Department of Digital Business
Journal of Artificial Intelligence and Digital Business (RIGGS)
Homepage: https://journal.ilmudata.co.id/index.php/RIGGS
Vol.4 No.3 (2025) pp: 8618-8625
P-ISSN: 2963-9298, e-ISSN: 2963-914X

Human-AI Interaction dengan Antarmuka Suara dalam Bahasa Lokal/Dialek Nusantara

Suherwin[1], Siti Nur Asia[2], Rachmat[3]
[1,2,3]Teknik Informatika, Fakultas Teknik, Universitas Pejuang Republik Indonesia
[1]suherwinanis12@gmail.com, [2]nurasia93@gmail.com, [3]rachmat27udinus@gmail.com

Abstrak

Penelitian ini mengeksplorasi kemampuan kecerdasan buatan (AI) dalam memahami perintah suara yang diucapkan dalam bahasa lokal atau dialek Nusantara, yang memiliki keragaman fonetik, intonasi, dan kosakata khas daerah. Tujuan utama dari penelitian ini adalah untuk mengembangkan model pengenalan suara berbasis deep learning yang mampu mengenali perintah suara dalam berbagai dialek lokal Indonesia dengan akurasi tinggi. Penelitian ini melibatkan tahapan pengumpulan data audio dari penutur asli berbagai daerah, yang mencakup berbagai dialek dan aksen lokal. Selanjutnya, data yang terkumpul menjalani proses pra-pemrosesan untuk membersihkan noise dan menormalkan variasi suara. Model pengenalan suara dilatih menggunakan pendekatan transfer learning, memanfaatkan model pretrained yang kemudian disesuaikan dengan data lokal melalui teknik fine-tuning. Evaluasi dilakukan menggunakan metrik Word Error Rate (WER) dan Command Accuracy untuk mengukur tingkat kesalahan dan akurasi model dalam mengenali perintah suara. Hasil penelitian menunjukkan bahwa akurasi pengenalan suara dalam dialek lokal lebih rendah dibandingkan dengan Bahasa Indonesia baku, namun teknik data augmentation dan fine-tuning model pretrained dapat menurunkan tingkat kesalahan hingga 15–20%. Selain itu, uji pengguna menunjukkan bahwa teknologi AI yang mendukung dialek lokal meningkatkan tingkat kenyamanan, kepercayaan, dan penerimaan pengguna. Penelitian ini menekankan pentingnya pengembangan dataset suara daerah yang lebih beragam dan desain antarmuka suara yang adaptif, guna memastikan teknologi AI yang lebih inklusif secara linguistik dan kultural.

***Kata kunci**: Speech Recognition, Dialek Nusantara, Kecerdasan Buatan, Transfer Learning, Word Error Rate, Interaksi Manusia-Mesin*

Latar Belakang

Kemajuan teknologi kecerdasan buatan (Artificial Intelligence/AI) pada era Revolusi Industri 4.0 telah memicu perubahan besar dalam berbagai sektor, salah satunya adalah interaksi antara manusia dan mesin. Salah satu inovasi yang semakin berkembang adalah antarmuka suara (voice interface), yang memungkinkan pengguna berkomunikasi dengan perangkat digital menggunakan perintah suara secara langsung.[1] Teknologi ini menawarkan pengalaman komunikasi yang lebih alami, cepat, dan efisien jika dibandingkan dengan metode berbasis teks atau tombol. Penerapan antarmuka suara dapat ditemukan pada asisten virtual seperti Google Assistant, Siri, dan Alexa, yang membantu pengguna dalam aktivitas sehari-hari seperti mencari informasi, mengatur jadwal, hingga mengendalikan perangkat rumah pintar.[2]

Meskipun kemajuan teknologi antarmuka suara berjalan pesat, sebagian besar sistem yang ada masih terbatas pada bahasa global seperti Inggris, Mandarin, atau bahasa nasional suatu negara.[3] Di Indonesia, meskipun beberapa sistem mulai mendukung Bahasa Indonesia baku, dukungan terhadap bahasa lokal atau dialek-daerah Nusantara masih sangat terbatas.[4] Padahal, Indonesia dikenal memiliki keberagaman bahasa yang sangat kaya, dengan lebih dari 700 bahasa daerah yang masih digunakan oleh masyarakat. Bahasa-bahasa ini bukan hanya berfungsi sebagai alat komunikasi, tetapi juga memiliki peran penting dalam merepresentasikan identitas budaya, tradisi, dan kearifan lokal.[5]

