



PENGEMBANGAN PEMODELAN DEEP
LEARNING LONG-SHORT TERM MEMORY
UNTUK PENGENALAN SUARA DIGIT
BILANGAN DESIMAL BERBAHASA
INDONESIA

DISERTASI

ADAM HUDA NUGERAHA

99217001

PROGRAM DOKTOR TEKNOLOGI INFORMASI

UNIVERSITAS GUNADARMA

2021

BAB I

PENDAHULUAN

1.1 Latar Belakang

Berbicara adalah bentuk komunikasi paling umum dan paling mudah yang dimiliki oleh manusia. Setiap manusia umumnya terlahir mempunyai organ-organ biologis yang dapat membuatnya berkomunikasi dengan cara berbicara dan belajar mengembangkan kemampuan tersebut sejak kecil mengikuti lingkungan dan budaya tempatnya berada, sehingga dengan berbicara manusia dapat saling berinteraksi dan menyampaikan informasi dengan cepat dan mudah, walaupun masih terbatas oleh jarak. Perkembangan teknologi memperluas dalam penyampaian informasi, sehingga manusia bias berbicara dalam jarak jauh dengan menggunakan telepon (kabel), telepon genggam (cdma dan gsm) dan komputer (voip), namun kemampuan teknologi ini hanya sebatas menjadi perantara sinyal-sinyal suara..

Teknologi pengenalan suara kemudian berkembang menjadi teknologi yang mempermudah manusia dan memberikan kenyamanan terhadap manusia dalam memanfaatkan informasi menggunakan perantara suara. Saat ini teknologi bukan hanya sekedar perantara dalam membantu peranan manusia dalam berkomunikasi dan bertukar informasi dalam melakukan berbagai interaksi dengan adanya alat-alat yang tersedia saat ini. Seperti halnya teknologi antarmuka yang mempermudah manusia dalam mengekspresikan apa yang dimaksud dalam upaya berinteraksi untuk mencapai tujuan yang sama. Antarmuka suara akan sangat mendukung aplikasi-aplikasi seperti memberikan perintah tertentu lewat suara untuk dijalankan, menterjemahkan sinyal suara menjadi teks kemudian menterjemahkan kembali menjadi suara dalam bahasa lain, memverifikasi *password* atau *pin* yang diucapkan,

dan lain-lain. Aplikasi-aplikasi ini mendorong para peneliti sejak 1970-an untuk mengembangkan algoritma pengolahan suara agar dapat dikenali dengan baik oleh komputer.

Suara adalah sinyal gelombang yang merambat melalui udara karena adanya getaran dalam molekul udara (McLoughlin, 2009). Di dalam udara suara merambat sebagai variasi tekanan tinggi dan rendah yang disebut dengan *amplitude*. *Microphone* menangkap sinyal suara ini dan diubah menjadi sinyal digital oleh *Analog to Digital Converter* (ADC), proses pengubahan ini di-digitalisasi dengan format standar tertentu yang telah ditetapkan (McLoughlin, 2009). *Pulse Code Modulation* (PCM) adalah format standar yang digunakan saat ini. Sinyal-sinyal ini disimpan dengan mengukur tingginya amplitud pada waktu tertentu, kemudian agar menjadi gelombang, diambil dua nilai dalam satu siklus, nilai positif dan negatif (Jurafsky dan Martin, 2008). Banyaknya nilai sinyal yang diambil dalam satu detik disebut dengan *sampling rate*. Semakin tinggi *sampling rate* menyebabkan semakin akurat nilai sinyal, sehingga dalam pengolahan suara umumnya diambil *sample rate* dua kali dari frekuensi minimum seperti yang ditetapkan dalam *Nyquist frequency* (Jurafsky dan Martin, 2008). Minimum *sampling rate* suara manusia berada pada kisaran dibawah 4000Hz, seperti yang digunakan pada jalur telepon, sehingga pengolahan suara dengan frekuensi 8000Hz merupakan nilai dengan *sampling rate* yang cukup untuk mewakili suara manusia.

Suara manusia terbentuk dari gelombang akustik yang bergetar ketika udara keluar dari paru-paru melalui tenggorokan dan mulut atau hidung. Terdapat tiga rongga utama yang membuat manusia bersuara yaitu rongga hidung, rongga mulut, dan rongga faring (Chu, 2003). Saluran vokal (*vocal tract*) mengacu pada rongga faring dan rongga mulut. Saluran nasal (*nasal tract*) berawal dari rongga langit-langit tenggorokan (*velum*) dan berakhir pada rongga hidung (Chu, 2003). Suara hidung adalah suara yang dihasilkan ketika langit-langit tenggorokan menutup. Dalam laring (*larynx*) atau pangkal tenggorokan terdapat komponen penting dalam pengolahan suara yang disebut dengan pita suara (*vocal folds*). Pita suara ini terdiri

atas pasangan otot dan membran yang menutup dan membuka dengan cepat (bergetar). Kecepatan pita suara dalam bergetar atau disebut juga dengan pitch ini unik dalam setiap manusia dan mendefinisikan ciri dari pembicara (Stevens. 2000). Secara umum pitch berada antara rentang 50 - 250 siklus per detik (Hz) untuk suara laki-laki dewasa sampai dengan lebih dari 120 - 500 siklus per detik (Hz) untuk suara perempuan, atau dalam domain waktu berada antara 4 ms - 20 ms untuk laki-laki dan 2 ms - 8 ms untuk perempuan (Chu. 2003). Bentuk dan formasi dari saluran vokal dan saluran nasal ini juga berubah secara terus menerus dalam waktu, ketika udara keluar dari paru-paru melalui tenggorokan, spektrum frekuensi yang terbentuk tergantung dari bentuk dan panjang dimensi saluran vokal (Chu. 2003). Spektrum frekuensi yang dihasilkan ini juga unik tergantung dari panjang saluran vokal tiap manusia, panjang saluran vokal (*vocal tube*) ini secara umum adalah 14,1 cm untuk perempuan hingga 16,9 cm untuk laki-laki (Stevens. 2000).

Pemanfaatan teknologi informasi sebagai sumber informasi yang cepat sangat menunjang dalam berbagai hal, seseorang bisa menemukan informasi yang diinginkan hanya dengan hitungan menit tanpa harus bersusah payah mendatangi pusat informasi ataupun membeli koran dan sebagainya. Perkembangan teknologi saat ini telah memberikan pengaruh yang sangat besar bagi dunia teknologi informasi. Sinyal-sinyal suara yang di-digitalisasi dalam bentuk amplitudo inilah yang diolah untuk dapat dikenali. Amplitudo sinyal suara selain dinyatakan dalam domain waktu dapat dilihat juga dalam domain frekuensi yang disebut dengan spektrogram. Spektrogram menggambarkan sinyal dalam bentuk dua dimensi dengan waktu sebagai sumbu-x, frekuensi dalam sumbu-y dan energi frekuensi digambarkan dengan intensitas warna. Sinyal suara diubah kedalam bentuk spektrogram menggunakan Fourier Transform, karena formula Fourier dapat mengekstrak kandungan frekuensi yang terdapat dalam sinyal suara.

Dalam kondisi nyata yang dihadapi oleh manusia, untuk satu pembicara walaupun mengucapkan suara atau kata yang sama, nilai-nilai dari sinyal suara ini tidak akan sama persis, cara pengucapan (panjang atau pendek dalam mengucapkan

kata), dialek (bahasa yang berbeda) pembicara, ekspresi yang menghasilkan intonasi tekanan tinggi rendah suara yang dikeluarkan saat berbicara dan emosi saat berbicara seringkali berubah-ubah setiap waktu. Suara (kata) yang sama diucapkan oleh pembicara yang berbeda juga memiliki keunikannya masing-masing (misal, jenis kelamin dan usia), karena setiap sinyal suara dihasilkan melalui saluran vokal dan saluran nasal yang berbeda-beda setiap manusia. Sinyal suara juga dapat bercampur dengan sinyal bunyi-bunyi lain (*noise*) seperti suara hujan, suara kendaraan, suara pendingin udara dalam ruangan dan lain-lain, tergantung dengan lingkungan tempatnya berbicara. Hal inilah yang membuat pengenalan suara menjadi topik penelitian yang terus berkembang hingga saat ini.

Pengenalan suara pada dasarnya membutuhkan basis data suara (*training*) dan proses pengenalan (klasifikasi). Basis data suara berisi koleksi sinyal-sinyal suara yang telah diberi label. Proses pengenalan adalah proses untuk mengidentifikasi sinyal-sinyal yang diucapkan dengan cara membandingkan dengan sinyal-sinyal suara yang terdapat dalam basis data (Chavan dan Sable. 2013). Sinyal-sinyal suara dalam bentuk amplitudo atau spektrogram sangat sulit dilakukan proses pengenalan karena selain banyaknya data yang harus diproses (untuk sample rate 8000Hz dalam satu detik terdapat 16000 data numerik), nilai-nilai amplitudo tersebut juga sangat berfluktuasi tergantung dari tekanan pengucapan, kualitas microphone, noise, dan lain-lain. Algoritma untuk mengekstraksi ciri (koefisien) dari nilai-nilai amplitudo ini kemudian dikembangkan agar mempermudah proses pengenalan sinyal suara. Ekstraksi ciri adalah mengubah sinyal-sinyal amplitudo atau spektrogram menjadi hanya beberapa vektor koefisien yang diperkirakan mengandung informasi yang penting. Beberapa algoritma ekstraksi ciri yang paling sering digunakan adalah *Linear Predictive Coding* (LPC) yang mengambil nilai koefisien berdasarkan prediksi dari nilai sebelumnya, biasanya sekitar 10 koefisien. *Mel-Frequency Cepstrum Coefficient* (MFCC) yang dikenalkan oleh Davis dan Mermelstein pada tahun 1980, yang mengambil koefisien frekuensi dari sinyal suara, sekitar 12 vektor koefisien

(Davis dan Mermelstein. 1990), dan *Perceptual Linier Prediction* (PLP) yang menambahkan filter *loudness* untuk menghilangkan *noise* dan berasumsi bahwa intensitas pendengaran manusia berada di sekitar 40 db (Ningthoujam dan Prathima. 2016).

