Journal of Data Analysis
Vol.2, No.2, December 2019, p. 61-70

Penerapan *Time Delay Neural Network* pada Model Akustik untuk Sistem *Voice-to-Text* Berbahasa Sunda

Alim Misbullah[1*], Nazaruddin[1], Marzuki[2] dan Zulfan[1]

[1] Jurusan Informatika, Fakultas MIPA, Universitas Syiah, Banda Aceh, Indonesia
[2]Jurusan Statistika, Fakultas MIPA, Universitas Syiah Kuala, Banda Aceh, Indonesia
E-mail: misbullah@unsyiah.ac.id*, anzaro@unsyiah.ac.id, marzuki@unsyiah.ac.id, zulfan.abdullah@unsyiah.ac.id

* = corresponding author

Abstrak

Penerapan metode *deep learning* dalam berbagai bidang terutama pada kasus pengenalan pola sudah menghasilkan akurasi yang sangat menjanjikan. Jaringan saraf tiruan atau *neural network* merupakan bagian dari *deep learning* yang digunakan untuk melatih model pada kasus pengenalan pola seperti model untuk sistem pengenalan ucapan (*voice-to-text*). *Neural network* akan menyimpan informasi dari setiap fitur data berupa bobot pada jaringan yang terhubung antar-*layer* pada model yang dibangun. Bobot pada jaringan tersebut diperbaharui berdasarkan banyaknya fitur dari data yang diinput. Sistem *voice-to-text* merupakan salah satu bidang pengenalan pola yang mengimplementasikan *neural network* untuk membangun model akustik. Model akustik pada sistem pengenalan ucapan dilatih menggunakan data audio berupa percakapan atau rekaman dari setiap individu untuk bahasa tertentu seperti bahasa Inggris. Penerapan *neural network* untuk sistem pengenalan ucapan berbahasa Inggris sudah banyak dilakukan bahkan sudah diimplementasikan dalam bentuk aplikasi karena mampu menghasilkan akurasi yang tinggi. Namun, penggunaan *neural network* untuk bahasa lokal masih jarang digunakan. Dalam tulisan ini, *time delay neural network* digunakan untuk membangun model akustik pada sistem pengenalan ucapan berbahasa Sunda. Berdasarkan hasil pengujian terhadap model akustik, *time delay neural network* mampu menghasilkan *WER* sampai dengan 0.57% setelah dilakukan penyesuaian pada *hyperparameter* dari *neural network*.

Informasi Artikel

Sejarah Artikel:
Diajukan, 19 Des 2019
Diterima, 2 Apr 2020

Kata Kunci:
Deep Learning,
Neural Network,
Model Akustik,
Sistem Pengenalan Ucapan

Abstract

Implementation of deep learning techniques has given promising results recently in any research area, especially for pattern recognition. Neural network as a part of deep learning has been widely used to build model for various pattern recognition field including speech recognition. In neural network, weights which is parameters among layers play important roles to capture information from input data. The parameters are updated frequently based on input features in each iteration. In speech recognition, neural network is implemented to build acoustic model that uses speech from different speakers as training data. The acoustic model is built for specific language such as English, Mandarin and Indonesian. In recent years, the speech recognition

Keyword:
Deep Learning
Neural Network
Acoustic Model
Speech Recognition System

system using deep neural network for English language has been developed well and use in many applications. But, implementation of deep neural network for local language is rarely done. In this research, time delay neural network is used to build acoustic model for speech recognition system of Sundanese language. Based on experimental result, the implementation of time delay neural network can reduce WER to be 0.57% with well-tuned hyperparameters of neural network.

1. Pendahuluan

Jaringan saraf tiruan (*neural network*) saat ini telah memberikan hasil yang menjanjikan pada berbagai bidang teknologi terutama untuk kasus pengenalan pola seperti sistem pengenalan ucapan (*voice-to-text*). Umumnya, struktur *neural network* terdiri dari *input layer*, *hidden layer* dan *output layer*. Setiap *node* pada *input layer* merepresentasikan vector untuk jumlah data input, sedangkan setiap *node* pada *hidden layer* berfungsi untuk mengontrol dapat atau tidaknya informasi dari *input layer* diteruskan ke *layer* berikutnya. Pada *output layer*, setiap *node* didefinisikan sebagai target dari kelas yang akan diprediksi.