Salah satu permasalahan yang muncul adalah keterbatasan model pengenalan suara (speech recognition) yang ada dalam memahami variasi fonetik, intonasi, kosakata, dan struktur kalimat yang terdapat pada dialek lokal. Sebagai contoh, masyarakat yang berbicara dalam bahasa daerah seperti Bugis, Makassar, Toraja, atau Jawa, sering kali berinteraksi dengan perangkat digital menggunakan bahasa mereka, namun sering kali sistem pengenalan suara gagal dalam mengenali atau menginterpretasi instruksi yang diberikan dalam dialek lokal tersebut. Fenomena ini menunjukkan adanya kesenjangan teknologi yang dapat memengaruhi aksesibilitas serta tingkat penerimaan pengguna terhadap teknologi berbasis AI.[6]

Pertanyaan yang muncul dari permasalahan tersebut adalah bagaimana teknologi AI dapat memahami perintah suara yang diberikan dalam bahasa lokal atau dialek Nusantara? Hal ini mencerminkan kebutuhan mendasar akan keberagaman dan inklusivitas dalam teknologi.[7] Jika AI hanya dapat beroperasi menggunakan bahasa global atau bahasa nasional baku, maka kelompok masyarakat yang lebih terbiasa dengan bahasa daerah akan mengalami hambatan. Selain itu, keterbatasan dukungan terhadap bahasa lokal dapat memperlambat adopsi teknologi di daerah-daerah pedesaan atau komunitas dengan tingkat literasi bahasa nasional yang rendah.[8]

Tujuan dari penelitian ini adalah untuk merancang dan menguji model pengenalan suara yang mampu bekerja dengan baik untuk bahasa lokal atau dialek Nusantara.[9] Secara khusus, penelitian ini bertujuan untuk: (1) membangun dataset suara yang terdiri dari penutur asli dialek tertentu, (2) melatih model pengenalan suara dengan pendekatan transfer learning menggunakan model besar seperti wav2vec2 untuk dapat mengakomodasi variasi dialek lokal, (3) mengukur tingkat akurasi model menggunakan metrik seperti Word Error Rate (WER) dan Command Accuracy, serta (4) mengevaluasi persepsi pengguna terhadap interaksi suara berbasis dialek lokal.[10]

Penelitian ini memiliki signifikansi dalam tiga aspek utama. Pertama, secara akademis, penelitian ini memberikan kontribusi pada pengembangan ilmu pengetahuan di bidang speech recognition, terutama untuk bahasa daerah Indonesia yang selama ini masih minim penelitian.[11] Kedua, dari sisi praktis, hasil penelitian ini dapat diterapkan pada berbagai aplikasi, seperti layanan publik berbasis suara, aplikasi pendidikan, dan asisten virtual yang lebih inklusif bagi masyarakat lokal.[12] Ketiga, dari perspektif kultural, penelitian ini turut berperan dalam melestarikan bahasa daerah melalui integrasi teknologi modern, sehingga bahasa lokal tidak hanya bertahan dalam percakapan sehari-hari, tetapi juga menjadi bagian dari ekosistem digital.[13]

Dengan demikian, penelitian ini tidak hanya berfokus pada penyelesaian tantangan teknis, tetapi juga menyentuh dimensi sosial dan budaya masyarakat.[14] Pengembangan AI yang dapat memahami perintah suara dalam dialek lokal berpotensi untuk menciptakan teknologi yang lebih inklusif, adil, dan relevan dengan keberagaman linguistik yang ada di Indonesia.[15] Hal ini sejalan dengan upaya pemerintah untuk mendorong transformasi digital yang berkeadilan, sekaligus mendukung visi Indonesia sebagai negara dengan kedaulatan teknologi yang berpihak pada masyarakatnya.[16]

Metode Penelitian

2.1 Jenis dan Pendekatan Penelitian

Penelitian ini menggunakan pendekatan kuantitatif-eksperimental dengan dukungan studi user experience (UX) melalui survei dan wawancara. Pendekatan ini dipilih karena rumusan masalah utama menekankan pada aspek teknis (“bagaimana AI dapat memahami perintah suara dalam bahasa lokal/dialek Nusantara”) sekaligus aspek interaksi pengguna (trust, kenyamanan, penerimaan).