Terdapat berbagai metode yang digunakan dalam pengenalan suara, yaitu: *Template Based* (1970), *Statistical Learning* (1980), *Machine Learning* (1990) dan *Deep Learning* (2000). Sakoe dan Chiba pada tahun 1978 mengenalkan algoritma *Dynamic Time Warping* (DTW) untuk mengenali suara digit dalam bahasa Jepang berdasarkan *template based*. Basis data *template* menggunakan ekstraksi ciri FBANK diambil dari 10 laki-laki mengucapkan digit 0-9 dalam bahasa Jepang yang diulang sebanyak 6 kali. Untuk pengenalan (*matching*) 10 pembicara yang sama mengucapkan digit 0-9 diulang sebanyak 5 kali. Tingkat akurasi yang dihasilkan mencapai 90% (Sakoe dan Chiba. 1978). Algoritma ini mampu mengatasi masalah perbedaan panjang data suara setiap pembicara. Algoritma ini telah digunakan untuk mengenali suara digit di beberapa bahasa, seperti bahasa Arab dengan ekstraksi ciri MFCC (Darabkh, Khalifeh, Bathech dan Sabah. 2013; Hachkar, Farchi, Mounir dan El-Abbadi. 2011), bahasa Inggris dengan ekstraksi ciri MFCC (Limkar, Rao dan Sagvekar. 2012) dan WMFCC (Chapaneri dan Jayaswal. 2013), bahasa Malaysia dengan ekstraksi ciri MFCC (Al-Haddad, Samad, Hussain, Ishak dan Mirvaziri. 2007), bahasa Gujarati dengan ekstraksi ciri MFCC (Pandit dan Bhatt. 2014) dan bahasa Spanyol dengan ekstraksi ciri MFCC (Terissi dan Gómez. 2005), walaupun semuanya menunjukkan tingkat akurasi yang cukup tinggi, diatas 90%, namun algoritma ini hanya mampu mengenali jika data suara pembicara telah berada dalam basis data (*speaker dependent*). Untuk data pembicara yang tidak terdapat dalam basis data, *template based* DTW menunjukkan tingkat akurasi yang rendah. Tingkat akurasi pengenalan adalah jumlah ketepatan suara digit yang diucapkan berbanding dengan jumlah data suara yang akan diterjemahkan menjadi teks, sehingga semakin tinggi akurasi pengenalan menunjukkan semakin tingginya hasil ketepatan sistem menterjemahkan sinyal-sinyal suara ini menjadi teks.

Sehingga proses dalam menterjemahkan sebuah informasi dapat berjalan lebih cepat dan efektif.

Ekstraksi ciri adalah mengubah sinyal-sinyal amplitudo atau spektrogram menjadi hanya beberapa vektor koefisien yang diperkirakan mengandung informasi yang penting. Beberapa algoritma ekstraksi ciri yang paling sering digunakan adalah Linier Predictive Coding (LPC) yang mengambil nilai koefisien berdasarkan prediksi dari nilai sebelumnya, biasanya sekitar 10 koefisien. Mel-Frequency Cepstrum Coefficient (MFCC) yang dikenalkan oleh Davis dan Mermelstein pada tahun 1980, yang mengambil koefisien frekuensi dari sinyal suara, sekitar 12 vektor koefisien (Davis dan Mermelstein. 1990), dan Perceptual Linier Prediction (PLP) yang menambahkan filter loudness untuk menghilangkan noise dan berasumsi bahwa intensitas pendengaran manusia berada di sekitar 40 db (Ningthoujam dan Prathima. 2016).

Pengenalan suara berdasarkan pengucapan dapat diklasifikasikan sebagai (Ningthoujam dan Prathima. 2016): Isolated Words, Connected Words dan Continuous Speech. Isolated words adalah mengenali suara per kata yang diucapkan, sistem pengenalan menunggu satu kata selesai diucapkan lalu proses pengenalan dimulai. Connected words mirip dengan isolated words, namun mampu mengenali lebih dari satu kata yang diucapkan, sedangkan pada continuous speech sistem mengenali terus menerus setiap kata yang diucapkan tanpa ada jeda menunggu. Terdapat berbagai metode yang digunakan dalam pengenalan suara, yaitu: Template Based (1970), Statistical Learning (1980), Machine Learning (1990) dan Deep Learning (2000).

Sakoe, Isotani, Yoshida, Iso dan Watanabe pada tahun 1989 mengenalkan algoritma *machine learning: Dynamic Programming Neural Network* untuk mengenali suara digit berbahasa Jepang. Data untuk pembelajaran (*training*) dengan ekstraksi ciri MFCC diambil dari 50 pembicara mengucapkan digit 0-9 dalam bahasa Jepang, dan data suara dari 57 pembicara berbeda digunakan saat

proses pengenalan. Tingkat akurasi mencapai 90% (Sakoe, Isotani, Yoshida, Iso dan Watanabe. 1990). Jumlah basis data suara yang digunakan saat proses pembelajaran (training) sangat penting dalam *machine learning*, semakin banyak data suara yang digunakan akan semakin meningkatkan akurasi pengenalan, dan seperti pada *statistical learning* dengan HMM, ekstraksi ciri yang digunakan dalam proses pembelajaran *machine learning* juga sangat menentukan tingkat akurasi. Proses pembelajaran *neural network* dengan spektrogram membutuhkan layer yang besar (*deep*), sehingga diperlukan neural network khusus untuk proses pengenalan suara. *Neural Network* untuk pengenalan suara dengan layer yang besar (diatas 100 hidden layer), mempunyai dua masalah utama untuk mencapai network yang optimal, yaitu: Vanishing Gradient (Hochreiter dan Schmidhuber. 1997) dan Dependency (Lekshmi dan Sherly. 2016). Vanishing Gradient adalah kondisi hilangnya nilai error (nilai selisih antara prediksi output dengan prediksi sesungguhnya, yang dibutuhkan untuk memperbaiki bobot layer) saat proses backward-pass dari hidden layer di ujung output ke hidden layer di dekat input (Hochreiter dan Schmidhuber. 1997). *Dependency* terjadi karena pada neural network, semua input layer diasumsikan tidak saling berhubungan (*independent*) (Lekshmi dan Sherly. 2016), sedangkan pada data suara, setiap 20 milidetik yang diambil pada dasarnya saling berhubungan dengan data 20 milidetik berikutnya.

Jumlah basis data suara yang digunakan saat proses pembelajaran (training) sangat penting dalam machine learning, semakin banyak data suara yang digunakan akan semakin meningkatkan akurasi pengenalan, dan seperti pada statistical learning dengan HMM, ekstraksi ciri yang digunakan dalam proses pembelajaran machine learning juga sangat menentukan tingkat akurasi. Ekstraksi ciri mana yang akan dipakai? siapa yang lebih baik dari berbagai macam ekstraksi ciri yang ada? atau apakah perlu menemukan ekstraksi ciri yang lebih baik? bagaimana jika tidak menggunakan ekstraksi ciri, kembali ke data mentah yang berupa amplitudo atau spektrogram?. Tingkat akurasi dengan MFCC terbukti lebih baik dari LPC. MFCC berbasis frekuensi. Spektrogram berbasis frekuensi. Neural Network dengan data

spektrogram?. Pertanyaan-pertanyaan ini yang kemudian mendorong para peneliti menemukan metode Deep Learning.

Data spektrogram dapat di input langsung ke neural network, dan diharapkan neural network sendiri yang menemukan pola ekstraksi ciri suara (model akustik pembicara). Untuk sample rate 8000Hz, vektor data spektrogram berjumlah 64 koefisien setiap 20 milidetik sehingga dalam satu detik terdapat sekitar 64x50 vektor data suara, dibandingkan dengan MFCC yang hanya 12x50 koefisien atau LPC yang 10x50 koefisien. Proses pembelajaran neural network dengan spektrogram membutuhkan layer yang besar (deep), sehingga diperlukan neural network khusus untuk proses pengenalan suara. Neural Network untuk pengenalan suara dengan layer yang besar (+ diatas 100 hidden layer), mempunyai dua masalah utama untuk mencapai network yang optimal, yaitu: Vanishing Gradient (Hochreiter dan Schmidhuber. 1997) dan Dependency (Lekshmi dan Sherly. 2016). Vanishing Gradient adalah kondisi hilangnya nilai error (nilai selisih antara prediksi output dengan prediksi sesungguhnya, yang dibutuhkan untuk memperbaiki bobot layer) saat proses backward-pass dari hidden layer di ujung output ke hidden layer di dekat input (Hochreiter dan Schmidhuber. 1997). Dependency terjadi karena pada neural network, semua input layer diasumsikan tidak saling berhubungan (independent) (Lekshmi dan Sherly. 2016), sedangkan pada data suara, setiap 20 milidetik yang diambil pada dasarnya saling berhubungan dengan data 20 milidetik berikutnya.

Long-Short Term Memory (LSTM) Network yang dikenalkan oleh Hochreiter dan Schmidhuber pada tahun 1997 memperbaiki masalah *vanishing gradient* dan keterhubungan antara input layer (*long-term dependencies*) yang dibuat khusus untuk mengatasi masalah di pengenalan suara (Hochreiter dan Schmidhuber. 1997). LSTM menjadi model *Deep Learning* yang sering digunakan hingga saat penelitian ini dilakukan. Google pada tahun 2010 mulai menggunakan LSTM untuk mengenali data suara berbahasa Inggris pada 22500 kosakata bahasa Inggris, kemudian menjadi awal dari sistem pengenalan suara di Google *Voice*

Search Task yang diimplementasikan di situs Google (Sak, Senior dan Beaufays. 2014). Microsoft menggunakan Deep Learning LSTM untuk sistem pengenalan suara multi-bahasa di internal Microsoft dengan data suara yang besar pada proses pembelajaran, 138 jam data suara berbahasa Perancis, 195 jam data suara berbahasa Italia (Deng, Li, Huang, Yao, Yu, Seide, Seltzer, Zweig, He dan Williams. 2013). Model LSTM dengan penambahan satu layer untuk pengenalan aksent digunakan untuk mengenali suara berbahasa Mandarin, dengan tingkat akurasi 90% (Yi, Ni, Wen, Liu dan Tao. 2016). LSTM digunakan untuk mengenali suara pada basis data TIMIT (basis data fonem berbahasa Inggris dengan dialek Amerika, direkam dari 630 pembicara yang dikumpulkan oleh Massachusetts Institute of Technology, SRI International dan Texas Instrument) menunjukkan tingkat akurasi yang lebih tinggi dibandingkan dengan metode lainnya (Graves, Mohamed dan Hinton. 2013).