Sistem *voice-to-text* memiliki 2 (dua) komponen utama untuk membangun sistem yaitu model akustik (*acoustic model*) dan model bahasa (*language model*). Model akustik merupakan sebuah model yang mengandung representasi statistik dari setiap *voice* yang berbeda dalam membentuk sebuah kata atau kalimat. Setiap representasi statistik tersebut menentukan sebuah label atau target yang disebut dengan suku kata (*phonemes*).

Struktur *neural network* pada model akustik digunakan untuk melatih model sehingga mampu menyimpan informasi dari setiap fitur data *voice* yang diinput. Struktur *neural network* menggunakan pendekatan *learning based* dalam melatih model akustik yang berarti bahwa setiap fitur akan melalui setiap *hidden layer* sehingga informasi yang ada di dalam fitur akan tercatat pada setiap jaringan yang terhubung antar-*layer* yang disebut dengan parameter (*weight*). Selanjutnya, parameter tersebut akan diperbaharui berdasarkan *error* yang didapat antara nilai keluaran (*output*) dan nilai target *(class)*.

Proses melatih model akustik menggunakan *neural network* memiliki 2 (dua) tahapan utama yaitu *feedforward* dan *backpropagation*. Pada tahapan *feedforward*, setiap data input berupa vektor akan melalui setiap *hidden layer* dalam *neural network* sampai dengan *output layer*. Setiap *node* pada *hidden layer* akan menjadi aktif atau tidak berdasarkan transformasi hasil perkalian linier antara vector input dan parameter matriks yang ada pada jaringan. Proses transformasi hasil perkalian tersebut menggunakan sebuah fungsi yang disebut sebagai fungsi aktifasi (*activation function*). Sedangkan pada tahap *backpropagation*, parameter matriks akan diperbaharui berdasarkan selisih antara nilai *output* dan nilai target yang ada yang disebut dengan nilai *error*. Nilai error ini menjadi acuan awal untuk menentukan nilai parameter baru pada setiap jaringan.

Penggunaan *neural network* untuk membangun model pada sistem *voice-to-text* sudah banyak ditemukan terutama untuk sistem *voice-to-text* berbahasa Inggris [1, 2, 3, 4]. Bahkan sistem *voice-to-text* berbahasa Inggris sudah banyak digunakan pada aplikasi dalam kehidupan seperti Google Asisten, Microfost Cortana, dan Alexa Amazon. Sedangkan, penelitian sistem *voice-to-text* untuk bahasa lokal seperti bahasa Sunda masih jarang dilakukan karena keterbatasan data *voice* yang tersedia dan bahkan akurasinya masih perlu ditingkatkan.

Uraian di atas mengantarkan peneliti untuk mengimplementasikan *neural network* untuk membangun model akustik pada sistem *voice-to-text* berbahasa Sunda sehingga dapat diperoleh akurasi yang lebih baik dari hasil penelitian yang telah ada sebelumnya. Hasil penelitian ini dapat digunakan sebagai *prototipe* awal dalam membangun aplikasi *voice-to-text* berbahasa Sunda untuk berbagai perangkat.

2. Tinjauan Kepustakaan

Penelitian-penelitian *voice-to-text* untuk bahasa Sunda sudah pernah dilakukan sebelumnya. Sakti dkk [14] mengembangkan pengenalan ucapan berbasis *Grapheme* terhadap bahasa Jawa, Sunda, Bali, dan Batak. Tujuan akhir dari penelitian-penelitian tersebut adalah ingin membuat alat penerjemah ucapan dari bahasa daerah ke bahasa Inggris dan bahasa Indonesia. Hasil penelitian menunjukkan bahwa dengan menggunakan *multilingual acoustic model*, performa dapat meningkatkan pada aksen bahasa Indonesia, tetapi tidak untuk bahasa daerah. Ini menunjukkan bahwa karakteristik akustik bahasa Indonesia cukup berbeda dari karakteristik akustik dari bahasa-bahasa daerah tersebut.