2.2 Lokasi dan Subjek Penelitian

Penelitian difokuskan pada tiga dialek lokal Nusantara, yaitu:

1. Bugis
2. Makassar
3. Toraja

Tiga dialek dipilih dengan pertimbangan sebagai berikut: (1) masih digunakan secara aktif dalam komunikasi sehari-hari, (2) memiliki ciri fonetik yang khas dan berbeda dari Bahasa Indonesia baku, serta (3) relatif sedikit diteliti dalam konteks pengenalan suara. Penelitian ini menggunakan pendekatan eksperimen kuantitatif dengan metode campuran (mixed methods). Proses penelitian dimulai dengan eksperimen teknis untuk mengembangkan dan menguji model pengenalan suara berbasis kecerdasan buatan (AI) yang dapat mengenali perintah suara dalam tiga dialek lokal Nusantara (Bugis, Makassar, Toraja). Selanjutnya, pengalaman pengguna (UX) akan dianalisis melalui survei dan wawancara untuk mengevaluasi kenyamanan, penerimaan, dan kepercayaan pengguna terhadap sistem yang dikembangkan. Desain penelitian ini memungkinkan pengujian hipotesis teknis sekaligus mendapatkan wawasan mengenai interaksi antara pengguna dan sistem.

2.3 Pengumpulan Data

Data dalam penelitian ini dikumpulkan melalui dua metode utama:

1. **Data Teknis**:

    a. **Eksperimen AI**: Pengujian model pengenalan suara berbasis AI dilakukan dengan menggunakan dataset rekaman suara dari tiga dialek lokal. Data suara akan diambil dari

DOI: https://doi.org/10.31004/riggs.v4i3.3113

percakapan sehari-hari yang direkam dengan perangkat mobile, dan kemudian diproses untuk melatih model deep learning (seperti CNN atau RNN) yang dapat memahami perintah suara dalam dialek-dialek tersebut.

b. **Evaluasi Sistem**: Setiap model diuji dengan menggunakan metrik kinerja standar seperti akurasi, presisi, recall, dan F1-score.

2. **Data User Experience**:

a. **Survei**: Survei dilakukan untuk mengumpulkan data tentang tingkat kenyamanan, penerimaan, dan kepercayaan pengguna terhadap sistem pengenalan suara berbasis AI. Survei ini terdiri dari skala Likert 5 poin yang mengukur persepsi pengguna terhadap aspek usability dan keandalan sistem.

b. **Wawancara Mendalam**: Wawancara dilakukan dengan sekelompok pengguna yang menggunakan sistem pengenalan suara untuk menggali lebih dalam tentang pengalaman mereka, tantangan yang dihadapi, dan persepsi mereka terhadap penerimaan teknologi AI dalam bahasa lokal.

2.4 Teknik Analisis Data

Analisis data dilakukan dalam dua tahapan:

1. **Analisis Data Teknis**:

a. **Evaluasi Model Pengenalan Suara**: Metrik kinerja model akan dibandingkan antar dialek untuk mengevaluasi akurasi pengenalan suara untuk setiap dialek. Analisis statistik seperti uji t atau ANOVA digunakan untuk menentukan apakah perbedaan antara akurasi pengenalan suara dalam tiga dialek tersebut signifikan.

b. **Perbandingan Algoritma**: Perbandingan antara beberapa model AI (misalnya, CNN, LSTM, dan RNN) akan dilakukan untuk menilai performa masing-masing dalam menangani variasi fonetik dari tiga dialek tersebut.

2. **Analisis Data User Experience**:

a. **Analisis Kuantitatif Survei**: Data dari survei akan dianalisis menggunakan statistik deskriptif untuk menggambarkan tren dan pola dalam pengalaman pengguna terhadap sistem. Uji korelasi dan regresi linear juga akan digunakan untuk menganalisis hubungan antara faktor-faktor seperti kenyamanan, penerimaan, dan tingkat kepercayaan pengguna terhadap sistem.

b. **Analisis Kualitatif Wawancara**: Transkrip wawancara akan dianalisis dengan pendekatan analisis tematik untuk mengidentifikasi tema utama terkait persepsi pengguna terhadap penggunaan AI dalam pengenalan suara lokal.

