# Peningkatan *Automatic Speech Recognition* Dialek Bugis-Makassar Menggunakan Data Ucapan Sintetis

**Muh Fatwah Fajriansyah Marlang*, Herdianti Darwis, Huzain Azis, Abdul Rachman Manga**

Prodi Teknik Informatika, Fakultas Ilmu Komputer, Universitas Muslim Indonesia, Indonesia.

13020230319@student.umi.ac.id, herdianti.darwis@umi.ac.id, huzain.azis@umi.ac.id, abdulrachman.manga@umi.ac.id

**Abstract.** Speech in the Bugis-Makassar dialect poses a major challenge for Automatic Speech Recognition (ASR) systems, particularly in detecting clitic particles that are essential to utterance meaning but do not exist in standard Indonesian. This study investigates the effect of synthetic speech data augmentation using OpenAI's gpt-4o-mini-tts on ASR performance for this dialect, testing four ratios of synthetic-to-real data (0%, 50%, 100%, and 150%) through fine-tuning of the wav2vec2-large-xlsr-indonesian model. Evaluation was conducted using Word Error Rate (WER), Character Error Rate (CER), and Particle Detection Rate (PDR) as a dedicated metric for clitic particle detection. Results show that augmentation at a 50% ratio produced consistent improvements across all metrics compared to baseline, with WER decreasing by 3.47 percentage points and PDR increasing by 3.62 percentage points. Further increases in ratio progressively reduced these gains, forming an inverted-U pattern, with performance at a 150% ratio returning to near-baseline levels. Acoustic analysis identified a significant pitch distribution gap between synthetic and real speech (a difference of 58.09 Hz, or 24.8%) as a contributing factor to performance degradation at higher ratios. These findings indicate that TTS-based synthetic data augmentation effectively supports low-resource dialectal ASR, but only when applied at a controlled, moderate ratio.

**Keywords:** *Automatic Speech Recognition, Bugis-Makassar Dialect, Text to Speech, Dialectal Particles, Low-Resource ASR.*

**Abstrak.** Tuturan dialek Bugis-Makassar merupakan tantangan besar bagi sistem *Automatic Speech Recognition* (ASR), khususnya dalam mendeteksi partikel klitik yang penting bagi makna ujaran namun tidak terdapat dalam Bahasa Indonesia standar. Penelitian ini menginvestigasi pengaruh augmentasi data sintetis menggunakan OpenAI *gpt-4o-mini-tts* terhadap performa ASR dialek tersebut, dengan menguji empat rasio data sintetis terhadap data asli (0%, 50%, 100%, dan 150%) melalui *fine-tuning* model *wav2vec2-large-xlsr-indonesian*. Evaluasi dilakukan menggunakan *Word Error Rate* (WER), *Character Error Rate* (CER), dan *Particle Detection Rate* (PDR) sebagai metrik khusus deteksi partikel klitik. Hasil menunjukkan bahwa augmentasi pada rasio 50% menghasilkan peningkatan konsisten pada seluruh metrik dibandingkan *baseline*, dengan WER turun 3,47 poin persentase dan PDR naik 3,62 poin persentase. Penambahan rasio secara bertahap mengurangi manfaat tersebut dan membentuk pola kurva *inverted-U*, di mana performa pada rasio 150% kembali mendekati *baseline*. Analisis akustik mengidentifikasi perbedaan distribusi *pitch* yang signifikan antara data sintetis dan asli (selisih 58,09 Hz atau 24,8%) sebagai faktor yang berkontribusi terhadap degradasi performa pada rasio tinggi. Temuan ini menunjukkan bahwa augmentasi data sintetis berbasis TTS efektif mendukung ASR dialek *low-resource*, namun hanya jika diterapkan pada rasio yang terkontrol dan moderat.