Akurasi atau tingkat akurasi pengenalan adalah jumlah ketepatan suara digit yang diucapkan berbanding dengan jumlah data suara yang akan diterjemahkan menjadi teks, sehingga semakin tinggi akurasi pengenalan menunjukkan semakin tingginya hasil ketepatan sistem menterjemahkan sinyal-sinyal suara ini menjadi teks. Lawrence Rabiner dan Biing Hwang Juang pada tahun 1986, mengusulkan penggunaan model Hidden Markov (HMM) yang telah diketahui sejak lama di kalangan matematikawan (pertama kali dikenalkan oleh L.E. Baum pada tahun 1966) untuk memodelkan banyak suara pembicara menjadi hanya beberapa model sehingga menghilangkan ketergantungan dengan basis data dan pengenalan suara yang speaker independent dapat dihasilkan. Speaker independent adalah pengenalan suara yang tidak tergantung pada siapa pembicaranya, pembicara pada basis data suara berbeda dengan pembicara pada basis data uji. Basis data pada HMM hanya dipakai saat proses pembelajaran (train) dan setelah didapatkan model, basis data tidak digunakan lagi. Proses pengenalan akan mencari kemiripan ke model-model yang dihasilkan sehingga dapat mempercepat proses. Rabiner dan Juang menunjukkan metode HMM ini dapat memodelkan pengucapan dari beberapa pembicara dan mampu menyamai tingkat akurasi dari template based.

Basis data dengan ekstraksi ciri MFCC diambil dari pembicara mengucapkan digit 0-9 dalam bahasa Inggris. Tingkat akurasi yang dihasilkan mencapai 98% (Rabiner dan Juang. 1986). Model ini hingga sekarang masih banyak digunakan untuk mengenali suara digit dari berbagai bahasa yang berbeda dengan rata-rata tingkat akurasi juga mencapai diatas 90%, seperti bahasa Arab dengan ekstraksi ciri MFCC (Hachkar, et al. 2011; Alotaibi, Alghamdi dan Alotaiby. 2010), bahasa Kroasia dengan ekstraksi ciri MFCC (Gulić, Lučanin dan Šimić. 2011), bahasa Perancis dengan ekstraksi ciri LPC (Tassy dan Miclet. 1986), bahasa India dengan ekstraksi ciri MFCC (Dhandhania, Hansen, Kandi dan Ramesh. 2012; Saxena dan Wahi. 2015), bahasa Malaysia dengan ekstraksi ciri MFCC (Al-Haddad, et al. 2007), dan bahasa Rumania dengan ekstraksi ciri MFCC (Cucu, Caranica, Buzo dan Burileanu. 2015), serta dalam bahasa Inggris dengan memodelkan ciri suara yang berbeda untuk suara digit yang tercampur dengan background suara yang tidak jernih (noise), seperti di keramaian, di jalan raya, di bandara, dan di dalam kereta dengan ekstraksi ciri BFCC dan WMFCC (Mukhedkar dan Alex. 2014). Basis data suara dimodelkan oleh HMM saat proses pembelajaran berdasarkan maksimum probabilitas dari ekstraksi ciri data suara, sehingga menentukan ekstraksi ciri suara yang tepat menjadi sangat penting dalam HMM.

Keberhasilan dari HMM dalam memodelkan data suara dengan ekstraksi ciri tertentu mendorong peneliti untuk memodelkan data suara dengan Neural Network (Machine Learning), dengan harapan menyederhanakan dan membuat model yang lebih baik. Pada neural network hanya menghasilkan satu model yang diasumsikan sudah optimal, dibandingkan dengan beberapa model yang dihasilkan oleh HMM (model fonem, model bahasa atau model word). Neural Network telah terbukti mampu memodelkan data yang tidak linier dan telah diaplikasikan pada berbagai macam bidang mulai dari memprediksi kebangkrutan, mengenali tulisan tangan hingga mendiagnosis penyakit (Zhang. 2000).

Pada tahun 2013 penelitian untuk pengenalan suara digit berbahasa Indonesia telah dilakukan dengan menggunakan CMU-Sphinx (Dewi, Firdausillah

dan Supriyanto. 2013). CMU-Sphinx merupakan sistem pengenalan suara berbasis Hidden Markov Model dengan fitur ekstraksi ciri MFCC yang dikembangkan oleh Carnegie Melon University (CMU) dan Sun Microsystem (Lamere, Kwok, Gouvea, Raj, Singh, Walker, Warmuth dan Wolf. 2003). Proses pembelajaran untuk mendapatkan model HMM diambil dari 7 orang laki-laki mengucapkan digit 0-9 sebanyak 3 kali dalam lingkungan yang bebas noise. Untuk proses pengenalan setiap pembicara yang sama diharuskan mengucapkan digit seperti dalam percakapan yang normal. Tingkat akurasi yang dicapai kurang dari 50%, disebabkan oleh faktor tekanan pengucapan, cara pengucapan dan panjang pengucapan yang berbeda antara model data yang dihasilkan oleh HMM dengan data suara yang diuji-cobakan (Dewi, et al. 2013).

Pada tahun 2016 dilakukan penelitian untuk mengenali suara digit berbahasa Indonesia, hal ini dilakukan dengan menggunakan CMU-Sphinx (Prakoso, Ferdiana dan Hartanto. 2016). Proses pembelajaran model HMM suara digit berbahasa Indonesia pada penelitian ini diambil dari 8 laki-laki dan 7 perempuan mengucapkan digit sebanyak 2 kali dalam lingkungan yang bebas noise. Proses pengenalan dilakukan dengan mengambil 6 pembicara yang mengucapkan 54 digit secara acak dari 99 digit yang disediakan. Tingkat akurasi mencapai 80% (Prakoso, et al. 2016).

1.2 Batasan dan Rumusan Masalah

Penelitian ini menggunakan 400 data suara mahasiswa yang terdiri dari 200 perempuan dan 200 laki-laki dengan rentang usia diantara 19-22 tahun. Data suara ini direkam di dalam ruangan kelas di kampus Universitas Gunadarma yang berwilayah di Bekasi. Setiap mahasiswa mengucapkan digit decimal dalam Bahasa Indonesia, dari 0-9 (nol, satu, dua, tiga, empat, lima, enam, tujuh, delapan, dan sembilan) berjumlah 10 pengucapan yang direkam menggunakan telepon genggam tanpa *microphone* tambahan, sehingga total data suara menjadi 799 (mahasiswa) x

10 (digit) = 7990 data. Data suara tersebut kemudian dipotong-potong sesuai digit yang diucapkan dan diberi label digit (0-9). Dimana data ini disebut dengan data latih.

Dalam proses pengenalan dan menghitung akurasi direkam 79 data mahasiswa berbeda yang terdiri dari 37 perempuan dan 42 laki-laki, dengan rentang usia diantara 19-22 tahun, dimana para mahasiswa melakukan pengucapan yang sama yaitu digit 0-9 dan data yang telah terkumpul akan melalui proses pemotongan yang sama, sehingga total suara pengenalan menjadi $79 \text{ (mahasiswa)} \times 10 \text{ (digit)} = 790$ data, yang mana data ini disebut sebagai data uji.

Berangkat dari upaya untuk membuat sistem pengenalan suara yang dapat membuat komputer memiliki kemampuan seperti manusia dalam melakukan pengenalan terhadap suara, sehingga dapat mempermudah dalam proses interaksi dengan komputer. Berdasarkan beberapa hasil penelitian suara yang dilakukan berbahasa Indonesia dan riset-riset tentang metode pengenalan suara yang telah diuraikan di latar belakang terdapat beberapa masalah yang akan di teliti, Antara lain:

- a. Bagaimana menemukan model arsitektur LSTM yang sesuai untuk pola suara digit decimal berbahasa Indonesia,
- b. Bagaimana tingkat akurasi dari model LSTM yang dibangun untuk mengenali data suara digit decimal berbahasa Indonesia,
- c. bagaimana waktu eksekusi dari model LSTM yang akan dibangun.

1.3 Tujuan Penelitian

Model Deep Learning dengan LSTM belum pernah digunakan sebelumnya untuk meneliti suara digit decimal berbahasa Indonesia, maka penelitian ini bertujuan dalam mengembangkan model LSTM untuk mengenali suara digit

decimal Bahasa Indonesia. Adapun tujuan khusus yang ingin dicapai dalam penelitian ini adalah:

- a. Mengembangkan Deep Learning dengan LSTM untuk mencari model pola suara digit desimal berbahasa Indonesia yang tidak tergantung kepada siapa pembicaranya (*speaker independent*),
- b. Menghasilkan tingkat akurasi yang lebih tinggi dari model LSTM yang dibangun untuk pengenalan suara digit desimal berbahasa Indonesia dibandingkan dengan penelitian sebelumnya,
- c. Menghasilkan waktu eksekusi yang optimal dari model LSTM yang dibangun.

1.4 Kontribusi dan Manfaat penelitian

Hasil penelitian ini memberikan kontribusi keilmuan untuk bidang penelitian pengenalan suara berbahasa Indonesia karena masih sangat jarang penelitian mengenai pengenalan suara berbahasa Indonesia, model LSTM yang dihasilkan dapat dijadikan model LSTM acuan untuk penelitian selanjutnya. Basis data suara digit decimal berbahasa Indonesia yang dikumpulkan sebanyak 7990 data untuk pembelajaran dan 790 data untuk pengenalan dalam penelitian ini.

Model Deep Learning dengan LSTM yang dihasilkan di implementasikan menjadi sebuah perangkat lunak yang dapat dimanfaatkan sebagai alat untuk memverifikasi pengucapan digit pin pada mesin ATM, password yang berupa digit lewat suara ataupun pengucapan pemilihan digit pada mesin operator telepon.