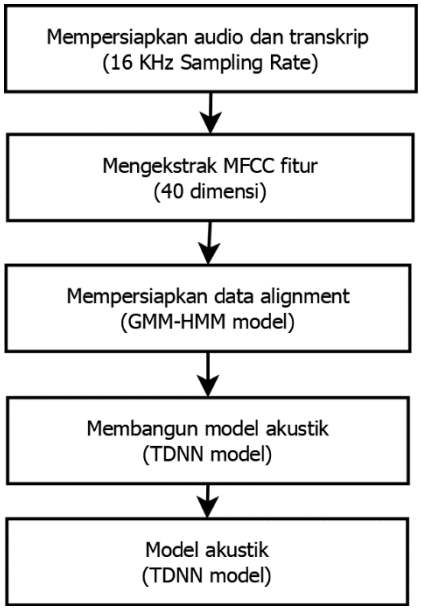
Rahmawati dkk [14] melakukan penelitian yang bertujuan untuk membandingkan fitur dan teknik pemodelan terbaik untuk mengenali dialek bahasa Jawa dan Sunda pada bahasa Indonesia menggunakan MFCC dan kombinasi dari MFCC + pitch. Selain itu juga ingin membandingkan hasil teknik pemodelan dengan menggunakan GMM dan *i-vector*. Berdasarkan hasil penelitian tersebut dapat disimpulkan bahwa fitur terbaik untuk mengenali dialek bahasa Jawa dan Sunda dari ucapan bahasa Indonesia adalah kombinasi dari fitur MFCC + *pitch* dengan teknik pemodelan terbaik adalah *i-vector*.

Teknik *Deep Neural Network* juga pernah diuji coba untuk pengenalan ucapan bahasa Sunda dialek Utara [15], bahasa Sunda dialek Tengah Timur [16] dan bahasa Sunda dialek Garut [17]. Namun dataset yang mereka uji jumlahnya sangat kecil sehingga tidak dapat ditarik kesimpulan yang pasti dari kinerja *Deep Neural Network* terhadap bahasa Sunda.

Kjartansson dkk [18] yang tergabung dalam *Google Research* telah menyediakan dataset untuk 5 (lima) bahasa daerah yaitu bahasa Jawa, Sunda, Sinhala, Nepal, dan Bengali Bangladesh. Setiap bahasa terdiri dari rata-rata sekitar 200.000 ucapan yang direkam oleh sukarelawan penutur asli *(native*) di wilayah masing-masing. Dataset tersebut diklaim sangat cocok untuk penelitian pemodelan akustik pada sistem pengenalan ucapan. Dataset tersebut diizinkan untuk digunakan oleh para peneliti untuk mengembangkan sistem pengenalan ucapan untuk bahasa-bahasa tersebut. Akurasi untuk bahasa Sunda dalam penelitian ini masih sebatas pada model GMM-HMM saja sedangkan untuk melatih model akustik, tidak dilanjutkan menggunakan *Deep Neural Netwok* sehingga performa model masih sangat mungkin untuk ditingkatkan. Model akustik dibangun menggunakan *Time Delay Deep Neural Networks* untuk dataset berbahasa Sunda [18]. Struktur *Time Delay Neural Network* diuji pada jumlah *node* berbeda dari setiap *layer*.

3. Metode Penelitian

Penelitian ini dibagi ke dalam 2 (dua) tahapan yaitu pelatihan dan pengujian model. Tahapan pertama adalah melatih model akustik untuk sistem *voice-to-text* berbahasa Sunda menggunakan *time delay neural network* disingkat dengan TDNN [4]. Model akustik akan dibangun menggunakan *framework* untuk sistem *voice-to-text* yaitu Kaldi toolkit [6]. Proses membangun model akustik dimulai dengan menyiapkan data *voice* dan transkrip, mengekstrak fitur, melakukan *alignment* dengan model GMM-HMM [5], dan membangun model dengan TDNN. Gambar 1 menunjukkan tahapan pertama yang dilakukan untuk membangun model akustik dalam penelitian ini.



Gambar 1 Tahapan membangun model akustik

Pembangunan model akustik untuk sistem *voice-to-text* terdiri dari beberapa tahapan dengan penjelasan sebagai berikut:

1. Mempersiapkan audio dan transkrip
   Data audio dan transkrip dipersiapkan dalam format *wav* dengan *sampling rate* 16000 Hz dan memiliki *mono channel*. Selanjutnya, dataset tersebut disesuaikan formatnya dengan Kaldi *toolkit* sehingga akan mudah dalam proses di tahapan selanjutnya.
2. Mengekstrak MFCC fitur
   Umumnya, fitur yang digunakan untuk melatih model akustik pada sistem *voice-to-text* adalah MFCC (*Mel-Frequency Cepstral Coefficient*) [7]. Fitur MFCC tersebut mudah digunakan dan mampu meningkatkan akurasi untuk *voice based* sistem [8, 9].
3. Mempersiapkan data alignment
   Data alignment adalah proses pemetaan *voice* dari setiap *frame* direntan waktu tertentu terhadap suku katanya. Proses tersebut dilakukan menggunakan GMM-HMM (*Gaussian Mixture Model – Hidden Markov* Model) model [10] yang juga dibangun menggunakan data *voice* yang memiliki transkrip.
4. Membangun model akustik
   Model akustik dapat dibangun menggunakan GMM-HMM model, namun akurasi yang dihasilkan tidak sebaik model yang dibangun menggunakan *neural network*. Pada tahapan ini, model akustik akan dibangun menggunakan *time delay neural network* (TDNN) [4]. Suku kata yang diperoleh pada tahapan sebelumnya akan menjadi target untuk setiap input fitur yang dimasukkan ke dalam struktur TDNN.

