, lebih banyak aktivitas referensi berhasil dipertahankan. Namun, sensitivitas yang lebih tinggi membuat prediksi aktivitas cenderung lebih agresif. *False Alarm* meningkat menjadi 13,30%, sehingga penurunan *miss* tidak sepenuhnya berubah menjadi penurunan *DER*. Hasil ini menggambarkan pertukaran antara sensitivitas dan ketelitian: sistem yang terlalu berhati-hati dapat kehilangan ucapan, sedangkan sistem yang terlalu aktif dapat menambahkan ucapan yang tidak.

### 4.6 Analisis Full CORA

*Full CORA* menghasilkan *DER* 29,11%. Nilai tersebut masih lebih baik daripada baseline, tetapi belum melampaui *Cardinality Only*. *Full CORA* menghasilkan *False Alarm* terendah sebesar 7,70%, namun *Speaker Miss* meningkat menjadi 12,49%. Dengan demikian, kombinasi dua supervisi mengubah keseimbangan kesalahan, tetapi belum memberikan keuntungan kumulatif.

*Cardinality* dan *overlap* memiliki hubungan logis yang sangat dekat. *Overlap* dapat diturunkan dari *cardinality* karena *frame* dianggap *overlap* ketika *cardinality* lebih dari satu. Ketika dua *loss* yang berkorelasi dioptimalkan bersamaan, gradiennya dapat saling memperkuat pada sebagian contoh tetapi bersaing pada contoh lain. Hasil juga dipengaruhi distribusi kelas, bobot *loss*, *threshold* keputusan, dan potensi konflik gradien.

Temuan ini penting karena menunjukkan bahwa menambahkan lebih banyak tugas bukan jaminan menghasilkan model yang lebih baik. Pembelajaran multitugas memerlukan desain interaksi antartugas, bukan sekadar penjumlahan fungsi *loss*.




Gambar 4. Perbandingan komponen kesalahan yang menunjukkan pertukaran antara *Speaker Miss* dan *False Alarm*

### 4.7 Makna Hasil bagi Pembaca Nonteknis

Secara sederhana, sistem harus membuat dua keputusan yang sama-sama sulit: jangan kehilangan suara yang benar-benar ada dan jangan menciptakan aktivitas yang sebenarnya tidak ada. *Cardinality* membantu model memperkirakan "berapa orang yang seharusnya terdengar", sedangkan *overlap* membantu menjawab "apakah lebih dari satu orang sedang berbicara". Eksperimen menunjukkan bahwa pertanyaan pertama memberikan informasi yang lebih kaya.

Penurunan *DER* pada SK3 memang relatif kecil dibandingkan peningkatan besar pada SK2, tetapi temuan ilmiahnya tetap penting. Eksperimen ablasi terkontrol membuktikan bahwa informasi struktural yang berbeda menghasilkan pola kesalahan berbeda. Pengetahuan tersebut memberikan arah yang lebih jelas untuk pengembangan selanjutnya, terutama penggunaan *cardinality* dan *overlap* pada tahap inferensi, bukan hanya selama pelatihan.

## 5. Kesimpulan atau Kontribusi Ilmiah

### 5.1 Kesimpulan

1. *Mamba-EEND-EDA* berhasil dikembangkan dengan mengganti *encoder* berbasis *self-attention* menggunakan *Bidirectional Mamba* dan mempertahankan *EDA* sebagai pembentuk *speaker attractor*. Integrasi tersebut menurunkan *DER* dari 18,5% menjadi 14,7% serta memperbaiki seluruh komponen kesalahan.