**Kata Kunci:** *Automatic Speech Recognition, Dialek Bugis-Makassar, Text to Speech, Partikel Dialek, ASR Low Resource.*

*13020230319@student.umi.ac.id

## A. Pendahuluan

Indonesia memiliki keragaman bahasa daerah yang tinggi, termasuk dialek Bugis-Makassar yang dituturkan luas di Sulawesi Selatan. Salah satu ciri khas dialek ini adalah keberadaan partikel klitik seperti mi, ji, ko, na, ki', ka', dan ta' yang berperan penting dalam menyampaikan makna pragmatis, seperti penekanan, kesantunan, dan hubungan sosial. Kesalahan mengenali partikel-partikel tersebut dapat mengubah makna kalimat secara signifikan.

Kemajuan *Automatic Speech Recognition* (ASR) telah menghasilkan sistem dengan akurasi tinggi untuk bahasa berdaya tinggi, tetapi performanya masih rendah pada bahasa daerah *low-resource* (Bartelds et al., 2023). Penelitian oleh (Azis et al., 2026) menunjukkan bahwa model ASR yang telah diadaptasi untuk Bahasa Indonesia masih menghasilkan *Word Error Rate* (WER) sebesar 87,73% dan *Particle Detection Rate* (PDR) hanya 2,9% pada dialek Bugis-Makassar. Temuan ini menunjukkan bahwa evaluasi ASR pada dialek tidak cukup hanya menggunakan WER, tetapi juga memerlukan metrik khusus seperti PDR untuk mengukur kemampuan mendeteksi partikel klitik.

Salah satu pendekatan yang banyak digunakan untuk mengatasi keterbatasan data adalah augmentasi menggunakan *Text-to-Speech* (TTS). Berbagai penelitian melaporkan bahwa data sintetis mampu meningkatkan performa ASR pada bahasa *low-resource* (Bartelds et al., 2023; Casanova et al., 2022; Yang et al., 2024). Namun, efektivitasnya dipengaruhi oleh *distributional mismatch* antara ucapan sintetis dan ucapan asli, terutama pada aspek pitch, prosodi, dan kecepatan bicara (Lim et al., 2025; Srinivasan et al., 2021; Yadav & Pradhan, 2021).

Berdasarkan kajian tersebut, masih belum terdapat penelitian yang mengkaji pengaruh rasio data sintetis terhadap kemampuan deteksi partikel klitik pada dialek Bugis-Makassar. Oleh karena itu, penelitian ini menginvestigasi pengaruh augmentasi TTS terhadap performa ASR dengan fokus pada WER dan PDR, menentukan rasio data sintetis yang optimal, serta menganalisis faktor akustik yang memengaruhi perubahan performa. Kontribusi penelitian ini meliputi evaluasi *pipeline* augmentasi TTS, analisis pengaruh rasio data sintetis, dan diagnosis *distributional mismatch* sebagai dasar pengembangan ASR untuk bahasa daerah *low-resource.*

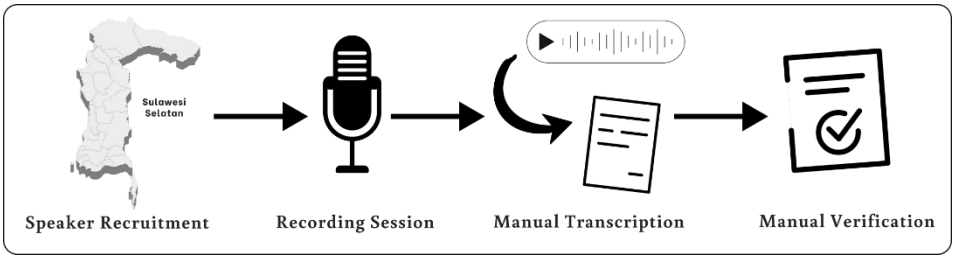
## B. Metode

Penelitian ini terdiri dari empat tahap utama: (1) pengumpulan dan persiapan dataset, (2) pembangunan korpus sintetis, (3) perancangan eksperimen augmentasi, dan (4) evaluasi performa model. Keseluruhan tahapan dirancang untuk memastikan konsistensi dan keterbandingan antar kondisi eksperimen.