BAB II

TELAAH PUSTAKA

2.1 Suara

Pada umumnya sebagian besar bahasa didunia ini memiliki 2 golongan dasar pengucapan yaitu: Konsonan yang dilafalkan ketika terjadi penyempitan di tenggorokan atau ketika terjadi pemblokiran di dalam bagian mulut (lidah, gigi, bibir) saat pengucapan, sedangkan Vokal dilafalkan ketika ada penyempitan atau pemblokiran udara yang keluar dari paru-paru (Huang, Acero, Hon dan Reddy, 2001).

Suara dapat dibagi ke dalam sub-kelompok berdasarkan karakteristik pengucapannya, Karakteristik ini diperoleh dari sejumlah anatomi pelafalan serta tempat dimana suara tersebut melewatiskan saluran vokal yang dimiliki oleh manusia. Otot-otot yang ada disekitar bagian mulut juga turut membantu pergerakan serta penempatan pengucapan. Beberapa instrumen penghasil suara adalah: Paru-paru, batang tenggorokan, Laring (organ produksi suara), rongga faring (tenggorokan), rongga mulut dan rongga hidung (Huang, et al. 2001).

Faring serta rongga mulut biasanya dikenal juga dengan istilah saluran vokal, sedangkan rongga hidung dikenal juga dengan istilah saluran hidung, Instrumen penghasil suara manusia Antara lain:

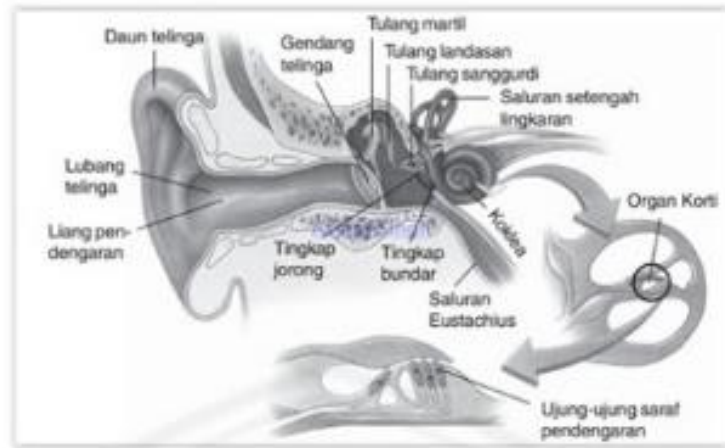
- a. Paru-paru: Sumber udara yang dikeluarkan ketika sedang berbicara.
- b. Pita suara (Laring/ Vocal Cords): lipatan vokal yang sangat rapat dan berisolasi satu sama lain ketika suara sedang dikeluarkan. Terdapat pada seluruh lipatan vokal yang menyatu satu sama lain, atau sering disebut juga dengan *glottis*.
- c. Langit-langit lunak (velum): berfungsi sebagai katup suara yang mampu membuka jalur udara masuk (sehingga bisa bersonansi) melalui rongga

hidung. Suara yang dihasilkan dengan lipatan terbuka termasuk pengucapan pada huruf “ m” dan huruf “n”.

- d. Langit-langit keras (Hard Palate): Permukaan panjang yang relatif keras dan terletak dibagian mulut paling atas. Ketika lidah ditempatkan pada bagian tersebut, pengucapan konsonan pun terjadi.
- e. Lidah: Anatomi pelafalan yang paling fleksibel. Ketika huruf hidup (vokal) diucapkan. Lidah ditempatkan pada bagian langit-langit, sedangkan pengucapan huruf mati (konsonan) lidah ditempatkan pada bagian permukaan yang keras.
- f. Gigi: Berfungsi untuk memperkokoh lidah ketika mengucapkan huruf-huruf konsonan tertentu.
- g. Bibir: Anatomi yang dapat dibulatkan atau dilebarkan untuk dapat menciptakan huruf vokal yang sempurna dan ditutup rapat-rapat agar dapat menghentikan aliran udara ketika mengucapkan huruf konsonan tertentu (p,b,m).

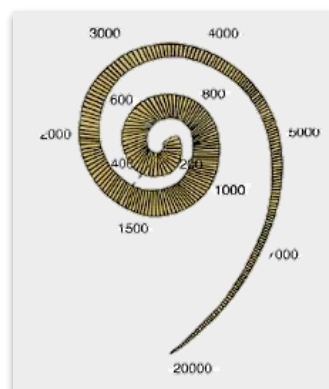
2.2 Pendengaran

Pendengaran manusia berada pada rentang frekuensi 20 Hz – 20,000 Hz, meskipun demikian frekuensi suara manusia ketika berbicara (speech) berada pada frekuensi 100Hz-4000Hz (National Institutes of Health, 2007). Frekuensi diatas 20.000Hz dianggap sebagai suara ultrasonik, meskipun suara ultrasonik berada pada jangkauan frekuensi yang dapat didengar oleh manusia, ada banyak hewan yang dapat mendengar pada frekuensi tersebut, misalnya, anjing yang dapat mendengar hingga frekuensi 50.000Hz dan kelelawar yang dapat mendengar hingga 100.000Hz. Suara yang berada dibawah frekuensi 20Hz disebut sebagai suara subsonik (National institutes of Health, 2007).



Gambar 2.1 Diagram Struktur Sistem Pendengaran Manusia

Diagram struktur sistem pendengaran manusia dapat dilihat pada gambar 2.1 Pinna (telinga) adalah permukaan yang mengelilingi kanal tempat suara disalurkan. Kanal sepanjang 2.5 cm ini juga berfungsi sebagai penguat sinyal (*Amplifier*) untuk suara dengan frekuensi 3.000- 4.000 Hz (National institutes of Health, 2007). Gelombang suara disalurkan oleh kanal melalui *erdrum* membran yang berperan sebagai pengubah gelombang akustik suara menjadi energi mekanik. Gelombang suara yang telah diterjemahkan menjadi getaran mekanik kemudian menuju koklea melawan serangkaian tulang yang disebut tulang kecil (*ossicles*) (Huang, et al, 2001). Tulang kecil (*ossicles*) sepanjang 20 mm ini juga berguna untuk memperkuat (*amplifier*) getaran sinyal (National institutes of Health, 2007).



Gambar 2.2 Frekuensi yang Terdeteksi oleh *Basilar Membran*

Koklea adalah organ yang berbentuk seperti siput. Getaran mekanik melalui *ossicles* menyebabkan internal membran, atau disebut juga dengan *basilar membrane*, bergetar pada berbagai macam set frekuensi di lokasi yang berbeda-beda, seperti terlihat pada gambar 2.3. *Basilar membrane* kemudian dikarakterisasi berdasarkan set-set frekuensi ini. Gerakan yang terjadi di *basilar membrane* ini kemudian dirasakan oleh sel-sel rambut yang menterjemahkan menjadi sinyal-sinyal yang dikirim ke otak (Huang, et al, 2001).

Setiap lokasi *basilar membrane* bereaksi nerbeda tergantung pada frekuensi dari gelombang suara yang diterima, sehingga sel-sel rambut yang berlokasi pada setiap titik yang berbeda sepanjang membran dirangsang oleh suara dari berbagai macam frekuensi. Saraf yang terhubung dengan sel rambut kemudian meneruskan spesifikasi frekuensi ini ke pusat pendengaran yang lebih tinggi. Berdasarkan aturan-aturan tersebut, sistem pendengaran manusia pada dasarnya berperilaku mirip sebagai penganalisa frekuensi, sehingga karakterisasi sistem pengenalan suara akan menjadi lebih sederhana jika dilakukan pada domain frekuensi (Huang, et al, 2001).

2.3 Digit Desimal

Digit desimal adalah bilangan dengan basis 10 yang terdiri dari angka 0 hingga 9. Digit diucapkan berbeda hampir dalam semua bahasa. Pengucapan digit untuk Bahasa Indonesia dapat dilihatpada Tabel 2.1 (Badan Pengembangan dan Pembinaan Bahasa, 2016)

Tabel 2.1 Pengucapan Digit Desimal Bahasa Indonesia

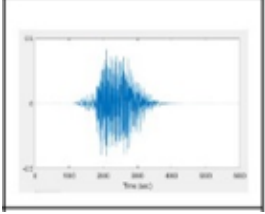
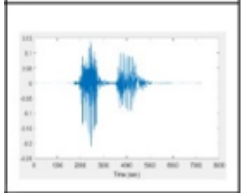
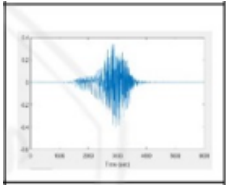
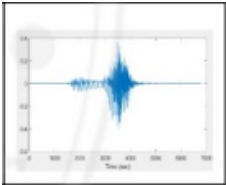
Digit	Indonesia	Pengucapan
0	Nol	e-nol
1	Satu	sa-tu
2	Dua	du-a
3	Tiga	ti-ga
4	Empat	em-pat
5	Lima	li-ma
6	Enam	e-nam
7	Tujuh	tu-juh
8	Delapan	de-la-pan
9	Sembilan	sem-bi-lan

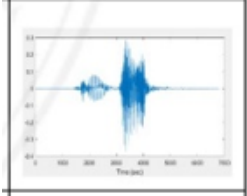
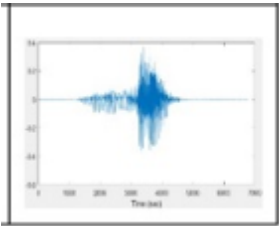
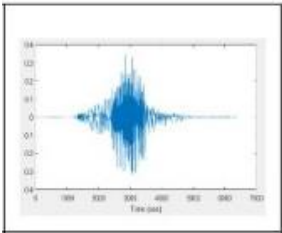
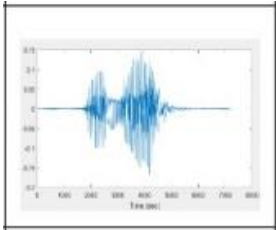
Sumber : Diolah penulis

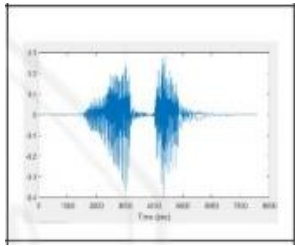
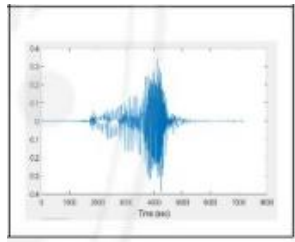
2.4 Sinyal Suara

Di dalam udara gelombang suara merambat sebagai variasi tekanan tinggi dan rendah yang disebut dengan amplitudo. Microphone menangkap sinyal suara dan diubah menjadi sinyal digital oleh *Analog to Digital Converter* (ADC). Sinyal-sinyal ini disimpan dengan mengukur tingginya amplitudo pada waktu tertentu (Jurafsky dan Martin, 2008). Banyaknya nilai sinyal yang diambil dalam satu detik disebut dengan sampling rate. Sampling rate yang digunakan dalam penelitian adalah 8000Hz, 16-bit untuk semua data latih dan data uji. Penggambaran amplitudo ini disebut dengan *wavefrom*. Nilai amplitudo berhubungan dengan volume keras atau pelan-nya suara yang diucapkan. Semakin tinggi nilai amplitudo menunjukkan semakin keras suara yang diucapkan.