2. *Bidirectional Mamba* meningkatkan efisiensi secara nyata: waktu inferensi turun 77,8%, RTF menjadi 0,10, penggunaan memori GPU berkurang 45,2%, dan *throughput* meningkat 324,8%.
3. Kombinasi *Bidirectional Mamba* dan *EDA* dapat membangun representasi *frame* serta *attractor* yang mendukung prediksi aktivitas beberapa pembicara secara bersamaan.
4. *Cardinality* dapat digunakan sebagai sinyal pembelajaran tambahan. *Cardinality Only* menghasilkan DER 28,87%, lebih rendah daripada *baseline* 29,19%, dan menjadi konfigurasi terbaik pada eksperimen empat pembicara.
5. *Overlap* dapat digunakan sebagai sinyal pembelajaran tambahan untuk meningkatkan sensitivitas terhadap ucapan tumpang tindih, tetapi peningkatan sensitivitas dapat menaikkan *False Alarm*.
6. *Full CORA* menghasilkan DER 29,11% dan belum melampaui *Cardinality Only*. Kombinasi dua supervisi belum memberi keuntungan kumulatif karena keseimbangan antara *Speaker Miss* dan *False Alarm* belum optimal.

## 5.2 Kontribusi Ilmiah

Kontribusi ilmiah disertasi ini tidak hanya berupa penyajian hasil eksperimen, tetapi berupa pengembangan arsitektur dan pembuktian konsep dengan teknik baru. Kontribusi tersebut adalah sebagai berikut.

1. Menghasilkan kerangka bertahap penanganan *overlap* yang menghubungkan overlap-aware resegmentation, pemodelan urutan berbasis Mamba, dan modul tambahan per *frame* dalam satu garis penelitian yang saling berkelanjutan.
2. Menghasilkan arsitektur *Mamba-EEND-EDA* yang mempertahankan *speaker attractor EEND-EDA* tetapi mengganti *encoder* perhatian dengan *Bidirectional Mamba*. Hasil eksperimen membuktikan bahwa *state-space model* dapat meningkatkan akurasi sekaligus efisiensi untuk diarisasi pembicara.
3. Mengusulkan *CORA-Mamba-EEND-EDA*, yaitu integrasi *Bidirectional Mamba*, *EDA attractor*, *cardinality-aware learning*, dan *overlap-aware learning* dalam satu kerangka multitugas untuk *overlap* kompleks empat pembicara.
4. Membuktikan secara empiris bahwa *cardinality* merupakan sinyal struktural yang lebih informatif daripada keputusan *overlap biner* dalam membimbing representasi temporal, karena *cardinality* membedakan diam, pembicara tunggal, serta *overlap* dua, tiga, dan empat pembicara.
5. Membuktikan bahwa manfaat pembelajaran multitugas tidak selalu bersifat kumulatif. Interaksi antara *cardinality loss* dan *overlap loss* dipengaruhi oleh bobot l*oss*, ketidakseimbangan kelas, *threshold*, serta potensi konflik gradien.
6. Memberikan kerangka analisis berdasarkan komponen *DER* yang memperlihatkan pertukaran antara sensitivitas dan ketelitian. Penurunan *Speaker Miss* dapat disertai kenaikan *False Alarm*, sehingga evaluasi model perlu mempertimbangkan profil kesalahan, bukan hanya *DER* total.

## 6. Tindak Lanjut

Kontribusi ilmiah penelitian membuka beberapa cakrawala pengembangan baru. Arah tersebut tidak hanya berhubungan dengan *speaker diarization*, tetapi juga dapat mendukung transkripsi otomatis, analisis interaksi manusia, sistem rapat cerdas, dan pemrosesan audio multimodal.