### Dataset

Dataset yang digunakan merupakan rekaman tuturan spontan penutur asli dialek Bugis-Makassar di Sulawesi Selatan, terdiri dari 1.288 *utterance* dari 22 *speaker* dengan total durasi sekitar 70 menit. Dataset dibagi menjadi *pool* pelatihan (1.106 *utterance*, 20 *speaker*) dan himpunan uji (182 *utterance*, 2 *speaker*) yang dipisahkan secara *speaker-independent*, *speaker* pada himpunan uji tidak pernah muncul dalam data pelatihan dalam bentuk apapun. Penutur dalam dataset berasal dari tujuh kabupaten/kota di Sulawesi Selatan, yaitu Makassar, Takalar, Palopo, Maros, Pangkep, Bone, dan Enrekang.

Penutur dalam dataset berasal dari tujuh kabupaten/kota di Sulawesi Selatan, yaitu Makassar, Takalar, Palopo, Maros, Pangkep, Bone, dan Enrekang. Gambar 1 menggambarkan alur pengumpulan data penelitian ini. Proses pengumpulan data dilakukan dengan merekrut penutur asli dari ketujuh kabupaten/kota tersebut, kemudian merekam tuturan mereka menggunakan mikrofon dan perangkat *smartphone*. Sesi perekaman dilakukan melalui dua pendekatan, yaitu mengajukan pertanyaan langsung kepada penutur atau memberikan satu topik tertentu untuk dibicarakan secara bebas. Setelah proses perekaman selesai, seluruh audio ditranskripsi secara manual, kemudian diverifikasi secara manual untuk memastikan akurasi teks referensi.




**Gambar 1.** Diagram Alur Pengumpulan Data

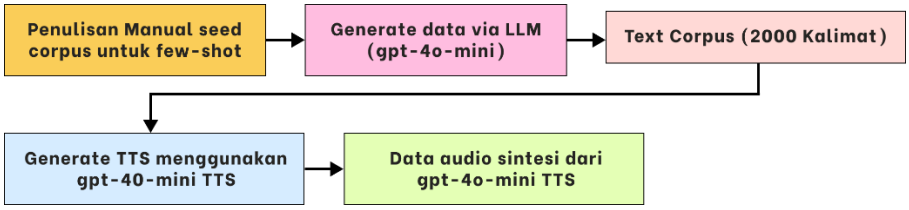
Komposisi dialek terdiri dari tuturan Bugis-Makassar, yang mencakup beberapa partikel klitik target dengan frekuensi kemunculan yang sangat bervariasi dari yang paling umum seperti ki' (354 kemunculan) dan di' (308 kemunculan) hingga yang sangat langka seperti pale' dan jeko yang masing-masing muncul kurang dari 15 kali. Tabel 1 menyajikan gambaran partikel klitik target beserta contoh penggunaannya dalam konteks kalimat, diambil dari himpunan uji (182 *utterance*).

**Tabel 1.** Contoh Penggunaan Partikel Klitik dalam Korpus (Himpunan Uji)

| Partikel | Kemunculan di Data Uji | Contoh Kalimat |
|---|---|---|
| ka’ | 42 | "alhasil sampai ka' di mesjid itu full mi" |
| na | 24 | "lucu na" |
| mi | 23 | "banyak sekali mi perubahannya guys ternyata" |
| ki’ | 22 | "baru habis itu ganti menu ki' toh?" |
| ji | 20 | "padahal sedekat itu ji loh dari makassar ke takalar" |
| ma’ | 16 | "di situ maghrib mi pergi ma' shalat tarwih" |
| ta’ | 14 | "bagaimana kuliah ta' terus mereka ternyata di unm masuk mi kuliah" |
| meki’ | 12 | "iyo cepat cepat meki' kerjai supaya ndada mi di pikir lagi" |
| di’ | 9 | "manai ais di' di dalam i kapang duduk" |
| to’ | 5 | "e lama lama di makassar to' maksud ku' lama ka' lagi" |
| pale’ | 5 | "kemarin tadi malam pale'" |
| pa’ | 4 | "ndak tau, sama pa' latipa bentar" |
| ko | 3 | "jan meki' kaget nah kalo tiba tiba berhenti di lante ko yang kosong" |
| mo | 3 | "saya bagian pembahasan mo ku kerja" |
| tawwa | 2 | "rame sekali tawwa di stadion teriaki lagi semua orang bilang ewako psm" |

