Sistem Pengenal Wicara Menggunakan Mel-Frequency Cepstral Coefficient

(Speech Recognition System Using Mel-Frequency Cepstral Coefficient)

KARISMA TRINANDA PUTRA

ABSTRACT

Human-machine interaction evolves toward a more adaptive and interactive system. There are several media that can be used in human-machine interaction systems, such as voice signals. The process includes converting analog signals into the appropriate meaning, which depend on the noise and reliability of signal characteristic extraction methods. In fact, variations of pronunciation by different people will result in a diversity of voice signal patterns. This research develops technology that can recognize and translate speech according to data that has been trained and can be modified based on user requirement. The voice signal will be separated from the silent signal using voice activity detection. Then, the voice signal is converted to the frequency domain before it is extracted using mel-frequency cepstral coefficients. Cepstral value from MFCC extraction will be identified as words using artificial neural network. This study utilizes a computer with a microphone as a sound recording device and pascal programming language as the basis for building applications. Based on the experimental results, the accuracy is 87% on the speech recognition process with 28 vocabulary sets. Accuracy decreases with more sets of vocabulary. However, the more pronounced speech variations, the greater the accuracy with an average number around 93%.

Keywords: voice activity detection, mel-frequency cepstral coefficient, artificial neural network.

PENDAHULUAN

Dalam sistem interaksi manusia dan mesin, ada beberapa media yang dapat digunakan untuk menyampaikan pesan atau perintah. Manusia memiliki beberapa cara untuk berinteraksi seperti dengan kontak mata, bahasa tubuh dan wicara. Cara yang paling sering digunakan oleh manusia adalah interaksi dengan wicara. Manusia memiliki aset kompleks untuk berbicara, meliputi organ reproduksi suara, organ pendengaran dan otak sebagai organ pemroses informasi. *Neuropsychiatrist* melaporkan bahwa manusia menggunakan sekitar 6.000 – 20.000 kata setiap hari (Mehl et al., 2007). Itulah mengapa wicara sangatlah penting bagi manusia. Dengan wicara, informasi dapat disampaikan dengan lebih detail. Dengan banyaknya variasi bahasa menyebabkan penerjemahan makna dari sinyal wicara menjadi semakin kompleks. Belum lagi masalah akurasi yang rendah karena sinyal suara rentan terhadap *noise*. Penelitian ini bertujuan untuk mengembangkan teknologi kecerdasan buatan yang dapat mengenali dan menerjemahkan wicara manusia.

Sinyal suara adalah suatu sinyal analog yang membutuhkan pemrosesan lebih lanjut sebelum dapat dipahami oleh mesin. Pada mulanya, sinyal suara dipisahkan dari sinyal *silent* (jeda) menggunakan *voice activity detection* (VAD). Kemudian ciri sinyal diekstrak menggunakan *mel-frequency cepstral coefficients* (MFCC). Sistem ini mengemulasikan sistem pendengaran manusia dengan menganalisis spektrum frekuensi ke dalam beberapa *spectral filter* tertentu. *Artificial neural network* (ANN) akan digunakan dalam proses identifikasi *cepstral* yang dihasilkan. ANN adalah sekelompok jaringan unit pemroses yang dapat memodelkan sesuatu berdasarkan sistem jaringan syaraf manusia. ANN mampu untuk memproses data statistik non-linier. ANN akan memberikan penilaian statistik tingkat kecocokan sinyal suara dengan kata-kata yang dilatihkan.

Penelitian ini akan berkontribusi terhadap pengembangan sistem interaksi manusia-mesin. Sistem interaksi manusia-mesin diharapkan

berkembang menuju sistem interaksi yang lebih alami dan adaptif. Artinya mesin dapat memodelkan sinyal wicara berdasarkan ciri khusus yang dimiliki setiap individu sehingga diharapkan interaksi yang terbentuk akan mendekati sistem interaksi antar manusia.

STUDI LITERATUR

Mesin harus memiliki antarmuka penggunaan yang memudahkan manusia untuk berinteraksi dengannya. Antarmuka adalah sebuah sarana dimana interaksi antara pengguna dan mesin terjadi. Secara alami, interaksi manusia terjadi melalui beberapa cara seperti bahasa isyarat, kontak mata dan wicara.