1. Mengoptimalkan bobot *cardinality loss* dan *overlap loss* menggunakan *dynamic loss weighting* atau metode pengendalian konflik gradien agar kedua tugas memberikan kontribusi yang seimbang.
2. Memanfaatkan keluaran *cardinality* dan *overlap* pada tahap inferensi untuk mengoreksi jumlah pembicara aktif, mengkalibrasi *threshold*, atau membatasi keputusan aktivitas pada setiap frame.
3. Menganalisis kinerja secara terpisah pada *overlap* dua, tiga, dan empat pembicara agar diketahui kondisi yang paling diuntungkan atau paling sulit bagi setiap konfigurasi.
4. Menguji generalisasi pada percakapan nyata dengan bahasa, ruangan, derau, perangkat rekam, dan pola pergantian pembicara yang lebih beragam.
5. Memperluas model untuk jumlah pembicara lebih dari empat dan menguji estimasi jumlah pembicara otomatis tanpa *oracle speaker count*.
6. Mengembangkan versi *streaming* atau *real-time* menggunakan *Mamba* kausal, pemrosesan berbasis *chunk*, dan strategi pemeliharaan identitas pembicara *antar-chunk*.
7. Mengintegrasikan informasi visual, seperti gerak bibir atau arah pandang, untuk membantu pemisahan pembicara pada *overlap* yang sangat kompleks.
8. Menerapkan hasil diarisasi pada transkripsi rapat otomatis dan analisis percakapan sehingga kesalahan diarisasi dapat dinilai berdasarkan dampaknya terhadap aplikasi akhir.

## Riwayat Hidup

Agus Komarudin adalah mahasiswa Program Studi Doktor Teknik Elektro dan Informatika, Institut Teknologi Bandung, dengan Nomor Induk Mahasiswa 33220005. Bidang penelitian doktornya adalah pemrosesan ucapan dan kecerdasan artifisial, khususnya speaker diarization pada percakapan multipembicara yang mengandung ucapan tumpang tindih.

Biodata lengkap, tempat dan tanggal lahir, riwayat pendidikan S1 dan S2, serta riwayat pekerjaan perlu dilengkapi sesuai dokumen resmi promovendus. Bagian ini dibatasi maksimal satu halaman sesuai ketentuan Ringkasan Disertasi.

**Biodata dan riwayat singkat**

| | |
|---|---|
| **Nama** | Agus Komarudin |
| **NIM** | 33220005 |
| **Tempat/Tanggal Lahir** | Bandung, 14 Juni 1978 |

| | |
|---|---|
| **Pendidikan S1** | Teknik Informatika, STMIK Mardira Indonesia, 2002 |
| **Pendidikan S2** | Teknik Elektro, Institut Teknologi Bandung, 2010 |
| **Pendidikan S3** | Program Doktor Teknik Elektro dan Informatika, Institut Teknologi Bandung |
| **Riwayat Pekerjaan** | Dosen Tetap di Fakultas Sains dan Informatika (FSI) Universitas Jenderal Achmad Yani (UNJANI) dari tahun 2010 hingga sekarang |
| **Bidang Keahlian** | Informatika, kecerdasan artifisial, pemrosesan ucapan, speaker diarization |

**Daftar Publikasi Terkait Penelitian**

Selanjutnya, terdapat juga Daftar Publikasi Terkait Penelitian yang berisi daftar publikasi yang telah dilakukan selama menjalani program doktor.

| No | Tahun | Judul | Jenis | Penerbit |
|---|---|---|---|---|
| 1 | 2026 | Application of the Mamba-Encoder Model to EEND-EDA Speaker Diarization | Jurnal Internasional (Q4) | Taiwan Ubiquitous Information CO LTD |

Komarudin, A., Munir, R., Lestari, D.P., dan Kistijantoro, A.I. (2026): Application of the Mamba-Encoder Model to EEND-EDA Speaker Diarization, *Journal of Information Hiding and Multimedia Signal Processing*.

**Ucapan Terima Kasih**

Promovendus menyampaikan puji dan syukur kepada Allah SWT atas rahmat dan karunia-Nya sehingga penelitian dan penyusunan disertasi ini dapat diselesaikan.

Ucapan terima kasih disampaikan kepada Prof. Dr. Ir. Rinaldi Munir, M.T. selaku Promotor, Dessi Puji Lestari, S.T., M.Eng., Ph.D. selaku Co-Promotor, dan Achmad Imam Kistijantoro, S.T., M.Sc., Ph.D. selaku Co-Promotor atas bimbingan, arahan, kritik, dan dukungan selama pelaksanaan penelitian doktor.