3.1. Persiapan Voice Dataset

Model akustik dan model bahasa dibangun menggunakan data *voice* berbahasa Sunda[1]. Data tersebut merupakan bagian dari projek Google Inc. yang dikumpulkan menggunakan *tool* yang disebut *DataHound* [12]. Tabel 1 menunjukkan informasi mengenai data *voice* berbahasa Sunda yang digunakan dalam penelitian ini.

[1] https://openslr.org/36/

Tabel 1 Distribusi data *voice* berbahasa Sunda

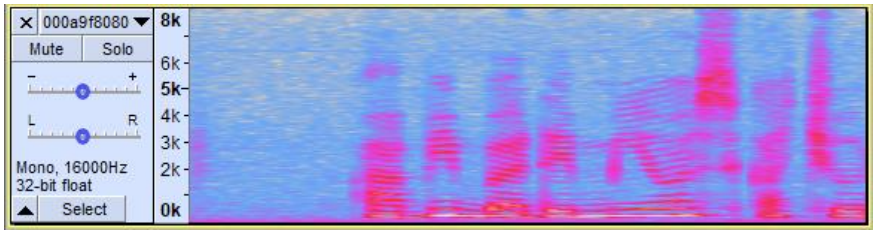
Dataset	# *Speaker*	# *Utterances*	Total Waktu (*hours*)
Training	537	216112	327
Tessting	6	1994	3,2
Total	543	218106	329,2

Selanjutnya, data *voice* tersebut dikonversi ke dalam format yang dapat digunakan oleh Kaldi. Format yang digunakan oleh Kaldi memuat informasi yang terdiri dari *text*, *wav.scp*, *utt2spk*, dan *spk2utt*. Isi dari setiap *file* yang diperlukan oleh Kaldi untuk membangun model akustik ditunjukkan di dalam Tabel 2. Setiap *file* terdiri dari 2 (dua) kolom yang kolom pertamanya mendefiniskan *utterance_id* untuk setiap ucapan. Bagian kolom pertama tersebut akan membentuk relasi antar *file* yang berikutnya digunakan untuk membangun model akustik.

Tabel 2 Format file untuk Kaldi *toolkit*

Format File	Contoh
text	`01f12-0cda366937 wartawan keur nyiar berita di situ cileunca` `01f12-20cbb934dd yana julio olohok ningali marc anthony keur diwawancara`
wav.scp	`01f12-0cda366937 flac -cds /asr_sundanese/data/0c/0cda366937.flac |` `01f12-20cbb934dd flac -cds /asr_sundanese/data/20/20cbb934dd.flac |`
utt2spk	`01f12-0cda366937 01f12` `01f12-20cbb934dd 01f12`
spk2utt	`01f12 01f12-0cda366937 01f12-20cbb934dd`

Setiap file *audio* disimpan dalam format *flac* sehingga memerlukan *library* khusus untuk membacanya. File *audio* tersebut memiliki *sampling rate* 16000 kHz dan *mono channel*. Gambar 2 menunjukkan *spectrogram* untuk salah satu data *audio* berbahasa Sunda. Bagian berwarna merah pada Gambar 1 merepresentasikan energi yang dihasilkan dari ucapan dan energi tersebut mengandung kata pada rentang waktu tertentu.



Gambar 2 Contoh *spectrogram* data *audio* berbahasa Sunda

3.2 Pembuatan Kamus dan Leksikal

Proses membangun model akustik merupakan bagian dari proses klasifikasi sehingga memerlukan label atau kelas yang akan menjadi target dari setiap *frame* pada audio. Label tersebut dapat dibangun menggunakan kamus kata yaitu jumlah kata unik yang terdapat di dalam transkrip dari *training* dan *testing* data. Jumlah kata unik untuk kamus kata berbahasa Sunda dalam penelitian ini adalah sebanyak 12.067 yang diekstrak dari transkrip *training* dan *testing* data. Selanjutnya, kata unik tersebut akan diurai menjadi suku kata (*phonemes*) atau leksikal yang merupakan cara baca dari setiap kata. Tabel 3 menunjukkan contoh kata dan suku kata yang