Tabel 2.2 Amplitudo Digit 0-9

<p>Pengucapan kata “ Nol “, Nama file diberi label “-0” untuk menandai data suara digit</p>	 <p>A waveform plot showing a single, sharp, high-amplitude pulse centered around 2000 samples. The y-axis ranges from -0.2 to 0.2, and the x-axis is labeled 'Time (ms)' with ticks from 0 to 6000.</p>
<p>Pengucapan kata “ Satu “, Nama file diberi label “-1” untuk menandai data suara digit</p>	 <p>A waveform plot showing two distinct pulses. The first pulse is around 1500 samples and the second is around 3500 samples. The y-axis ranges from -0.2 to 0.2, and the x-axis is labeled 'Time (ms)' with ticks from 0 to 6000.</p>
<p>Pengucapan kata “ Dua “, Nama file diberi label “-2” untuk menandai data suara digit</p>	 <p>A waveform plot showing a single, sharp, high-amplitude pulse centered around 2000 samples. The y-axis ranges from -0.2 to 0.2, and the x-axis is labeled 'Time (ms)' with ticks from 0 to 6000.</p>
<p>Pengucapan kata “ Tiga “, Nama file diberi label “-3” untuk menandai data suara digit</p>	 <p>A waveform plot showing a single, sharp, high-amplitude pulse centered around 2000 samples. The y-axis ranges from -0.2 to 0.2, and the x-axis is labeled 'Time (ms)' with ticks from 0 to 6000.</p>

<p>Pengucapan kata “ Empat “, Nama file diberi label “-4” untuk menandai data suara digit</p>	
<p>Pengucapan kata “ Lima “, Nama file diberi label “-5 ” untuk menandai data suara digit</p>	
<p>Pengucapan kata “ Enam “, Nama file diberi label “-6 ” untuk menandai data suara digit</p>	
<p>Pengucapan kata “ Tujuh “, Nama file diberi label “-7 ” untuk menandai data suara digit</p>	

<p>Pengucapan kata “ Delapan “, Nama file diberi label “-8 ” untuk menandai data suara digit</p>	
<p>Pengucapan kata “ Sembilan “, Nama file diberi label “-9 ” untuk menandai data suara digit</p>	

Sumber: Diolah penulis

2.5 Spektogram

Pendengaran manusia pada dasarnya berperilaku mirip sebagai penganalisa frekuensi, sehingga karakterisasi sistem pengenalan suara akan menjadi lebih sederhana jika dilakukan pada domain frekuensi (Huang, et al. 2001). Penggambaran frekuensi pada sinyal suara disebut dengan spektogram. Sinyal suara terdiri dari campuran beberapa macam frekuensi yang berbeda menjadi satu. Untuk melihat frekuensi-frekuensi yang dihasilkan dalam satuan waktu tertentu, sinyal suara dapat diubah ke dalam domain frekuensi menggunakan alat bantu matematis yang dikenal sebagai *Fourier Transform*. Dengan mentransformasi sinyal menggunakan *Fourier* dapat dilihat frekuensi yang terdapat pada setiap frame sinyal. Joseph Fourier (1768-1830), menemukan bahwa setiap sinyal terdiri dari gabungan penjumlahan dari setiap kombinasi sinusoidal sinyal (Manolakis dan

Ingle. 2011). Formula Fourier untuk melihat sinyal dalam bentuk frekuensi seperti berikut:

$$X[k] = \sum_{n=0}^{N-1} x[n] e^{-2\pi \frac{ink}{N}}$$

Dimana :

e^{in} = $\cos n + i \sin n$

N = Jumlah sample input

X [k] = urutan ke-k komponen output DFT (X[0], X[1],,X[n-1])

K = indeks output DFT dalam domain frekuensi (0,1,....., n-1)

x[n] = urutan ke-n sample input (x[0],x[1],....., x[n-1])

n = indeks sample input dalam domain waktu (0,1,....., n-1)

Frekuensi suara (Pitch) unik setiap 20 milidetik (Huang, et al. 2001), karena itu pengubahan sinyal analog ke sinyal diskrit dilakukan pada setiap 20 milidetik. Pemotongan setiap 20 milidetik ini mengakibatkan adanya diskontinuitas tersebut pada proses pemotongan dilakukan proses khusus yang disebut dengan windowing. Hamming window paling sering digunakan untuk sinyal suara, formulasi *Hamming* diberikan sebagai berikut (Smith, 2011):

$$w[n] = 0.54 + 0.46 \cos\left(\frac{2\pi n}{N-1}\right)$$

Dimana ,

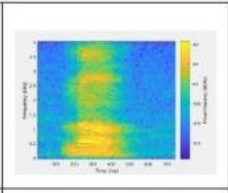
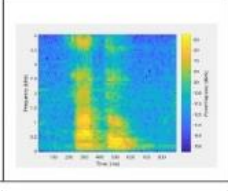
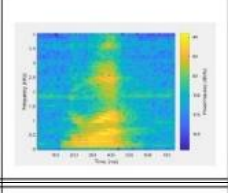
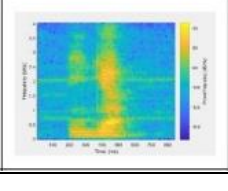
n = indeks sample input dalam domain waktu (0,1,, n-1)

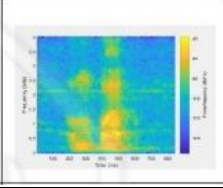
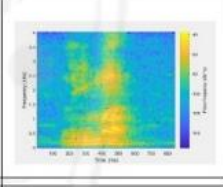
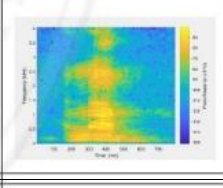
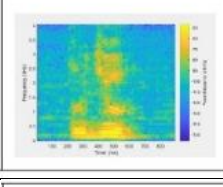
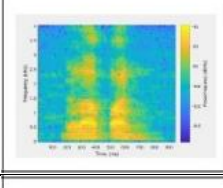
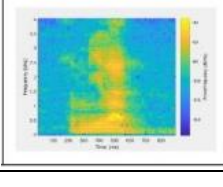
N = jumlah sample input

Gabungan koefisien *Fourier* setiap frame yang digabungkan menjadi satu, membuat penggambaran yang disebut dengan spektrogram. Dengan menggunakan formula *Fourier* ini, membuat setiap sinyal mempunyai spektrum sehingga sinyal suara dapat dianalisa dalam dominan frekuensi.

Serangkaian set frekuensi dari sinyal adalah spektrum sinyal. Nilai log absolut dari koefisien *Fourier* dapat diidentifikasi sebagai energi (positif dan negatif) frekuensi dari setiap frame sinyal suara (Osgood. 2009). Setiap sinyal suara yang dapat didengar manusia terdiri dari gabungan berbagai macam frekuensi, dan seberapa besar perbedaan frekuensi yang didengar manusia dihasilkan dari energi frekuensi yang ditangkap. Pada tabel 2.3 diberikan sinyal suara dalam bentuk spektrogram pengucapan digit 0-9 yang telah dipotong sesuai dengan pengucapan

Tabel 2.3 Spektrogram digit 0-9

Pengucapan kata 'Nol'. Nama file diberi label '-0' untuk menandai data suara digit	
Pengucapan kata 'Satu'. Nama file diberi label '-1' untuk menandai data suara digit	
Pengucapan kata 'Dua'. Nama file diberi label '-2' untuk menandai data suara digit	
Pengucapan kata 'Tiga'. Nama file diberi label '-3' untuk menandai data suara digit	

Pengucapan kata 'Empat'. Nama file diberi label '-4' untuk menandai data suara digit			
Pengucapan kata 'Lima'. Nama file diberi label '-5' untuk menandai suara digit			
Pengucapan kata 'Enam'. Nama file diberi label '-6' untuk menandai data suara digit			
Pengucapan kata 'Tujuh'. Nama file diberi label '-7' untuk menandai data suara digit			
Pengucapan kata 'Delapan'. Nama file diberi label '-8' untuk menandai data suara digit			
Pengucapan kata 'Sembilan'. Nama file diberi label '-9' untuk menandai data suara digit			

2.6 Mel-Frequency Ceptral Ceptrum (MFCC)

MFCC adalah representasi yang di definisikan sebagai nilai *real-cepstrum* dari *short-time* sinyal yang berasal dari spektrogram sinyal suara (Huang, et al, 2001), Perbedaan dari cepstrum adalah bahwa pada MFCC skala frekuensi nonlinier digunakan, yang mendekati perilaku sistem pendengaran manusia.