Bahasa isyarat adalah gerakan satu atau beberapa bagian tubuh yang berguna untuk menyampaikan informasi tertentu. Untuk menghadirkan *interface* yang atraktif, sebuah robot dilengkapi dengan kamera untuk mendeteksi pergerakan tangan dan isyarat dari pengguna (Fardana et al., 2013). Secara visual, penerjemahan perintah membutuhkan penanganan khusus karena teknologi ini sangat tergantung dengan pencahayaan, kedalaman gambar dan deteksi obyek (Olson et al., 2010). Dalam kasus tertentu seperti penerjemahan perintah kompleks, sistem komunikasi dengan bahasa isyarat akan menemui keterbatasannya. Penginderaan visual hanya mengijinkan robot untuk memahami sebuah perintah sederhana dan sulit diaplikasikan dalam sistem komunikasi yang kompleks. Sedangkan sistem interaksi menggunakan bahasa tubuh memiliki keterbatasan dalam menerjemahkan perintah, yang membutuhkan detail informasi tambahan.

Kontak mata adalah salah satu bentuk dari komunikasi non-verbal yang disebut *oculesic*. Penelitian yang telah dilakukan oleh Purwanto et al. (2009), menggunakan kedipan mata untuk mengontrol pergerakan kursi roda. Antarmuka jenis ini sangat cocok digunakan untuk menerjemahkan perintah sederhana yang membutuhkan respon cepat. Penggunaan kontak mata sebagai antarmuka memiliki kelemahan terutama pada akurasi yang sangat terpengaruh oleh kondisi lingkungan (Damaryam & Dunbar, 2005).

Antarmuka berbasis wicara mengijinkan mesin untuk memahami perintah yang diberikan oleh pengguna menggunakan perintah suara. Antarmuka ini dapat mengakomodasi perintah sederhana maupun kompleks. Sistem ini menggunakan sistem penerjemah wicara. Sistem penerjemah wicara akan memproses sinyal suara menjadi data dan menerjemahkannya menjadi makna kata-kata yang sesuai. Pengenalan wicara memiliki banyak variasi penggunaan sebagai contoh pengendalian robot mobil (Jangmyung & MinCheol, 2013) maupun penugasan robot industri (Teller et al., 2010).

Secara umum, sistem penerjemah wicara dibagi kedalam 2 proses, meliputi ekstraksi ciri dan pengenalan pola. Salah satu sarana identifikasi untuk merepresentasikan ciri sinyal wicara adalah dengan meggunakan *cepstral*. Reprentasi *cepstral* dari spektrum wicara memberikan reprentasi dari sifat-sifat spektral lokal sinyal untuk analisis *frame* yang diketahui. *Mel-frequency cepstral coefficients* menjadi kandidat untuk analisis spectral (Kumar et al., 2010). Sedangkan pada proses pengenalan pola *cepstral*, akan digunakan *neural network*. ANN menghasilkan akurasi pengenalan lebih baik dibandingkan metode yang telah ada sebelumnya (Waibel et al., 1988).

METODE

Sistem yang diusulkan terdiri dari beberapa fase pemrosesan sinyal. Proses pengenalan wicara terdiri dari *voice activity detection, pre-emphasis filtering, frame blocking, windowing, fast fourier transform, mel-frequency cepstral coefficients* dan *pattern identification* menggunakan *neural network*.

Voice Activity Detection (VAD)

Sinyal suara mentah diproses menggunakan *voice activity detection* untuk memisahkan sinyal suara dengan sinyal jeda. Setiap sinyal yang merepresentasikan kata diproses lebih lanjut sedangkan sinyal jeda akan dihapus. Untuk memisahkan kondisi sinyal wicara, digunakan perhitungan power sinyal dan *zero crossing rate*. Power sinyal melambangkan seberapa kuat sinyal dalam satuan waktu tertentu sedangkan *zero crossing* rate melambangkan seberapa sering sinyal suara melewati titik nol dalam satuan waktu tertentu.

Pre-emphasize Filtering

Pre-emphasize adalah salah satu tipe filter yang digunakan sebelum sinyal suara diproses lebih lanjut. Filter ini diaplikasikan pada frekuensi tinggi yang secara umum mengalami

pelemahan selama proses produksi suara. Tujuannya adalah untuk mengkompensasi bagian frekuensi tinggi yang terkompresi selama mekanisme produksi suara manusia dan menguatkan frekuensi tinggi yang masih diperlukan untuk pengolahan selanjutnya.