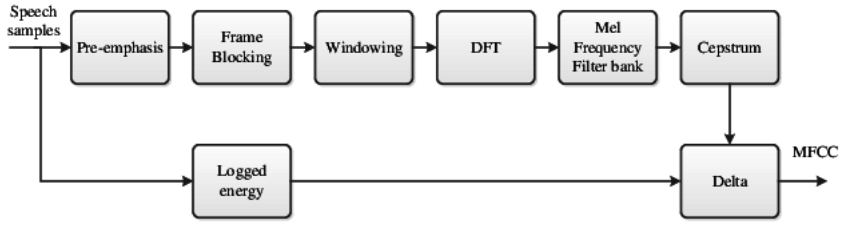
dibangun dari kamus kata. Aturan yang digunakan untuk membangun leksikal adalah berdasarkan huruf vocal dan konsonan yang ada pada sebuah kata, seperti *sayang* akan menjadi /sa/, /ya/, /ng/. Selanjutnya, setiap suku kata akan menjadi label atau kelas untuk proses klasifikasi dalam membangun model akustik. Jumlah suku kata yang berhasil diesktrak adalah sebanyak 276 yang berarti bahwa jumlah kelas untuk proses klasifikasi adalah sebanyak suku kata yang ada.

Tabel 3 Kata dan suku kata berbahasa Sunda

Kata	Suku Kata
adang	a da ng
kadek	ka de k
kakurung	ka ku ru ng
kamari	ka ma ri
trapani	t ra pa ni
teuas	te u a s

3.2. Ekstraksi Fitur

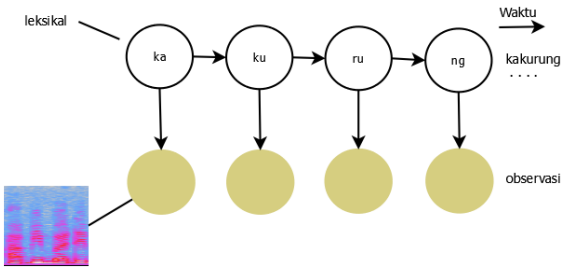
Tahapan ekstraksi fitur merupakan tahapan penting dalam proses membangun model akustik. Fitur yang digunakan untuk membangun model akustik dari data *voice* adalah MFCC (*Mel-Frequency Cepstral Coefficient*). Fitur MFCC akan diekstrak menggunakan *sliding window* sebesar 25 ms (*milliseconds*) dan *overlapping* sebesar 10 ms untuk total waktu dari sebuah data *voice*. Fitur tersebut menggunakan 40 dimensi untuk setiap *frame* yang diekstrak dan disimpan dalam bentuk vector untuk setiap frame. Gambar 3 mengilustrasikan proses yang dilalui untuk memperoleh fitur MFCC dari sebuah data *voice*.



Gambar 3 Ilustrasi ekstraksi fitur MFCC (Sumber [13])

3.3 Data Alignment

Data *alignment* merupakan tahapan untuk mendapatkan target dari setiap *frame* yang telah diekstrak menjadi fitur. Di tahapan ini, model GMM-HMM digunakan untuk mencocokkan setiap *frame* dengan suku kata yang merepresentasikannya. Pada sistem *voice-to-text*, GMM akan memodelkan distribusi fitur untuk setiap suku kata yang ada dengan menghitung nilai *likelihood* untuk setiap distribusi *gaussian*. Setiap suku kata dapat dibentuk oleh beberapa *frame* karena adanya *overlapping* pada tahapan esktraksi fitur. Sedangkan HMM merupakan rantai *markov* yang memuat variabel tertutup antara fitur *voice* dengan suku katanya seperti yang ditunjukkan pada Gambar 4.



Gambar 4 Ilustrasi HMM untuk observasi terhadap leksikalnya

3.4 Pelatihan Model Akustik

Model akustik merupakan sebuah model statistik yang dibangun untuk mendapatkan korelasi antara *frames* dan suku katanya. Secara umum, model-model yang digunakan pada sistem *voice-to-text* dapat direpresentasikan dalam bentuk fungsi distribusi berikut