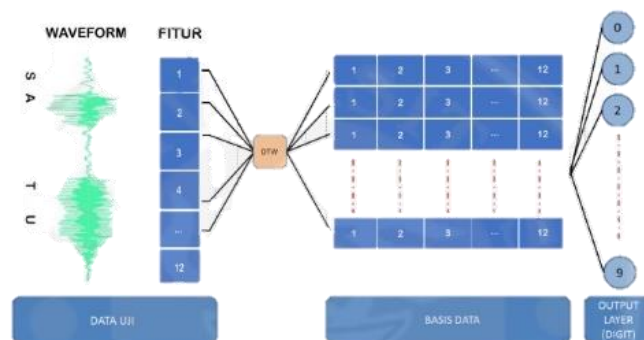
Spektrogram sendiri telah berhasil mendapatkan berbagai macam frekuensi yang ada dalam sinyal suara, namun mengikuti struktur koklea pada pendengaran

manusia yang menangkap frekuensi pada beberapa titik yang berbeda ditemukan formulasi yang mendekati perilaku sistem pendengaran manusia yang disebut dengan Skala mel (*mel-scale*). Skala mel pada dasarnya adalah melakukan filter pada koefisien frekuensi yang didapatkan dari spektrogram. Filter yang dilakukan adalah linier pada frekuensi 1000Hz dan logaritmik pada frekuensi di atasnya (O'shaughnessy, 2000).

2.7 Dynamic Time Warping (DTW)

DTW adalah algoritma untuk mencari kemiripan antara dua data yang time-sequence. Algoritma DTW awalnya dibuat khusus untuk menangani masalah perbedaan panjang data pada pengenalan suara (Sakoe dan Chiba, 1978). Algoritma DTW dapat mencari kemiripan antara dua data dengan panjang vektor berbeda, hal ini membuat algoritma DTW sangat cocok digunakan untuk data suara, karena data suara adalah data dengan panjang waktu yang berbeda-beda dalam setiap pengucapannya, walaupun pembicara yang sama mengucapkan digit yang sama.

Pada model template based, data suara yang di uji langsung dicari kemiripannya dengan data suara yang terdapat dalam basis data. Model pengenalan suara dengan DTW pada penelitian ini dapat dilihat pada gambar 2.3.



Gambar 2.3 Proses Pengenalan Suara dengan DTW

Untuk melakukan suara dengan DTW, data suara diubah menjadi spektrogram atau MFCC. Koefisien spektrogram atau koefisien MFCC ini kemudian dibandingkan dengan koefisien MFCC atau spektrogram yang terdapat

dalam basis data. Jarak DTW terdekat antara fitur yang terdapat dalam basis data merupakan klasifikasi dari suara yang diucapkan.

2.8 Hidden Markov Model (HMM)

Hidden Markov Model (HMM) adalah sebuah model untuk data yang saling berangkaian (*sequence data*). Model *sequence* atau klasifikasi *sequence* adalah sebuah model yang bertugas menentukan label/kelas pada setiap unit dalam data, atau memetakan antara data yang diinput dengan kelas yang akan diklasifikasi. Sebuah model HMM dapat juga dengan probabilitas model. Jika diberikan data, HMM akan menghitung distribusi probabilitas terhadap label/kelas yang ada dan kemudian menghitung label yang terbaik (Jurafsky dan Martin. 2008).

HMM merupakan model yang paling umum digunakan untuk pengenalan suara, karena akurasi yang dihasilkan mampu menyamai DTW dan menghilangkan ketergantungan terhadap basis data. Untuk mendapatkan model, langkah awal dilakukan dengan menghitung bobot observasi dengan algoritma *forward-backward*, menghitung *hidden state* dengan algoritma Viterbi, dan menghitung probabilitas bobot transisi dengan algoritma baum-welch pada setiap data input.

2.9 Long-Short Term Memory (LSTM)

Pembelajaran untuk menyimpan informasi yang penting menggunakan Recurrent Neural Network tidak pernah mencapai network dengan bobot optimal, hal ini terutama disebabkan oleh nilai *error* yang menghilang saat proses *backward* yang disebut dengan *vanishing gradient*. Hochreiter dan Schmidhuber pada tahun 1997 menganalisa masalah *vanishing gradient* ini dan mengusulkan arsitektur baru untuk Recurrent Neural Network yang disebut dengan Long-Short Term Memory. 4 buah hidden layer baru yang terdapat dalam LSTM Cell ditambahkan pada neural network dan disebut dengan gerbang (*gates*). Gerbang-gerbang ini berfungsi seperti neural network baru yang belajar untuk menyaring informasi yang dianggap penting.

Untuk melakukan pengenalan suara dengan *deep learning* menggunakan LSTM, ada dua langkah yang harus dilakukan. Langkah pertama adalah proses pembelajaran dan langkah kedua adalah klasifikasi atau pengujian. Pada proses pembelajaran berbasis data diubah menjadi spektrogram atau MFCC kemudian dilakukan proses pembelajaran LSTM untuk mendapatkan model LSTM yang optimal. Dalam memvalidasi model LSTM yang telah di-training, perlu dilakukan proses pengujian. Langkah-langkah pada pengujian sama dengan proses pembelajaran, namun menggunakan model LSTM yang telah dilakukan proses pembelajaran dan output berupa teks digit yang dikenali. Basis data yang digunakan adalah basis data uji yang berbeda dengan basis data pada saat proses pembelajaran

Proses pembelajaran Deep Learning - LSTM terdiri dari dua proses, forward-pass dan backward-pass. Forward pass menghitung output berdasarkan data yang diinput dengan nilai bobot-bobot neural (model) yang secara acak dibangkitkan dalam tahap awal pembelajaran. Nilai output ini pasti tidak akan sesuai dengan nilai output yang diharapkan. Nilai selisih antara output neural network dan output target ini kemudian dikembalikan (backward-pass) ke dalam neural network dengan mengubah nilai bobot-bobot neural agar pada iterasi berikutnya didapatkan nilai output yang mendekati dengan output yang diharapkan.

2.10 Penelitian LSTM untuk Pengenalan Suara

LSTM untuk bahasa Bengali belum pernah digunakan sebelumnya, penelitian untuk pengenalan berbahasa Bengali sebelumnya menggunakan HMM untuk memodelkan suara. LSTM kemudian digunakan untuk bahasa Bengali (Bangladesh) yang diteliti oleh Daneshvar dan Veisi pada tahun 2017. Pada model Daneshvar ini digunakan 2 layer LSTM tanpa Fully-Connected Layer dengan masing-masing berjumlah 100 LSTM-Cell pada setiap LSTM layer untuk mengenali kata dalam bahasa Bengali. Fitur ekstraksi ciri yang digunakan adalah

MFCC. Data suara yang mereka gunakan berjumlah 2000 kata, dari 15 pembicara dengan rentang usia 20-24 tahun.

Model Daneshvar ini melakukan proses pembelajaran berdasarkan fonem, ada 30 fonem dalam bahasa Bengali, sehingga output dari softmax layer yang digunakan berjumlah 30. Tingkat akurasi pengenalan mencapai 86.8%. Model LSTM yang dikembangkan dapat dilihat pada gambar 2.14 (Daneshvar dan Veisi. 2016).

Nahid, Purkaystha dan Islam pada tahun 2017 memodelkan multi-layer LSTM untuk mengenali bahasa Parsi dan membandingkannya dengan Hidden Markov Model (HMM). Mereka mencoba multi-layer dengan 3 layer LSTM. Model dengan masing-masing LSTM layer berjumlah 150 LSTM-Cell mendapatkan hasil akurasi pengenalan yang tertinggi untuk bahasa Parsi. Kesimpulan perbandingan penelitian dengan LSTM Nahid dan LSTM Daneshvar dapat dilihat pada tabel 2.4.

Tabel 2.4 Tinjauan Penelitian untuk Model LSTM

N o	Peneliti	Judul dan Metode	Basis Data	Hasil	Peluang Pengembangan
1	Md Mahadi Hasan Nahid, Bishwajit Purkaystha, Md Saiful Islam. 2017, IEEE	Bengali Speech Recognition : A Double Layered LSTM- RNN Approach.	15 orang laki-laki mengucapkan n 2000 kata, Speaker Independent	Tingkat akurasi pengenalan kata 86.8%	Model untuk bahasa Indonesia.

		Ekstraksi Ciri: MFCC.			
20th International Conference of Computer and Information Technology (ICCIT)					
2	Mohammad Danashvar and Hadi Veisi 2016	Persian Phoneme Recognition using Long Short-Term Memory Neural Network Ekstraksi Ciri : MFCC.	Farsdat Database, Persian speech database yang terdiri dari 6080 file wav diucapkan oleh 300 persian.	Speaker Independent . Tingkat Akurasi pengenalan kata 82,45%	Model untuk bahasa Indonesia
2016 International Conference on Information and Knowledge Technology (IKT).IEEE.					

2.11 Penelitian Pengenalan Suara Digit Berbahasa Indonesia

Penelitian pengenalan suara digit berbahasa Indonesia telah dilakukan pada tahun 2013 oleh Dewi, Firdausillah dan Supriyanto, menggunakan toolkit yang telah tersedia dari Carnegie Mellon University, yaitu CMU-Sphinx. Toolkit ini berbasis Hidden Markov Model dengan ekstraksi ciri menggunakan MFCC

(Lamere, et al. 2003). Data yang digunakan berjumlah 7 orang laki-laki yang mengucapkan digit 0 sampai digit 9 didalam ruangan yang bebas noise. Dalam proses perekaman ke-7 pembicara ini diharuskan berbicara dalam kondisi yang resmi dan masing-masing mengucapkan digit sebanyak 3 kali, sehingga terdapat 210 data suara. Untuk menguji sistem ASR yang dibuat menggunakan CMU-Sphinx ini, ketujuh pembicara yang sama mengucapkan digit ke sistem dan diharuskan mengucapkan dengan nada dan tekanan seperti pada saat mereka melakukan percakapan yang normal. Perbedaan intonasi, panjang pengucapan yang berbeda dan tekanan pengucapan menyebabkan sistem ASR hanya mampu mencapai tingkat akurasi 50% (Dewi, et al. 2013).

Penelitian untuk suara digit berbahasa Indonesia yang dilakukan pada tahun 2016 oleh Prakoso, Ferdiana dan Hartanto, juga menggunakan toolkit CMU-Sphinx. Data yang digunakan berjumlah 15 orang yang terdiri 8 orang laki-laki dan 7 orang perempuan mengucapkan digit 0 sampai digit 9 sebanyak dua kali didalam ruangan yang bebas noise, sehingga terdapat 300 data suara (Prakoso, et al. 2016). Untuk pengujian sistem ASR, 6 orang pembicara mengucapkan 54 digit yang disediakan dan diuji dalam kondisi noise yang berbeda-beda, tingkat akurasi dalam penelitian ini mencapai 86% (Prakoso, et al. 2016).