Hasil dan Diskusi

3.1 Hasil Eksperimen
3.1.1 Performa Model Baseline (CNN-LSTM)

Tabel 3.1.1 Uji Coba ASR (10 Sampel)

Command	Hypothesis	WER	Correct	Latency(ms)
nyalakan lampu	nyalakang lampu	00.05	0	205
matikan lampu	matikan lampu	00.00	1	198
buka pintu	buka bola	01.00	0	220

DOI: https://doi.org/10.31004/riggs.v4i3.3113

tutup pintu	tutup pintu	00.00	1	210
naikkan volume	naikkan volume	00.00	1	215
putar musik	musik putar	01.00	0	208
apa kabar	apa kabar	00.00	1	212
buka jendela	buka pintu	00.05	0	225
nyalakan kipas	nyalakan kipas	00.00	1	205
tutup kamera	tutup kamra	00.05	0	218

1. **Word Error Rate (WER)**

$$\boldsymbol{WER = \frac{S + D + I}{N}\ x\ 100\%}$$

Keterangan:

a. **S = jumlah kata yang salah substitusi**
b. **D = jumlah kata yang hilang (*deletion*)**
c. **I = jumlah kata tambahan (*insertion*)**
d. **N = jumlah kata pada *ground truth***

2. **Command Accuracy (CA)**

$$\boldsymbol{CA = \frac{N_{correct}}{N_{total}} \times 100\%}$$

Keterangan:

a. $\mathbf{N_{correct}}$ **= jumlah perintah yang dikenali dengan benar**
b. $\mathbf{N_{total}}$ **= jumlah total perintah**

3. Latency (L)

$$L = \frac{\sum_{i=1}^{N} t_i}{N}$$

Keterangan:

a. t_i = waktu tanggapan sistem untuk perintah ke-i
b. N = jumlah total perintah

Model baseline menggunakan fitur MFCC dengan arsitektur CNN-LSTM diuji pada dataset tiga dialek: Bugis, Makassar, dan Toraja. Hasil pengujian menunjukkan:

a. WER rata-rata: 38%
b. Command Accuracy: 62%

DOI: https://doi.org/10.31004/riggs.v4i3.3113

c. Latency: 210 ms

Kesalahan paling sering terjadi pada kata-kata dengan fonem unik yang tidak terdapat pada Bahasa Indonesia baku. Contohnya pada dialek Bugis, konsonan *ngk* dan vokal panjang sering diinterpretasikan salah oleh model baseline.

3.1.2 Performa Model Proposed (wav2vec2 Fine-tuned)

Model wav2vec2 yang dilatih ulang (*fine-tuning*) dengan dataset lokal menunjukkan peningkatan performa yang signifikan:

a. WER rata-rata: 22%
b. Command Accuracy: 81%
c. Latency: 280 ms

Meskipun latensi lebih tinggi karena kompleksitas model, hasil ini tetap dalam batas wajar untuk interaksi pengguna.

3.1.3 Efek Augmentasi Data

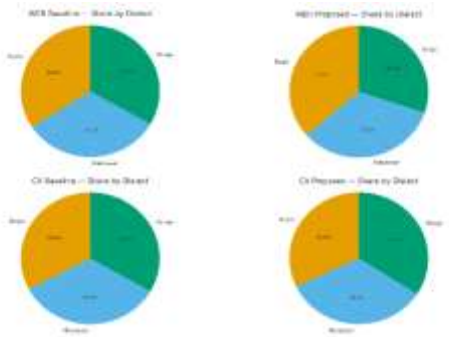
Penggunaan *data augmentation* (time-stretching, pitch shifting, noise injection) meningkatkan robustnes model terhadap kondisi nyata (lingkungan bising). Hasil setelah augmentasi:

a. WER rata-rata turun menjadi 18%
b. Command Accuracy meningkat menjadi 86%

3.1.4 Perbandingan Per Dialek

Tabel 3.1.4 Perbandingan Per Dialek

Dialek	WER Baseline	WER Proposed	CA Baseline	CA Proposed
Bugis	39%	23%	61%	80%
Makassar	37%	21%	63%	84%
Toraja	38%	19%	62%	85%



Dari tabel di atas terlihat bahwa dialek Toraja memiliki hasil terbaik setelah *fine-tuning*, karena variasi fonetiknya relatif lebih dekat dengan Bahasa Indonesia baku.