$$W^* = \arg\max_W P(W|X) \quad (1)$$

$$W^* = \arg\max_W \frac{P(X|W)P(W)}{P(X)} \quad (2)$$

$$W^* = \arg\max_W P(X|W)P(W) \quad (3)$$

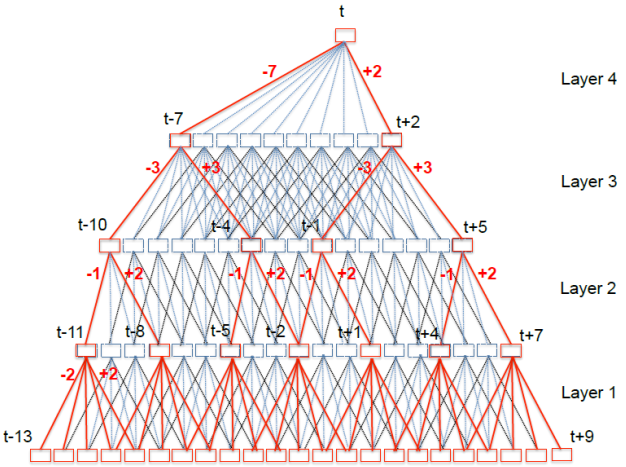
dimana W^*merupakan urutan kata. Fungsi $P(X|W)$ dan $P(W)$ secara berturut-turut adalah model akustik dan model bahasa. Tujuan umum dari fungsi (3) adalah menemukan urutan kata yang sesuai dengan memaksimalkan keluaran dari model akustik dan model bahasa terhadap kata tersebut. Umumnya, fungsi $P(X|W)$ atau model akustik dapat dibangun menggunakan pendekatan GMM-HMM [10], namun model tersebut tidak mampu bekerja optimal untuk data *audio* yang banyak dan *speaker* berbeda-beda [1]. Sedangkan fungsi $P(W)$ adalah model bahasa yang digunakan untuk mendapatkan nilai probabilitas terbaik di antara urutan kata. Gambar 5 mengilustrasikan bagaimana penggunaan model bahasa.

ketua jurusan melakukan kunjungan ke daerah (1)

ketua jurusan kunjungan melakukan ke daerah (2)

Gambar 5 Penggunaan model bahasa untuk urutan kata

Model GMM-HMM pada penelitian ini hanya digunakan untuk melakukan *alignment* dari data *voice*. Selanjutnya, data *alignment* tersebut digunakan untuk melatih model akustik menggunakan *neural network*. Jenis *neural network* yang digunakan untuk melatih akustis adalah TDNN (*Time Delay Neural Networks*) [4]. Pada struktur standar *deep neural network*, waktu permulaan dan akhir dari sebuah *utterance* harus ditentukan sebelum dilakukan proses perkalian linier. Data fitur akan diinput secara keseluruhan ke dalam struktur *deep neural network* sekaligus dalam satu waktu. Sedangkan pada stuktur TDNN, data fitur tidak diinput secara keseluruhan melainkan dibagi ke dalam beberapa bagian pada waktu yang berbeda. Gambar 6 menunjukkan struktur TDNN yang dibangun berdasarkan waktu berbeda untuk setiap *layer*.



Gambar 6 Struktur *Time Delay Neural Network* (Sumber [4])

Selanjutnya, fungsi $P(X|W)$ yang dimodelkan menggunakan GMM-HMM diganti menggunakan TDNN. Secara umum, formulasi untuk setiap *layer* pada *neural network* didefinisikan sebagai berikut.

$$\mathbf{a} = \mathbf{W}_{\boldsymbol{a}} \cdot \mathbf{x} + \mathbf{b} \quad (4)$$

$$\mathbf{z} = relu(\mathbf{a}) \quad (5)$$

$$\mathbf{y} = \text{softmax}(\mathbf{W}_{\boldsymbol{z}} \cdot \mathbf{z} + \mathbf{b}) \quad (6)$$

dimana **a** merupakan hasil perkalian linier antara vektor input **x** dan parameter matrik $\mathbf{W}_{\boldsymbol{a}}$ serta dijumlahkan dengan bias **b**. Sedangkan nilai **z** diperoleh setelah dilakukan aktivasi terhadap nilai **a** menggunakan *relu* sebagai fungsi aktivasi. Nilai **y** merupakan nilai keluaran akhir yang diperoleh dari hasil aktivasi proses perkalian linier antara vektor **z** dan parameter matriks $\mathbf{W}_{\boldsymbol{z}}$ menggunakan fungsi *softmax*. Nilai **y** tersebut kemudian digunakan untuk menghitung selisih terhadap target kelas yang didefiniskan sebelumnya menggunakan data *alignment*. Nilai selisih dapat dihitung menggunakan formulasi berikut.

$$Cross\ Entropy\ Loss = -(y_i \log(\hat{y}_i)) + (1 - y_i)\log(1 - \hat{y}_i) \quad (7)$$

dimana y_i dan $\hat{y}_i$ merupakan nilai prediksi dari *neural network* dan target yang didefinisikan secara berturut-turut.