**Pembangunan Korpus Sintetis**

Korpus sintetis dibangun dengan menyusun kalimat seed dari tuturan penutur asli sebagai contoh few-shot, kemudian menghasilkan kalimat baru menggunakan *Large Language Model* (LLM) melalui teknik *few-shot prompting* yang difokuskan pada partikel langka seperti pale’, tawwa, jeko, dan mo. Pendekatan ini sejalan dengan penelitian yang menunjukkan bahwa audio sintetis dapat meningkatkan pengenalan kata *out-of-vocabulary* (OOV) yang jarang muncul dalam data pelatihan (Zheng et al., 2021). Seluruh kalimat disintesis menjadi audio menggunakan OpenAI GPT-4o-mini-TTS dalam konfigurasi *tuned* (instruksi prosodi eksplisit, *speed* = 1.15). Seluruh audio di-*resample* ke 16.000 Hz sebelum digunakan un tuk pelatihan. Gambar 2 menunjukkan *pipeline* pembangunan korpus sintetis dialek Bugis-Makassar yang terdiri dari empat tahap: penyusunan kalimat *seed*, generasi kalimat berbasis LLM dan sintesis audio TTS.




**Gambar 2.** Diagram *Pipeline* Korpus Sintetis

Analisis akustik menunjukkan bahwa data sintetis memiliki rata-rata pitch yang lebih rendah dibandingkan data asli (175,72 Hz vs. 233,81 Hz) dengan variabilitas yang juga lebih kecil. Mengingat model *self-supervised* berbasis wav2vec 2.0 secara umum diketahui mengenkode informasi akustik seperti pitch pada lapisan-lapisan awal representasinya (Chiu et al., 2025), perbedaan distribusi ini perlu dipertimbangkan dalam menginterpretasikan hasil augmentasi.

**Model Dasar**

Model yang digunakan sebagai fondasi dalam penelitian ini adalah *wav2vec2-large-xlsr-indonesian*, yaitu model ASR berbasis arsitektur wav2vec 2.0 yang dibangun di atas XLSR-53. XLSR-53 dipra-latih (*pretrained*) secara *self-supervised* dari gelombang suara mentah (*raw waveform*) dalam 53 bahasa dengan total lebih dari 56.000 jam data audio tanpa label, mencakup data dari MLS, CommonVoice, dan BABEL (Conneau et al., 2021). Model ini kemudian diadaptasi lebih lanjut untuk Bahasa Indonesia standar sebelum digunakan dalam penelitian ini. Model dilatih menggunakan mekanisme *Connectionist Temporal Classification* (CTC) untuk tugas pengenalan suara otomatis. Karena XLSR-53 dipra-latih dari data Bahasa Indonesia standar, model ini tidak memiliki paparan terhadap partikel klitik khas dialek Bugis-Makassar sebelum proses *fine-tuning* dalam penelitian ini.

**Desain Eksperimen**

Eksperimen dirancang menggunakan empat kondisi rasio data sintetis terhadap data asli, yang diinspirasi oleh rekomendasi literatur mengenai interval rasio yang bersih dan mudah diinterpretasikan. Target rasio yang ditetapkan adalah 0%, 50%, 100%, dan 150% relatif terhadap jumlah data asli. Keempat kondisi ini selanjutnya disebut sebagai E0, E1, E2, dan E3. Rincian komposisi data pada setiap kondisi disajikan pada Tabel 2.