Terima kasih juga disampaikan kepada Institut Teknologi Bandung, Program Studi Doktor Teknik Elektro dan Informatika, institusi tempat promovendus bekerja, para dosen, peneliti, rekan sejawat, serta semua pihak yang telah menyediakan lingkungan akademik, fasilitas, data, dan dukungan untuk penelitian ini.

Penghargaan dan terima kasih yang mendalam disampaikan kepada istri tercinta, Wida Damayanti, S.Kom., dan anak tersayang, Daffa Abiyyi Muflih, atas doa, kesabaran, pengertian, kasih sayang, serta dukungan yang terus diberikan selama masa studi. Semoga hasil penelitian ini bermanfaat bagi perkembangan ilmu pengetahuan dan penerapan teknologi pemrosesan ucapan di Indonesia.

## Daftar Pustaka


Bullock, L., Bredin, H., dan Garcia-Perera, L. P. (2020): Overlap-Aware Diarization: Resegmentation Using Neural End-to-End Overlapped Speech Detection, ICASSP 2020, 7114–7118. https://doi.org/10.1109/ICASSP40776.2020.9053096

Fujita, Y., Kanda, N., Horiguchi, S., Nagamatsu, K., dan Watanabe, S. (2019a): End-to-end neural speaker diarization with permutation-free objectives, Proceedings of INTERSPEECH. https://doi.org/10.21437/Interspeech.2019-2899

Fujita, Y., Kanda, N., Horiguchi, S., Xue, Y., Nagamatsu, K., dan Watanabe, S. (2019b): End-To-End Neural Speaker Diarization with Self-Attention, 2019 IEEE Automatic Speech Recognition and Understanding Workshop, 296–303. https://doi.org/10.1109/ASRU46091.2019.9003959

Gu, A., dan Dao, T. (2024): Mamba: Linear-Time Sequence Modeling with Selective State Spaces, arXiv:2312.00752, 1–36.

Horiguchi, S., Fujita, Y., Watanabe, S., Xue, Y., dan Nagamatsu, K. (2020): End-to-end speaker diarization for an unknown number of speakers with encoder-decoder based attractors, Proceedings of INTERSPEECH. https://doi.org/10.21437/Interspeech.2020-1022

Horiguchi, S., Fujita, Y., Watanabe, S., Xue, Y., Garcia, P., dan García-Perera, L. P. (2021): Encoder-Decoder Based Attractors for End-to-End Neural Diarization, IEEE/ACM Transactions on Audio, Speech, and Language Processing, 30, 1493–1507. https://doi.org/10.1109/TASLP.2022.3162080

Miro, X. A., Bozonnet, S., Evans, N., Fredouille, C., Friedland, G., dan Vinyals, O. (2012): Speaker Diarization: A Review of Recent Research, IEEE Transactions on Audio, Speech and Language Processing, 20(2), 356–370. https://doi.org/10.1109/TASL.2011.2125954

Park, T. J., Kanda, N., Dimitriadis, D., Han, K. J., Watanabe, S., dan Narayanan, S. (2022): A review of speaker diarization: Recent advances with deep learning, Computer Speech and Language, 72. https://doi.org/10.1016/j.csl.2021.101317

Raj, D. (2021): Multi-Class Spectral Clustering with Overlaps for Speaker Diarization, 2021 IEEE Spoken Language Technology Workshop. https://doi.org/10.1109/SLT48900.2021.9383602

Singh, P., Villalba, J., dan Dehak, N. (2025): Count Your Speakers! Multitask Learning for Multimodal Speaker Diarization, Proceedings of INTERSPEECH, 1588–1592. https://doi.org/10.21437/Interspeech.2025-334

Takashima, Y., Fujita, Y., Watanabe, S., Horiguchi, S., Garcia, P., dan Nagamatsu, K. (2021): End-to-End Speaker Diarization Conditioned on Speech Activity and Overlap Detection, 2021 IEEE Spoken Language Technology Workshop. https://doi.org/10.1109/SLT48900.2021.9383555

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., dan Polosukhin, I. (2017): Attention Is All You Need, Advances in Neural Information Processing Systems, 5999–6009