Tabel 2.5 Tinjauan Penelitian untuk Digit berbahasa Indonesia

No	Peneliti	Judul dan Metode	Basis Data	Hasil	Peluang Pengembangan
1	Ika Novita Dewi, Fahri Firdausilla h, Catur Supriyant	Sphinx -4 Indonesia Isolated Digit Speech Recognition Ekstraksi	7 Orang laki-laki mengucapkan digit 0-9. Testing dengan pembicara	Tingkat akurasi pengenalan kata 50%	Meningkatkan akurasi. Mengembangkan sistem yang speaker independent.

	o, 2013. IEEE	Ciri : MFCC HMM	yang sama (speaker Dependent)		
2013, Journal of Theoretical and Applied Information Technology					
2	Hamdan Prakoso, Ridi Ferdiana, Rudy Hartanto, 2016, IEEE.	Indonesian Automatic Speech Recognition System Using CMU Sphinx Toolkit and Limited Dataset, Ekstraksi Ciri: MFCC. HMM.	8 orang laki- laki dan 7 perempuan mengucapkan digit 0-9. Testing dengan 6 orang pembicara yang sama (Speaker Dependent)	Tingkat Akurasi pengenalan digit 86%	Meningkatkan akurasi. Mengembangka n sistem yang speaker independent.
2016 International Symposium on Electronics and Smart Devices (ISESD)					

3	J. M. T. S., D. Puspitaningrum, and B. Susilo, 2016, J.Rekrusif	Penerapan Speech Recognition Pada Permainan Teka-Teki Silang Menggunakan Metode Hidden Markov Model (HMM) Berbasis Desktop,	8 orang laki-laki dan 7 perempuan mengucapkan digit 0-9. Testing dengan 6 orang pembicara yang sama (Speaker Dependent)	Tingkat Akurasi pengenalan digit 70%	Meningkatkan akurasi. Mengembangkan sistem yang speaker independent.

4	Zaurarista Dyarbirru, Andi Sofyan Anas	Sistem Pengenalan Suara Digit dengan Metode Wavelet- MFCC dan Korelasi	8 orang laki-laki dan 7 perempuan mengucapk an digit 0-9. Testing dengan 6 orang pembicara yang sama (Speaker Dependent)	Tingkat Akurasi pengenal an digit 63%	Meningkatkan akurasi. Mengembang kan sistem yang speaker independent.

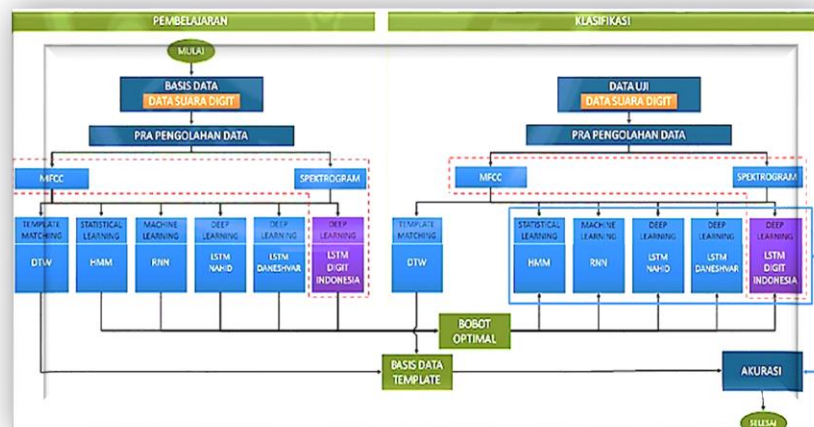
5	S. Amalia, 2017, J. Tek. Elektro ITP	Pengenalan Digit 0 Sampai 9 Menggunakan Ekstraksi Ciri MFCC dan Jaringan Syaraf Tiruan Backpropagati on,	10 orang yang berbeda mengucapk an digit 0-9. (Speaker Dependent)	Tingkat Akurasi pengenal an digit 82.2%	Meningkatkan akurasi. Mengembang kan sistem yang speaker independent.
2017 JURNAL IPTEK TERAPAN Research of Applied Science and Education					

BAB III

METODOLOGI PENELITIAN

3.1 Kerangka Metode Penelitian yang diusulkan

Penelitian ini bertujuan untuk mendapatkan model arsitektur LSTM untuk suara digit desimal berbahasa Indonesia sehingga tingkat akurasi pengenalan dapat lebih tinggi dibandingkan dengan penelitian sebelumnya. Akurasi atau tingkat akurasi pengenalan adalah jumlah ketepatan suara digit desimal yang diucapkan dibagi dengan jumlah data suara yang akan diterjemahkan menjadi teks, sehingga semakin tinggi akurasi pengenalan menunjukkan semakin tingginya hasil ketepatan sistem menterjemahkan sinyal-sinyal suara ini menjadi teks. Model LSTM telah diterapkan untuk pengenalan suara berbahasa Bengali (Nahid, et al. 2017) (LSTM Nahid) dan Parsi (Daneshvar dan Veisi. 2016) (LSTM Daneshvar), kedua model ini diteliti untuk dibandingkan tingkat akurasi pengenalan dengan model yang diusulkan untuk mengenali digit desimal bahasa Indonesia. Kerangka metode yang diusulkan diilustrasikan pada gambar 3.1.



Gambar 3.1 Kerangka Metode Penelitian

3.2 Pengumpulan Data Latih Dan Data Uji

Tahapan penelitian dilakukan seperti yang terlihat pada gambar 3.1, meliputi tahap pengumpulan data, tahap pra-pengolahan data yang mengekstraksi fitur suara menggunakan spektrogram atau MFCC, tahap pembelajaran untuk mencari model, tahap klasifikasi untuk menguji model yang dihasilkan dan evaluasi dengan menghitung akurasi pengenalan. Pada tahap pengumpulan data, dilakukan perekaman data suara digit desimal berbahasa Indonesia mulai dari digit 0 hingga digit 9, kemudian dilakukan pemotongan data suara sesuai dengan digit yang diucapkan dan dilanjutkan dengan pemberian label untuk setiap file digit yang disimpan. Tahap berikutnya adalah mentransformasi sinyal suara menjadi matriks koefisien spektrogram dan mengambil ciri data suara tersebut menjadi matriks koefisien MFCC. Proses pembelajaran dilakukan dengan menggunakan DTW, HMM, RNN, model LSTM Nahid, model LSTM Daneshvar dan model LSTM yang diusulkan terhadap matriks koefisien spektrogram dan MFCC. Untuk melihat tingkat akurasi, pengujian kemudian dilakukan terhadap DTW, HMM, RNN, model LSTM Daneshvar, model LSTM Nahid dan model LSTM yang diusulkan.

Data digit suara 0-9 yang digunakan dalam penelitian ini diambil di dalam ruang kelas ujian komputer Universitas Gunadarma. Proses perekaman dilakukan dalam jangka waktu 4 minggu, sehingga didapatkan sekitar 799 mahasiswa. Pada minggu pertama perekaman, didapatkan 100 pembicara masing-masing mengucapkan digit 0 sampai 9 sehingga terdapat 1000 data suara. 1000 data ini kemudian dilakukan proses klasifikasi LSTM menggunakan fitur spektrogram dengan 900 data latih dan 100 data uji, akurasi yang dihasilkan masih rendah sekitar 60%. Minggu kedua dilakukan kembali perekaman data dengan total data yang didapatkan 3000 data suara. Klasifikasi LSTM yang dihasilkan dari 3000 data ini meningkat menjadi 70%. Perekaman dilanjutkan dengan mendapatkan 5000 data suara dari 500 pembicara, hasil klasifikasi LSTM meningkat menjadi 80%. Pada saat 6000 data suara diklasifikasi oleh LSTM, akurasi sudah mencapai 90%, begitu juga dengan 7000 data suara. Perekaman data kemudian dilanjutkan di sisa minggu ke-empat sehingga didapatkan 7990 data suara dari 799 pembicara. Pembicara

adalah mahasiswa Gunadarma dengan rentang usia 19-22 tahun, terdiri dari 389 perempuan dan 410 laki-laki. Data suara yang diambil masing-masing mengucapkan digit dengan pelafalan dapat dilihat pada tabel 2.1, sehingga total basis data suara yang didapatkan sebanyak 7990 data. Data ini kemudian disebut dengan data latih.

Proses perekaman dilakukan dengan telepon genggam *iPhone 6* menggunakan aplikasi *voice memos* yang terdapat pada telepon genggam, setiap mahasiswa mengucapkan digit 0 hingga digit 9 sekaligus dengan jeda antara setiap digit, kemudian data tersebut disimpan. Perekaman terjadi dalam lingkungan pengucapan yang terdapat suara AC, suara pintu tertutup dan terbuka, suara mahasiswa lain batuk, suara keyboard, walaupun tidak disemua data. Semua *noise-noise* ini pada *background* sinyal ikut terekam oleh *microphone* telepon genggam. Data suara tersebut kemudian di-transfer ke *laptop* dan disimpan dalam format wav, dengan *sample-rate* 8000Hz, karena suara manusia berbicara berada pada rentang 34- 3400Hz, dan menurut *Nyquist Theorem* harus mengambil s dua kali dari rentang paling tinggi, maka *sample-rate* yang ditentukan adalah 8000Hz.

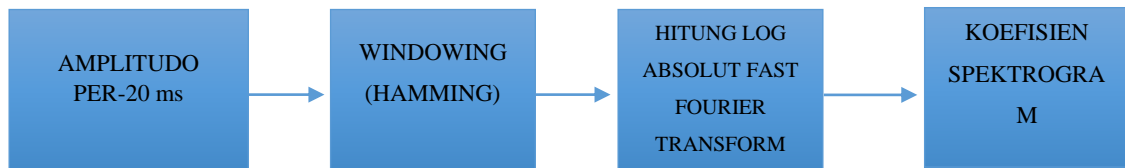
Data-data tersebut lalu dipotong sesuai dengan digit yang diucapkan, masing-masing digit disimpan dengan satu file dan di-labelkan sesuai dengan digit yang diucapkan. Pengucapan kata “no!” dari data suara mahasiswa A, data ini akan disimpan dalam satu file dengan diberi label seperti “A-O”, sehingga terdapat 10 file pengucapan digit untuk satu mahasiswa. *Noise-noise* yang ada, tidak dihilangkan dari sinyal suara. Untuk melihat tingkat akurasi diambil, sekitar 10 % data dari data latih. Suara 79 mahasiswa menjadi data uji, terdiri dari 37 perempuan dan 42 laki-laki yang tidak terdapat dalam data latih, masing-masing mengucapkan 10 digit (0-9), sehingga total data untuk pengujian berjumlah 790 data. Proses perekaman dilakukan dengan cara yang sama seperti pada pembentukan data latih. Dengan proses pengumpulan basis data seperti ini, didapatkan model yang *speaker independent*, karena pembicara pada data latih berbeda dengan pembicara pada data uji.