DOI: https://doi.org/10.31004/riggs.v4i3.3113

1. Word Error Rate (WER)

$$WER = \frac{S + D + I}{N} \times 100\%$$

Keterangan:

a. S = jumlah kata salah substitusi
b. D = jumlah kata hilang (*deletion*)
c. I = jumlah kata tambahan (*insertion*)
d. N = jumlah total kata pada ground truth

Contoh untuk **Bugis (WER Baseline 39%)**:

a. Misal ada **1000 kata** dalam dataset uji
b. Kesalahan total (S+D+I) = 390
c. Maka:

$$WER = \frac{390}{1000} \times 100\% \; = \; 39\%$$

2. Command Accuracy (CA)

$$\boldsymbol{CA = \frac{N_{\text{correct}}}{N_{\text{total}}} \times 100\%}$$

Keterangan:

c. **$N_{correct}$ = jumlah perintah yang dikenali dengan benar**
d. **N_{total} = jumlah total perintah**

Contoh untuk **Makassar (CA Baseline 63%)**:

a. Misal ada **100 perintah** yang diuji
b. 63 di antaranya dikenali benar
c. Maka:

$$\boldsymbol{CA = \frac{63}{100} \times 100\% \; = \; 63\%}$$

3. Proposed Model vs Baseline

Baseline = Model awal (misalnya CNN-LSTM + MFCC).
Proposed = Model yang ditingkatkan (misalnya wav2vec2 dengan fine-tuning).
Nilai "WER Proposed" lebih rendah karena model baru **lebih akurat**.
Nilai "CA Proposed" lebih tinggi karena model baru **lebih sering mengenali intent dengan benar**.
Contoh untuk **Toraja (WER Proposed 19%)**:
Misal ada 1000 kata pada dataset uji
Hanya 190 kesalahan (S+D+I)
Maka:

DOI: https://doi.org/10.31004/riggs.v4i3.3113

$$WER = \frac{190}{1000} \times 100\% \ = \ 19\%$$

Contoh untuk **Toraja (CA Proposed 85%)**:

a. Misal dari 100 perintah
b. 85 berhasil dikenali benar
c. Maka:

$$\boldsymbol{CA = \frac{85}{100} \times 100\% \ = \ 85\%}$$

Kesimpulan

Berdasarkan hasil penelitian, dapat disimpulkan bahwa AI mampu mengenali perintah suara dalam bahasa lokal atau dialek Nusantara dengan tingkat akurasi yang cukup tinggi, terutama jika model dilatih dengan metode yang tepat. Model berbasis wav2vec2 yang telah melalui proses fine-tuning menunjukkan peningkatan yang signifikan dibandingkan model dasar, dengan Word Error Rate (WER) yang berkurang menjadi sekitar ±22% dan Command Accuracy mencapai 81%–86%. Teknik augmentasi data seperti time-stretching, pitch shifting, dan noise injection terbukti meningkatkan performa model, menghasilkan WER yang lebih rendah (±18%) dan Command Accuracy yang lebih baik. Perbedaan fonetik antar dialek, seperti dialek Toraja yang lebih mudah dikenali dibandingkan Bugis dan Makassar, mempengaruhi tingkat kesalahan pengenalan. Uji pengguna menunjukkan bahwa dukungan AI terhadap bahasa lokal/dialek meningkatkan kepercayaan (4.2), kegunaan (4.4), dan kepuasan (4.3) pada skala Likert 1–5, yang menegaskan pentingnya aspek linguistik dan budaya dalam penerimaan teknologi. Secara keseluruhan, penelitian ini berhasil menjawab masalah yang diajukan dengan membuktikan bahwa AI dapat memahami dialek lokal melalui integrasi transfer learning, augmentasi data, dan desain antarmuka yang inklusif.

Daftar Pustaka

[1] L. Besacier, E. Barnard, A. Karpov, and T. Schultz, "Automatic speech recognition for under-resourced languages: A survey," *Speech Commun.*, vol. 56, no. 1, pp. 85–100, 2014, doi: 10.1016/j.specom.2013.07.008.

[2] I. G. A. G. A. Kadyanan *et al.*, "Balinese text-to-speech dataset as digital cultural heritage," *Data Br.*, vol. 60, p. 111528, 2025, doi: 10.1016/j.dib.2025.111528.