**Tabel 2.** Komposisi data pada setiap kondisi eksperimen

| Kondisi | Rasio Sintetis | Data Asli | Data Sintetis | Total |
|---|---|---|---|---|
| **E0** | 0% | 996 kalimat | 0 kalimat | 996 kalimat |
| **E1** | 50% | 996 kalimat | 498 kalimat | 1494 kalimat |
| **E2** | 100% | 996 kalimat | 996 kalimat | 1992 kalimat |
| **E3** | 150% | 996 kalimat | 1494 kalimat | 2490 kalimat |

**Metrik Evaluasi**

Evaluasi dilakukan menggunakan tiga metrik yang saling melengkapi. *Word Error Rate* (WER) mengukur proporsi kata yang ditranskripsi secara tidak tepat terhadap total kata dalam referensi, dihitung berdasarkan jumlah operasi substitusi (*S*), penghapusan (*D*), dan penyisipan (*I*) minimum yang diperlukan untuk mengubah hipotesis menjadi referensi:

$$\text{WER} = \frac{S + D + I}{N} \times 100\% \quad \ldots (1)$$

Pada persamaan (1), S, D, dan I masing-masing merupakan jumlah operasi substitusi, penghapusan, dan penyisipan, sedangkan N merupakan total kata dalam referensi.

*Character Error Rate* (CER) menggunakan formulasi yang identik namun dihitung pada level karakter, sehingga lebih sensitif terhadap token pendek seperti partikel klitik:

$$\text{CER} = \frac{S_c + D_c + I_c}{N_c} \times 100\% \quad \ldots (2)$$

Pada persamaan (2), Sc, Dc, dan Ic masing-masing jumlah operasi substitusi, penghapusan, dan penyisipan pada level karakter, sedangkan Nc merupakan total karakter dalam referensi.

*Particle Detection Rate* (PDR) merupakan metrik khusus yang dikembangkan untuk penelitian ini, didefinisikan sebagai jumlah partikel yang terdeteksi dengan benar ($H$) dibagi total partikel dalam referensi ($N_p$), dinyatakan dalam persentase:

$$\text{PDR} = \frac{H}{N_p} \times 100\% \quad \ldots(3)$$

Pada persamaan (3), H merupakan jumlah partikel yang terdeteksi dengan benar, sedangkan Np merupakan total partikel dalam referensi. PDR dihitung setelah normalisasi apostrof pada kedua sisi untuk menghindari penalti ganda. Metrik ini secara langsung mengukur kemampuan inti yang menjadi fokus penelitian ini, karena WER agregat tidak dapat membedakan apakah model berhasil atau gagal mendeteksi partikel klitik secara spesifik.

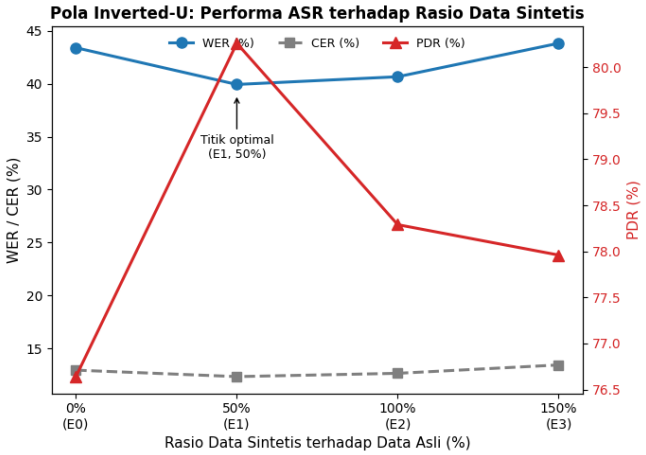
## C. Hasil Penelitian dan Pembahasan

### Hasil Augmentasi Data Sintetis

Eksperimen augmentasi data sintetis dilakukan pada empat kondisi rasio yang menghasilkan pola performa yang jelas dan konsisten. Tabel 3 menyajikan hasil evaluasi keempat kondisi berdasarkan metrik WER, CER, dan PDR.

**Tabel 3.** Hasil evaluasi performa model pada keempat kondisi eksperimen

| Kondisi | Rasio Sintetis (%) | WER (%) | CER (%) | PDR (%) |
|---|---|---|---|---|
| E0 | 0 | 43.40 | 12.94 | 76.64 |
| E1 | 50 | 39.93 | 12.33 | 80.26 |
| E2 | 100 | 40.66 | 12.64 | 78.29 |
| E3 | 150 | 43.82 | 13.43 | 77.96 |