Frame Blocking

Sinyal harus diproses dalam satuan waktu tertentu (*short frame*), karena sinyal suara terus berubah sebagai hasil dari pergeseran artikulasi dari organ reproduksi suara. Panjang *frame* yang digunakan adalah sekitar 25 milidetik. Di satu sisi, ukuran dari *frame* harus sepanjang mungkin untuk dapat menunjukkan resolusi frekuensi yang baik. Tetapi di lain sisi, ukuran *frame* juga harus cukup pendek untuk dapat menunjukkan resolusi waktu yang baik. Proses *frame blocking* ini dilakukan terus sampai seluruh sinyal dapat diproses. Selain itu, proses ini dilakukan secara *overlapping* untuk setiap *frame*-nya. Panjang daerah *overlap* yang digunakan adalah 30% dari panjang *frame*. *Overlapping* dilakukan untuk menghindari hilangnya ciri atau karakteristik suara pada perbatasan perpotongan setiap *frame*.

Windowing

Proses *framing* dapat menyebabkan terjadinya *magnitude leakage* (kebocoran spektral) atau *aliasing*. *Aliasing* adalah sinyal baru yang memiliki frekuensi yang berbeda dengan sinyal aslinya. Efek ini dapat terjadi karena rendahnya jumlah *sampling rate*, ataupun karena proses *frame blocking* sehingga menyebabkan sinyal menjadi *discontinue*. Untuk mengurangi kemungkinan terjadinya kebocoran spektral, maka hasil dari *frame blocking* harus melewati proses *windowing*. Sebuah fungsi *window* yang baik harus menyempit pada bagian *main lobe* dan melebar pada bagian *side lobe*-nya.

Fast Fourier Transform (FFT)

Analisis *fourier* adalah metode yang digunakan untuk menganalisa komponen level sinyal dalam domain frekuensi. Representasi level frekuensi biasa disebut *spectrogram*. Pada *spectogram*, ada hubungan yang sangat dekat antara waktu dan frekuensi. Hubungan antara frekuensi dan waktu berbanding terbalik dengan hubungan proporsional. Ketika menggunakan resolusi waktu yang tinggi, resolusi frekuensi yang dihasilkan akan lebih rendah. *Fast fourier transform* adalah solusi yang dapat digunakan dalam proses analisis sinyal.

Mel-Frequency Cepstral Coefficients

Frequency wrapping umumnya dilakukan dengan menggunakan *filter bank*. *Filter bank* adalah salah satu bentuk dari *filter* yang digunakan dengan tujuan untuk mengetahui ukuran energi dari frekuensi pada *band* tertentu dalam sinyal suara. *Filter bank* diterapkan pada domain domain frekuensi. *Filter* ini berjumlah 24 *channel* yang dilakukan secara linear terhadap frekuensi 0-4 kHz. *Filter bank* menggunakan representasi dengan mengalikan antara spektrum sinyal dengan *koefisien filter bank*. Selanjutnya *discrete cosinus transform* diaplikasikan pada hasil *filter wrapping.*

Pattern Identification

Neural network digunakan dalam proses identifikasi pola. Masukan ANN berupa koefisien *cepstral* sedangkan keluarannya berupa kata yang terkait. Pada tahap ini, kosakata yang sesuai dilatihkan menggunakan *back-propagation*. Nilai keluaran tertinggi akan merepresentasikan kosakata yang sesuai dengan sinyal input. ANN yang digunakan memiliki 2 *hidden layer* dan fungsi aktivasi *sigmoid* untuk setiap *neuron*-nya.

Proses pelatihan adalah proses penambahan pengetahuan pada ANN sehingga pengetahuan tersebut dapat digunakan secara optimal dalam proses pengenalan obyek. Pada proses ini, nilai parameter pembobot akan diperbarui secara kontinyu sehingga keluaran akan mengikuti nilai keluaran target yang dilatihkan.