Tahapan pengujian merupakan tahapan untuk menguji model yang telah dilatih menggunakan data testing. Model akustik yang telah dilatih akan diuji menggunakan data testing sebanyak 1994 ucapan dari 6 *speakers* berbeda. Pada proses pengujian model akustik, mode bahasa juga diperlukan untuk memprediksi urutan kata yang tepat. Model bahasa yang digunakan pada penelitian ini adalah model yang dibangun menggunakan pendekatan *n-gram* [21].

4. Analisis Pembahasan

Model akustik dalam penelitian ini dibangun menggunakan Kaldi *toolkit* [6]. Proses pelatihan model dimulai dari mengekstraksi fitur MFCC kemudian melatih model GMM-HMM untuk memperoleh data *alignment*. Model *i-vector* [19] juga dibangun untuk menyimpan informasi dari setiap *speaker* sehingga kinerja model akustik menjadi lebih baik. Selanjutnya, model akustik dari struktur TDNN dilatih menggunakan *learning rate* sebesar 0.015 dan *epochs* sebesar 3. Tabel 4 menunjukkan WER (*Word Error Rate*) untuk model akustik menggunakan GMM-HMM dan TDNN.

Tabel 4 Performa model akustik

Model Akustik	WER (%)
Monophone	43,58
Triphone	7,19
TDNN – 13L128N	0,59

** WER adalah persentase jumlah kata error yang diprediksi oleh model*

Hasil pengujian pada Tabel 4 menunjukkan bahwa model akustik yang dilatih menggunakan struktur TDNN – 13 *hidden layer* dan 128 *hidden nodes* dapat menurunkan WER mutlak sebesar 6.6%. Pengujian dilanjutkan dengan mengubah *hidden nodes* menjadi lebih besar dan menambahkan *layer SpecAugment* [20] untuk mendapatkan hasil yang lebih baik. Hasil pengujian untuk struktur TDNN berbeda ditunjukkan oleh Tabel 5.

Tabel 5 Performa model akustik menggunakan struktur TDNN

Model Akustik	WER (%)
TDNN – 13L128N	0,59
TDNN – 13L128N + SpecAug	0,57
TDNN – 13L256N + SpecAug	0,51

Penurunan WER mutlak sebesar 0.02% diperoleh dari TDNN – 13 *hidden layers* 128 *hidden nodes* dan dikombinasikan dengan *layer SpecAugment*. Sedangkan hasil terbaik dari model akustik menggunakan TDNN diperoleh dari struktur TDNN dengan 13 *hidden layers* 256 *hidden nodes* dikombinasikan dengan *layer SpecAugment*.

5. Kesimpulan dan Saran

Penelitian ini dapat disimpulkan bahwa penggunaan *Time Delay Neural Network* pada sistem *voice-to-text* dapat meningkatkan kinerja akustik model dibandingkan dengan model GMM-HMM biasa. Penambahan jumlah *hidden nodes* pada struktur TDNN dapat meningkatkan performa model akustik mejadi lebih baik. Penelitian selanjutnya akan dilakukan penambahan data training dengan teknik data augmentasi sehingga model akustik menjadi lebih baik. Selain itu, penggunaan struktur *neural network* yang lain seperti CNN (*Convolutiinal Neural Network*) dan LSTM (*Long Short-Term Memory*) juga akan dilakukan di penelitian selanjutnya.

Daftar Pustaka

[1] Deng, L., Hinton, G., & Kingsbury, B. (2013, May). New types of deep neural network learning for speech recognition and related applications: An overview. In *2013 IEEE International Conference on Acoustics, Speech and Signal Processing* (pp. 8599-8603). IEEE.

[2] Vanhoucke, V., Devin, M. and Heigold, G., 2013, May. Multiframe deep neural networks for acoustic modeling. In *2013 IEEE International Conference on Acoustics, Speech and Signal Processing* (pp. 7582-7585). IEEE.

[3] Hinton, G., Deng, L., Yu, D., Dahl, G., Mohamed, A.R., Jaitly, N., Senior, A., Vanhoucke, V., Nguyen, P., Kingsbury, B. and Sainath, T., 2012. Deep neural networks for acoustic modeling in speech recognition. *IEEE Signal processing magazine*, *29*.