**Gambar 3.** Perbandingan WER, CER, dan PDR pada Berbagai Rasio Data Sintetis

Gambar 3 memvisualisasikan pola inverted-U tersebut, di mana performa terbaik (WER dan CER terendah, PDR tertinggi) tercapai pada kondisi E1 (rasio 50%), sebelum menurun secara konsisten seiring peningkatan proporsi data sintetis. Hasil menunjukkan bahwa kondisi E1 dengan rasio sintetis 50% menghasilkan performa terbaik pada seluruh metrik evaluasi, dengan peningkatan WER sebesar 3,47 poin persentase, CER sebesar 0,61 poin persentase, dan PDR sebesar 3,62 poin persentase dibandingkan baseline E0. Sebaliknya, peningkatan rasio data sintetis secara bertahap pada E2 dan E3 menunjukkan penurunan manfaat yang konsisten hingga performa pada E3 kembali mendekati atau melampaui tingkat kesalahan baseline. Pola ini membentuk kurva berbentuk inverted-U yang mengindikasikan adanya titik optimal dalam penggunaan data sintetis.

**Diagnosis *Mismatch* Akustik**

Untuk menjawab RQ3, dilakukan analisis akustik langsung terhadap karakteristik audio data asli dan data sintetis. Hasil analisis pitch disajikan pada Tabel 4.

**Tabel 4.** Perbandingan distribusi pitch antara data asli dan data sintetis

| Metrik | Data Asli | Data Sintetis | Selisih |
|---|---|---|---|
| Mean pitch (Hz) | 233.81 | 175.72 | - 58.09 (24.8%) |
| Median pitch (Hz) | 238.29 | 175.45 | - 62.83 (26.4%) |
| Std. deviasi pitch | 47.48 | 30.05 | - 17.43 (36.73%) |

Analisis menunjukkan bahwa data sintetis memiliki distribusi pitch yang secara sistematis berbeda dari data asli, baik dalam hal rata-rata (selisih 58,09 Hz atau 24,8%) maupun variabilitas (selisih standar deviasi 36,7%). Perbedaan variabilitas yang lebih besar dari perbedaan rata-rata mengindikasikan bahwa ucapan sintetis bukan hanya lebih rendah nadanya, tetapi juga jauh lebih seragam dan monoton dibandingkan tuturan asli yang lebih ekspresif. Perbedaan distribusi pitch yang teramati (58,09 Hz, 24,8%) berkorelasi langsung dengan pola penurunan performa pada rasio sintetis tinggi (E3), sejalan dengan konsep *distributional mismatch* dalam augmentasi data dari distribusi berbeda (Lim et al., 2025), meskipun pitch merupakan salah satu dari beberapa elemen prosodi yang berkontribusi terhadap *mismatch* tersebut. Salah satu kemungkinan mekanisme yang mendasari korelasi ini adalah bahwa model *self-supervised* berbasis wav2vec 2.0 secara umum diketahui mengenkode informasi akustik seperti pitch pada lapisan-lapisan awal representasinya (Chiu et al., 2025), sehingga ketidaksesuaian distribusi pitch berpotensi mengganggu representasi tersebut ketika data sintetis mendominasi proses *fine-tuning*.

**Pembahasan**

Temuan penelitian ini secara keseluruhan menunjukkan bahwa augmentasi data sintetis TTS memberikan manfaat yang bersifat kondisional, efektif pada rasio rendah namun merugikan pada rasio tinggi. Hasil ini konsisten dengan temuan (Bartelds et al., 2023) yang menunjukkan manfaat augmentasi TTS untuk ASR *low-resource*, sekaligus memperluas pemahaman tersebut dengan menunjukkan bahwa manfaat tersebut memiliki batas optimal yang bergantung pada tingkat *mismatch* akustik antara data sintetis dan data asli.

Pada rasio rendah (E1, 50%), data asli yang masih mendominasi setiap *batch* pelatihan menjaga representasi akustik model tetap terjangkar pada pola suara asli, sementara data sintetis memberikan paparan tambahan terhadap konteks kalimat baru yang mengandung partikel langka. Sebaliknya, pada rasio tinggi (E3, 150%), data sintetis mendominasi distribusi pelatihan sehingga perbedaan akustik yang terukur, khususnya *gap pitch* sebesar 58,09 Hz dan perbedaan variabilitas prosodi sebesar 36,7% mulai mengganggu representasi yang telah terbentuk dari data asli. Mekanisme ini konsisten dengan konsep *distributional mismatch* dalam konteks augmentasi data dari distribusi berbeda (Lim et al., 2025).