HASIL PENGUJIAN

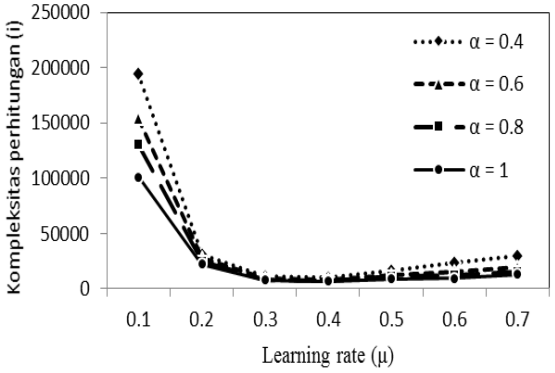
Setiap variabel *artificial neural network* sangat berpengaruh terhadap kompleksitas perhitungan. Kompleksitas perhitungan semakin bertambah dengan semakin kecilnya konstanta pembelajaran (μ). Nilai μ melambangkan kecepatan pembelajaran *neuron* dalam memperbarui nilai *weight* dan *bias* agar sesuai dengan target yang ingin dicapai. Melalui percobaan yang telah dilakukan, nilai μ yang optimal adalah sekitar 0,3. Semakin besar nilai μ tidak berpengaruh banyak terhadap kompleksitas perhitungan. Akan tetapi nilai μ yang lebih kecil dapat menambah kompleksitas perhitungan secara dramatis. Konstanta kemiringan kurva (α) juga mempengaruhi kompleksitas perhitungan. Semakin rendah nilai α, maka kompleksitas perhitungan semakin bertambah. Kompleksitas perhitungan akan mempengaruhi lama proses pembelajaran. Dengan perhitungan yang lebih efisien,

resource komputer dapat lebih dihemat dan pekerjaan dapat diselesaikan lebih cepat.

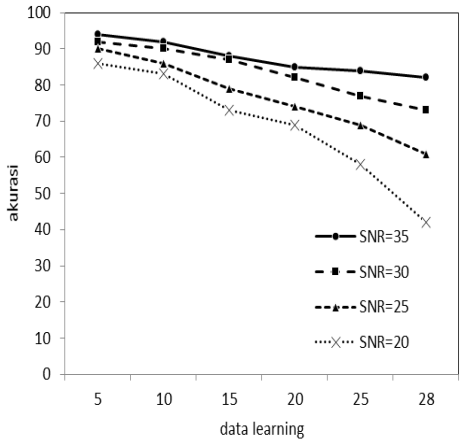
Pada pengujian akurasi proses klasifikasi, didapat data bahwa semakin kecil nilai beta yang diberikan, semakin tinggi akurasinya. akurasi mulai membaik dengan nilai $\beta \leq 10\text{-}5$. Perlu diingat bahwa nilai β yang semakin kecil akan meningkatkan kompleksitas perhitungan sahingga waktu pembelajaran akan semakin panjang. Nilai β yang terlalu kecil malah akan memperlambat proses pembelajaran untuk mencapai konvergen.

Selanjutnya dilakukan proses pengujian akurasi pengenalan wicara. Tujuan dari pengujian ini adalah menguji pengaruh banyaknya variasi kata dan *noise* terhadap akurasi yang dihasilkan. Pengujian dilakukan dengan 2 *hidden layer* yang masing-masing *layer* memiliki jumlah *neuron* sesuai dengan *output neuron*. Sesuai dengan Gambar 2, semakin bertambahnya variasi data pembelajaran, maka akurasi *neural network* semakin menurun. Dengan semakin banyaknya sampel pembelajaran, proses konvergensi akan semakin lama dan ada kemungkinan proses tersebut gagal. Kegagalan biasa diakibatkan kombinasi *weight* dan *bias* ANN yang tidak dapat mengakomodasi variasi pola yang diberikan. Memperbesar jumlah *neuron* dapat menjadi alternatif agar proses pembelajaran dapat mencapai konvergen dengan konsekuensi kompleksitas perhitungan semakin meningkat.

Kegagalan penerjemahan juga bisa disebabkan pengucapan kata yang mirip, misalkan kata ‘maju’ dan ‘laju’. Karena tidak termasuk ke dalam pengucapan 28 kosakata yang telah ditentukan sebelumnya, maka kata ‘laju’ maupun kata lain yang mirip dalam hal pengucapan harus dilatihkan dengan target keluaran bukan merupakan salah satu dari kosakata yang dimaksud. Perlu dilakukan proses pelatihan yang lebih banyak untuk kata-kata dengan pengucapan yang mirip sehingga nantinya ANN mampu menghasilkan diferensiasi yang akurat.