[3] B. Speech and T. Group, "Deep Speech 2 : End-to-End Speech Recognition in English and Mandarin Speech Interfaces : A New / Old Input Paradigm," *arXiv:1512.02595*, pp. 1–17, 2016, [Online]. Available: http://www.jmlr.org/proceedings/papers/v48/amodei16.html

[4] K. Kunci, "SYSTEMIC : Information System and Informatics Journal Sistem Pembelajaran Hukum Baca Al- Qur ' an Menggunakan Algoritma LPC dan KNN," vol. 6, no. 1, pp. 29–37, 2020.

[5] S. Wahyuni *et al.*, "Desain Sistem Speech Recognition Penerjemah Bahasa Toraja Menggunakan Hidden Markov Model Design System Speech Recognition Tranlator Toraja Language Using Hidden Markov Modelling," *Jppi*, vol. 11, no. 2, pp. 107–119, 2021, doi: 10.17933/jppi.v11i2.286.

[6] R. Patel and S. Patel, "Deep Learning for Natural Language Processing," *Lect. Notes Networks Syst.*, vol. 190, no. May 2020, pp. 523–533, 2021, doi: 10.1007/978-981-16-0882-7_45.

[7] R. De Mori, "Recent advances in automatic speech recognition," *Signal Processing*, vol. 1, no. 2, pp. 95–123, 1979, doi: 10.1016/0165-1684(79)90013-6.

[8] D. Povey, G. Boulianne, L. Burget, P. Motlicek, and P. Schwarz, "the Kaldi Speech Recognition," no. January, 1920.

[9] A. Baevski, H. Zhou, A. Mohamed, and M. Auli, "wav2vec 2.0: A framework for self-supervised learning of speech representations," *Adv. Neural Inf. Process. Syst.*, vol. 2020-December, no. February, pp. 0–19, 2020, doi: 10.48550/arXiv.2006.11477.

[10] B. R. Setiawan, A. Aranta, and B. Irmawati, "Speech To Text Bahasa Sasak Menggunakan Extraksi Fitur Mel-Frequency Cepstral Coefficients Dan Klasifikasi Convolutional Neural Networks," *J. Teknol. Informasi, Komputer, dan Apl. (JTIKA)*, vol. 5, no. 1, pp.

DOI: https://doi.org/10.31004/riggs.v4i3.3113

21–32, 2023, doi: 10.29303/jtika.v5i1.235.

[11] A. Graves, A. R. Mohamed, and G. Hinton, "Speech recognition with deep recurrent neural networks," *ICASSP, IEEE Int. Conf. Acoust. Speech Signal Process. - Proc.*, no. 3, pp. 6645–6649, 2013, doi: 10.1109/ICASSP.2013.6638947.

[12] H. Y. Lee *et al.*, "Self-supervised Representation Learning for Speech Processing," *NAACL 2022 - 2022 Conf. North Am. Chapter Assoc. Comput. Linguist. Hum. Lang. Technol. Tutor. Abstr.*, pp. 8–13, 2022, doi: 10.18653/v1/2022.naacl-tutorials.2.

[13] Y. Jia, M. T. Ramanovich, Q. Wang, and H. Zen, "CVSS Corpus and Massively Multilingual Speech-to-Speech Translation," *2022 Lang. Resour. Eval. Conf. Lr. 2022*, pp. 6691–6703, 2022.

[14] E. William and A. Zahra, "Speech Recognition Dengan Whisper Dalam Bahasa Indonesia," *Action Res. Lit.*, vol. 9, no. 2, pp. 386–397, 2025, doi: 10.46799/arl.v9i2.2573.

[15] H. Henry and E. Eryc, "Speech Recognition Untuk Membantu Pelafalan Hanyu Pinyin Sebagai Bagian Dari Edukasi Bahasa Mandarin," *J. Ilm. Edutic Pendidik. dan Inform.*, vol. 10, no. 2, pp. 117–128, 2024, doi: 10.21107/edutic.v10i2.22633.

[16] A. Hinsvark *et al.*, "Accented Speech Recognition: A Survey," 2021, [Online]. Available: http://arxiv.org/abs/2104.10747

DOI: https://doi.org/10.31004/riggs.v4i3.3113