Perlu dicatat bahwa penelitian ini menggunakan satu *run* pelatihan per kondisi. Berdasarkan karakterisasi *noise floor* dari enam baseline P0 independen yang menunjukkan rentang variasi alami sebesar 4,47 poin persentase, temuan ini perlu dipandang sebagai hasil indikatif yang memerlukan replikasi multi-*seed* untuk konfirmasi statistik yang lebih kuat.

Perlu dicatat pula bahwa penelitian ini berfokus pada analisis distribusi pitch sebagai indikator *distributional mismatch* akustik. Studi lain yang mengisolasi kontribusi masing-masing elemen prosodi menunjukkan bahwa durasi dan energi ujaran justru memberikan kontribusi yang lebih besar dibandingkan pitch dalam menurunkan *substitution error* pada augmentasi data sintetis (Lim et al., 2025). Dengan demikian, gap pitch yang teridentifikasi dalam penelitian ini kemungkinan merupakan salah satu dari beberapa faktor akustik yang berkontribusi terhadap pola *inverted-U* yang teramati, dan analisis elemen prosodi lain seperti durasi dan energi ujaran dapat menjadi arah penelitian lanjutan untuk memperkuat diagnosis mekanisme *mismatch* akustik secara lebih komprehensif.

## D. Kesimpulan

Penelitian ini menginvestigasi efek augmentasi data sintetis berbasis TTS terhadap performa ASR dialek Bugis-Makassar dengan fokus pada deteksi partikel klitik. Hasil eksperimen menunjukkan bahwa augmentasi data sintetis memberikan manfaat yang bersifat kondisional, efektif pada rasio rendah namun merugikan pada rasio tinggi. Pada rasio 50% (E1), seluruh metrik evaluasi meningkat secara konsisten dibandingkan baseline, dengan WER turun 3,47 poin persentase dan PDR naik 3,62 poin persentase. Sebaliknya, penambahan rasio secara bertahap membentuk pola *inverted-U* hingga performa pada rasio 150% (E3) kembali mendekati tingkat kesalahan baseline.

Analisis akustik mengidentifikasi *distributional mismatch* sebagai salah satu mekanisme yang berkontribusi terhadap degradasi pada rasio tinggi. Perbedaan distribusi pitch antara data sintetis (175,72 Hz) dan data asli (233,81 Hz), disertai variabilitas prosodi yang 36,7% lebih rendah pada data sintetis, berpotensi mengganggu representasi akustik model ketika proporsi data sintetis mendominasi pelatihan (Chiu et al., 2025; Lim et al., 2025).

Untuk penelitian selanjutnya, disarankan replikasi multi-*seed*, analisis per-partikel yang lebih granular, pengujian kondisi sintetis-saja, serta perbandingan antar sub-dialek Makassar dan Bugis

**Ucapan Terimakasih**

Penulis mengucapkan terima kasih yang sebesar-besarnya kepada Universitas Muslim Indonesia atas dukungan institusional yang diberikan selama pelaksanaan penelitian ini, termasuk penyediaan fasilitas dan lingkungan akademik yang mendukung proses penelitian dari tahap pengumpulan data hingga penyusunan artikel ini. Ucapan terima kasih juga disampaikan kepada Fakultas Ilmu Komputer atas bimbingan, arahan, serta dukungan akademik yang diberikan kepada penulis selama proses penelitian berlangsung. Semoga penelitian ini dapat memberikan kontribusi yang bermanfaat bagi pengembangan teknologi *Automatic Speech Recognition* untuk bahasa daerah di Indonesia, khususnya dialek Bugis-Makassar