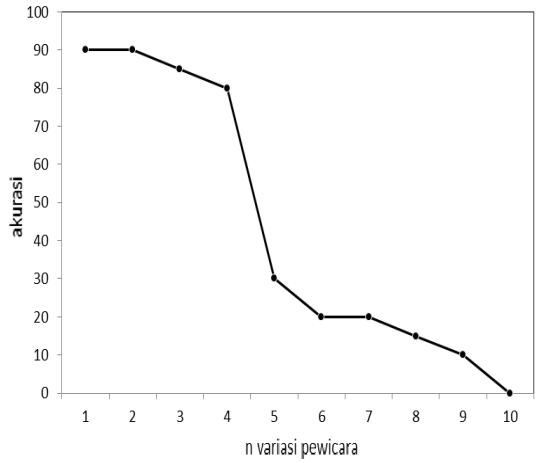


GAMBAR 1. Pengaruh variabel (kemiringan kurva *sigmoid* dan *lerning rate*) dalam topologi ANN

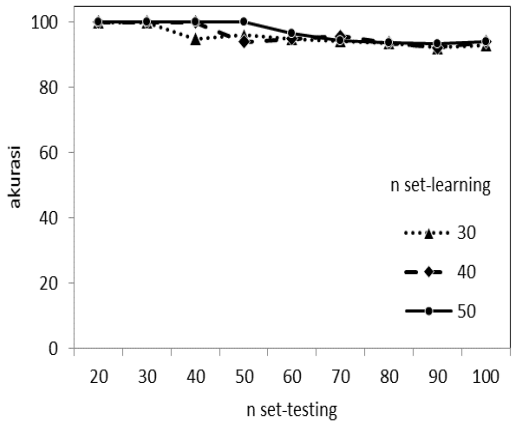


GAMBAR 2. Pengaruh *noise* terhadap akurasi pengenalan sinyal wicara

Pada pengujian selanjutnya, dilihat pengaruh variasi SNR terhadap akurasi identifikasi pola sinyal wicara. SNR didapat dengan mengurangkan power rata-rata sinyal wicara dengan power rata-rata sinyal *noise*. Semakin tinggi SNR, akurasi pengenalan pola juga akan meningkat. SNR yang semakin besar megindikasikan bahwa perbedaan sinyal wicara dan sinyal *noise* semakin lebar. Hal ini mempengaruhi akurasi VAD dalam proses pemotongan sinyal wicara. Nilai SNR yang besar akan mempermudah batasan yang diberikan untuk membedakan pola sinyal wicara dengan jeda. Sebaliknya SNR yang rendah akan mengakibatkan data yang diproses VAD menjadi tidak akurat. Akibatnya ANN banyak mengalami kesalahan dalam proses identifikasi. Untuk meningkatkan akurasi, data pembelajaran dapat ditingkatkan dengan penambahan data pola wicara. Akurasi tertinggi didapat dengan SNR 35 dB sebesar 82 %.



GAMBAR 3. Pengaruh variasi pewicara terhadap akurasi pengenalan sinyal wicara



GAMBAR 4. Hubungan banyaknya variasi kosakata dan data set pembelajaran terhadap akurasi

Akurasi sangat ditentukan oleh banyaknya variasi keluaran yang diharapkan. Semakin tinggi variasi obyek klasifikasi, maka semakin rendah akurasi yang dihasilkan sesuai dengan Gambar 3. Setiap kata yang diucapkan oleh pewicara yang berbeda memiliki ciri khusus yang berbeda satu sama lain. Hal ini sangat bergantung pada warna suara masing-masing individu. Akan tetapi pada dasarnya, MFCC dan ANN dapat digunakan untuk mempelajari pola data untuk pewicara beserta sinyal wicaranya. Hanya saja, akurasi ANN akan menurun seiring dengan semakin banyaknya pola data yang mampu untuk diidentifikasi.

Akurasi sangat menurun pada pengujian dengan variasi lima pewicara atau lebih. Hal ini dapat disebabkan karena sinyal wicara untuk satu kata bisa dikatakan memiliki ciri yang hampir mirip. Untuk mempelajari pola warna suara masing-masing individu, perlu dilakukan pengambilan sampel sinyal yang lebih banyak dan kompleks. Hal ini yang menyebabkan sinyal wicara satu kata yang singkat tidak dapat dijadikan acuan untuk mengidentifikasi identitas pewicara. Pengenalan pewicara dapat dilakukan dengan sampel data yang lebih kompleks dan detail.

Pada Gambar 4, akurasi pengenalan kata semakin menurun sesuai dengan semakin banyaknya kosakata yang diujikan. Dengan menambah sampel pembelajaran, akurasi dapat meningkat walaupun tidak signifikan. Akurasi diapat sekitar 93 % untuk 100 kosakata dengan 50 data pembelajaran. Dengan jumlah data pembelajaran sebanyak ini, ANN memiliki kemungkinan untuk sulit mencapai konvergen dikarenakan menipisnya jumlah kombinasi *weight* yang dapat digunakan dalam mempelajari pola. Untuk mengatasi hal ini, maka jumlah *neuron* dapat ditambah dengan konsekuensi kompleksitas perhitungan semakin meningkat.