[4] Peddinti, V., Povey, D. and Khudanpur, S., 2015. A time delay neural network architecture for efficient modeling of long temporal contexts. In *Sixteenth Annual Conference of the International Speech Communication Association*.

[5] Seltzer, M.L., Yu, D. and Wang, Y., 2013, May. An investigation of deep neural networks for noise robust speech recognition. In *2013 IEEE international conference on acoustics, speech and signal processing* (pp. 7398-7402). IEEE.

[6] Povey, D., Ghoshal, A., Boulianne, G., Burget, L., Glembek, O., Goel, N., Hannemann, M., Motlicek, P., Qian, Y., Schwarz, P. and Silovsky, J., 2011. The Kaldi speech recognition

toolkit. In *IEEE 2011 workshop on automatic speech recognition and understanding* (No. CONF). IEEE Signal Processing Society.

[7] Ittichaichareon, C., Suksri, S. and Yingthawornsuk, T., 2012, July. Speech recognition using MFCC. In *International Conference on Computer Graphics, Simulation and Modeling* (pp. 135-138).

[8] Dimitriadis, D., Maragos, P. and Potamianos, A., 2005. Robust AM-FM features for speech recognition. *IEEE signal processing letters*, *12*(9), pp.621-624.

[9] Tiwari, V., 2010. MFCC and its applications in speaker recognition. *International journal on emerging technologies*, *1*(1), pp.19-22.

[10] Su, D., Wu, X. and Xu, L., 2010, March. GMM-HMM acoustic model training by a two level procedure with Gaussian components determined by automatic model selection. In *2010 IEEE International Conference on Acoustics, Speech and Signal Processing* (pp. 4890-4893). IEEE.

[11] Kjartansson, O., Sarin, S., Pipatsrisawat, K., Jansche, M. and Ha, L., 2018. Crowd-Sourced Speech Corpora for Javanese, Sundanese, Sinhala, Nepali, and Bangladeshi Bengali.

[12] Hughes, T., Nakajima, K., Ha, L., Vasu, A., Moreno, P.J. and LeBeau, M., 2010. Building transcribed speech corpora quickly and cheaply for many languages. In *Eleventh Annual Conference of the International Speech Communication Association.*

[13] Tran, V.H., Nguyen, L.T.T., Hoang, T. and Tran, X.T., 2013. Design and Implementation of a SoPC System for Speech Recognition.

[14] Sakti, S. and Nakamura, S., 2014. Recent progress in developing grapheme-based speech recognition for Indonesian ethnic languages: Javanese, Sundanese, Balinese and Bataks. In *Spoken Language Technologies for Under-Resourced Languages*.

[15] Rahmawati, R. and Lestari, D.P., 2017, October. Java and Sunda dialect recognition from Indonesian speech using GMM and I-Vector. In *2017 11th International Conference on Telecommunication Systems Services and Applications (TSSA)* (pp. 1-5). IEEE.

[16] Arwandani, G., Osmond, A.B. and Nugrahaeni, R.A., 2018. Deep Neural Network Untuk Pengenalan Ucapan Pada Bahasa Sunda Dialek Utara. *eProceedings of Engineering*, *5*(3).

[17] Fathurrahman, D.N., Osmond, A.B. and Saputra, R.E., 2018. Deep Neural Network Untuk Pengenalan Ucapan Pada Bahasa Sunda Dialek Tengah Timur (majalengka). *eProceedings of Engineering*, *5*(3).

[18] Hakim, L.A., Osmond, A.B. and Saputra, R.E., 2018. Recurrent Neural Network Untuk Pengenalan Ucapan Pada Bahasa Sunda Selatan Dialek Garut. *eProceedings of Engineering*, *5*(3).

[19] Kjartansson, O., Sarin, S., Pipatsrisawat, K., Jansche, M. and Ha, L., 2018. Crowd-Sourced Speech Corpora for Javanese, Sundanese, Sinhala, Nepali, and Bangladeshi Bengali.

[20] Senior, A. and Lopez-Moreno, I., 2014, May. Improving DNN speaker independence with i-vector inputs. In *2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (pp. 225-229). IEEE.

[21] Park, D.S., Chan, W., Zhang, Y., Chiu, C.C., Zoph, B., Cubuk, E.D. and Le, Q.V., 2019. Specaugment: A simple data augmentation method for automatic speech recognition. *arXiv preprint arXiv:1904.08779*.

[22] Stolcke, A., 2002. SRILM-an extensible language modeling toolkit. In *Seventh international conference on spoken language processing*.