**Daftar Pustaka**

Azis, H., Fajriansyah M, M. F., Darwis, H., Purnawansyah, Hasanuddin, T., & Sugiarti. (2026). Development of a Low-Resource Automatic Speech Recognition System for the Makassar Dialect. *2026 20th International Conference on Ubiquitous Information Management and Communication (IMCOM)*, 1–8. https://doi.org/10.1109/IMCOM69009.2026.11360893

Bartelds, M., San, N., McDonnell, B., Jurafsky, D., & Wieling, M. (2023). Making More of Little Data: Improving Low-Resource Automatic Speech Recognition Using Data Augmentation. In A. Rogers, J. Boyd-Graber, & N. Okazaki (Eds.), *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)* (pp. 715–729). Association for Computational Linguistics. https://doi.org/10.18653/v1/2023.acl-long.42

Casanova, E., Shulby, C. D., Korolev, A., Júnior, A. C., da Silva Soares, A., Aluísio, S. M., & Ponti, M. A. (2022). ASR data augmentation in low-resource settings using cross-lingual multi-speaker TTS and cross-lingual voice conversion. *Interspeech*. https://doi.org/10.21437/Interspeech.2023-496

Chiu, A. Y. F., Fung, P. K. C., Li, R. T. Y., Li, J., & Lee, T. (2025). *A Large-Scale Probing Analysis of Speaker-Specific Attributes in Self-Supervised Speech Representations*. https://doi.org/10.48550/arXiv.2501.05310

Conneau, A., Baevski, A., Collobert, R., Mohamed, A., & Auli, M. (2021). Unsupervised Cross-Lingual Representation Learning for Speech Recognition. *Interspeech 2021*, 2426–2430. https://doi.org/10.21437/Interspeech.2021-329

Fajar, A. H., & Rejeki, Y. S. (2021). Perancangan Fasilitas Kerja Ergonomis pada Stasiun Persiapan Menggunakan Analisis Virtual Environment Modelling. *Jurnal Riset Teknik Industri*, *1*(2), 121–130. https://doi.org/10.29313/jrti.v1i2.413

Lim, Y., Kim, D., & Kim, S. H. (2025). Acoustic and linguistic effects in synthesized speech augmentation for speech recognition. *ETRI Journal*, *47*(6), 1061–1070. https://doi.org/10.4218/etrij.2024-0050

Naufal Rafif Rizaldi, Djamaluddin, & Ajrina Febri Suahati. (2024). Implementasi Sistem Informasi Berbasis Enterprise Resource Planning (ERP) dengan Menggunakan Software Accurate. *Jurnal Riset Teknik Industri*, 169–178. https://doi.org/10.29313/jrti.v4i2.5483

Rosdiana, H., Ashari, Y., & Fauzi Isniarno, N. (n.d.). Pengaruh Intensitas Curah Hujan terhadap Kegiatan Pemompaan pada Sump Berdasarkan Water Balance di PT XYZ. *Jurnal Sangkuriang Bandung*. Retrieved https://journal.sbpublisher.com/index.php/minetech

Srinivasan, A., Singh, D., Yarra, C., Illa, A., & Ghosh, P. K. (2021). A Robust Speaking Rate Estimator Using a CNN-BLSTM Network. *Circuits, Systems, and Signal Processing*, *40*(12), 6098–6120. https://doi.org/10.1007/s00034-021-01754-1

Yadav, I. C., & Pradhan, G. (2021). Pitch and noise normalized acoustic feature for children's ASR. *Digital Signal Processing*, *109*, 102922. https://doi.org/10.1016/j.dsp.2020.102922

Yang, G., Yu, F., Ma, Z., Du, Z., Gao, Z., Zhang, S., & Chen, X. (2024). Enhancing Low-Resource ASR through Versatile TTS: Bridging the Data Gap. *ICASSP 2025 - 2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 1–5. https://doi.org/10.1109/ICASSP49660.2025.10889894

Zheng, X., Liu, Y., Gunceler, D., & Willett, D. (2021). Using Synthetic Audio to Improve the Recognition of Out-of-Vocabulary Words in End-to-End Asr Systems. *ICASSP 2021 - 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 5674–5678. https://doi.org/10.1109/ICASSP39728.2021.9414778