KESIMPULAN

Pada penelitian ini, telah dirancang *voice activity detection* dengan akurasi sekitar ≥ 85% untuk SNR ≥ 25. Keberhasilan *voice activity detection* dalam menyeleksi sinyal wicara akan menentukan keberhasilan proses pengenalan wicara. Pada proses pengenalan pola ciri sinyal wicara, didapat nilai koefisien $\alpha = 1$, $\mu = 0.4$ and $\beta = 10\text{-}5$ sebagai nilai parameter *neural network*. Nilai ini dapat menghasilkan kombinasi kompleksitas pemrosesan *neural*

network yang paling rendah dan akurasi yang tinggi. Sistem ekstraksi ciri berbasis *mel-frequency cepstral coefficient* dikombinasikan dengan identifikasi pola berbasis *neural network* menghasilkan akurasi 87 % untuk 28 kosakata yang diujikan. Akurasi akan semakin menurun dengan semakin tingginya SNR maupun semakin banyaknya pola kosakata yang dipelajari. Dengan menambah jumlah data set pembelajaran, maka akurasi dapat meningkat menjadi 93% (50 data set).

DAFTAR PUSTAKA

Chen, D., and Manning, C. D. (2014), A Fast and Accurate Dependency Parser using Neural networks. Proceedings of the 2014 conference on Empirical Methods in Natural language processing (EMNLP).

Damaryam, G., Dunbar, G. (2005). A Mobile Robot Vision System for Self navigation using the Hough Transform and neural networks. Proceedings of the EOS Conference on Industrial Imaging and Machine Vision, Munich, pp. 72.

Fardana, A.R., Jain, S., Jovancevic, I., Suri, Y., Morand, C. and Robertson, N.M. (2013). Controlling a Mobile Robot with Natural Commands based on Voice and Gesture. Proceedings of the IEEE International Conference on Robotics and Automation (ICRA).

Jangmyung, L., MinCheol, L. (2013). A Robust Control of Intelligent Mobile Robot Based on Voice Command. Proceedings of the 6th International Conference, ICIRA.

Kumar, P., Biswas, A., Mishra, A .N., and Chandra, M. (2010). Spoken Language Identification Using Hybrid Feature Extraction Methods. Journal of Telecommunications. Volume 1. Issue 2.

Mehl, M. R., Vazire, S., Ramírez-Esparza, N., Slatcher, R. B., Pennebaker, J. W. (2007). Are women really more talkative than men. Science 317 (5834), 82-82.

Olson, C. F., Matthies, L. H., Schoppers, M., Maimone, M. W. (2010). Rover navigation using stereo ego-motion. Robotics and Autonomous Systems 43 (4): page 215–229.

Purwanto, D., Mardiyanto, R., Arai, K. (2009). Electric wheelchair control with gaze direction and eye blinking. Proceedings of the 14th International Symposium on Artificial Life and Robotics, Oita, Japan.

Socher, R., Bauer, J., Manning, C., D., Yan-Tak Ng., A. (2013). Parsing with compositional vector grammars, Proceedings of the ACL conference.

Teller, S., Walter, M. R., Antone, M., Correa, A., Davis, R., Fletcher, L., Frazzoli, E., Glass, J., How, J. P., Huang, A. S., Jeon, J. H., Karaman, S., Luders, B., Roy, N., Sainath, T. (2010). A Voice-Commandable Robotic Forklift Working Alongside Humans in Minimally-Prepared Outdoor Environments. Proceedings of the Robotics and Automation (ICRA).

Waibel, A., Hanazawa, T., Hinton, G., Shikano, K., Lang, K. (1988). Phoneme Recognition: Neural Networks vs Hidden Markov Models. Proceedings of the International Conference on Acoustics, Speech, and Signal Processing (ICASSP).

PENULIS:

Karisma Trinanda Putra

Prodi Teknik Elektro, Fakultas Teknik, Universitas Muhammadiyah Yogyakarta, Yogyakarta, Jalan Lingkar Selatan, Tamantirto, Kasihan, Bantul, Yogyakarta 55183.

Email: karisma@ft.umy.ac.id