3.3 Pra Pengolahan Data

Pada tahap ini, data latih dan data uji, diubah menjadi spektrogram dan fitur MFCC. Data suara yang masih berbentuk amplitudo sangat sulit langsung digunakan untuk pengenalan suara. Nilai-nilai amplitudo tersebut sangat berfluktuasi tergantung dari tekanan pengucapan, kualitas *microphone*, *noise*, dan lain-lain. Algoritma untuk mengekstraksi ciri dari nilai-nilai amplitudo ini kemudian dikembangkan agar mempermudah proses pengenalan sinyal suara. MFCC digunakan untuk mengekstraksi ciri suara pada penelitian ini, karena hampir semua penelitian dengan DTW, HMM dan RNN menggunakan fitur MFCC pada digit bahasa Jepang (Sakoe dan Chiba. 1978), pada digit bahasa Spanyol Gómez. 2005), pada digit bahasa Arab (Alotaibi et al. 2010), pada digit bahasa Kroasia (Gulić, Lučanin dan Šimić. 2011), pada digit bahasa Hindi (Dhandhanian, Hansen, Kandi dan Ramesh. 2012; Saxena dan Wahi. 2015), pada digit bahasa Myanmar (Tun dan Srijuntongsiri. 2016), dan pada digit bahasa Gujarati (Pandit dan Bhatt. 2014). Spektrogram mulai digunakan oleh (Sak, et al. 2014) untuk pemodelan dengan LSTM dengan harapan agar LSTM dapat menemukan sendiri model akustik dari pembicara tanpa perlu fitur ekstraksi ciri seperti MFCC, sehingga spektrogram juga digunakan pada penelitian pengenalan suara digit desimal berbahasa Indonesia ini saat proses pra-pengolahan data.

3.3.1 Spektrogram

Sinyal suara dapat dilihat dalam domain frekuensi yang menggambarkan kerapatan spektrum sinyal atau disebut dengan spektrogram. Untuk merubah sinyal dari domain waktu (amplitudo) ke domain frekuensi (spektrogram) dapat dilakukan dengan langkah seperti pada gambar 3.2. Sinyal suara dipotong per 20 milidetik, lalu dilakukan *windowing* dengan menggunakan metode hamming agar didapatkan rentang nilai magnitude di dalam sinyal, kemudian data yang telah di-*windowing* ini ditransformasi *Fourier* dan menghitung magnitude *Fourier* (log absolut). Untuk setiap 20 milidetik didapatkan 64 koefisien spektrogram.



Gambar 3.2 Spektrogram

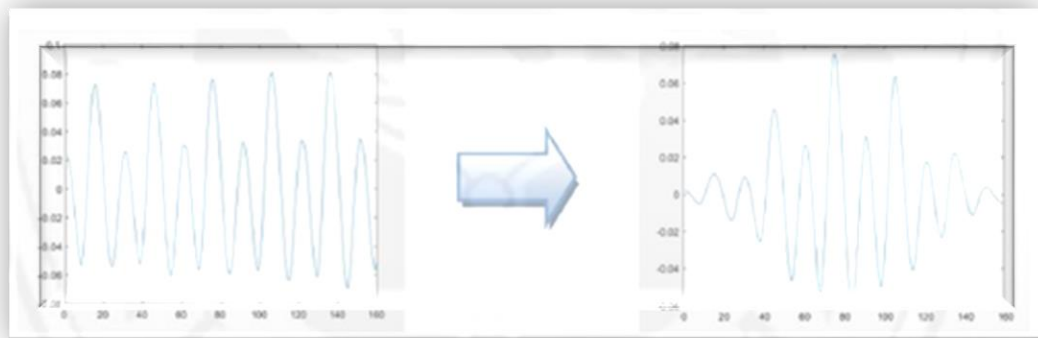
Pemotongan 20 milidetik atau 0.02 detik pada sample rate 8000Hz menghasilkan 160 sample. Jumlah sample didapatkan dari formula berikut:

Sample = sample-rate x time (*detik*)

=8000 x 0.02

=160

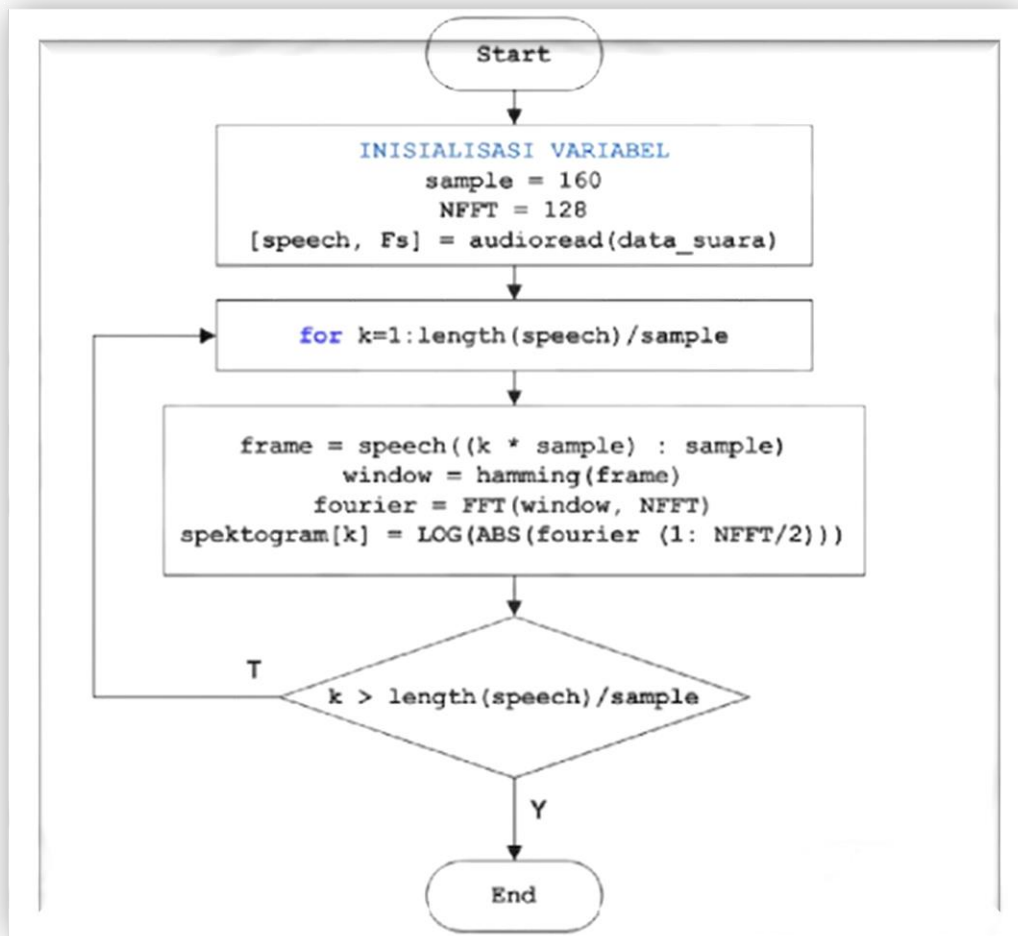
Pada gambar 3.3 diperlihatkan hasil pemotongan dari 20 milidetik atau 160 sample sinyal suara setelah dilakukan proses *windowing* dengan menggunakan *hamming*.



Gambar 3.3 Windowing (*Hamming*)

Koefisien magnitude dari *Fourier* kemudian didapatkan dari sinyal suara yang telah di *windowing*. Nilai FFT size yang digunakan adalah 128, sehingga didapatkan 64 (128 dibagi 2) koefisien *Fourier*. Untuk data suara dengan *sample rate* 8000Hz, nilai fft size 128, 256, 512 tidak terlalu berpengaruh, sehingga bisa menggunakan nilai fft *size* yang paling kecil agar proses transformasi fourier lebih cepat.

Flowchart mengubah sinyal amplitudo menjadi spektrogram pada Matlab dapat dilihat pada gambar 3.4.



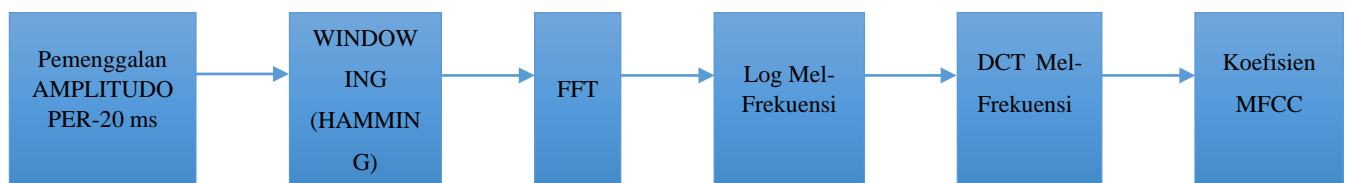
Gambar 3.4 Flowchart Spektrogram

3.3.2 Mel-Frequency Ceptral Ceptrum (MFCC)

Ekstraksi ciri pada dasarnya adalah mengubah sinyal-sinyal amplitudo atau spektrogram menjadi hanya beberapa vektor koefisien yang diperkirakan mengandung informasi yang penting. Mel-Frequency Cepstrum Coefficient (MFCC) yang dikenalkan oleh Davis dan Mermelstein pada tahun 1990, mengambil sekitar 12 vektor frekuensi dari sinyal suara. Koefisien MFCC ini merupakan fitur ekstraksi ciri suara paling sering digunakan karena menunjukkan akurasi yang

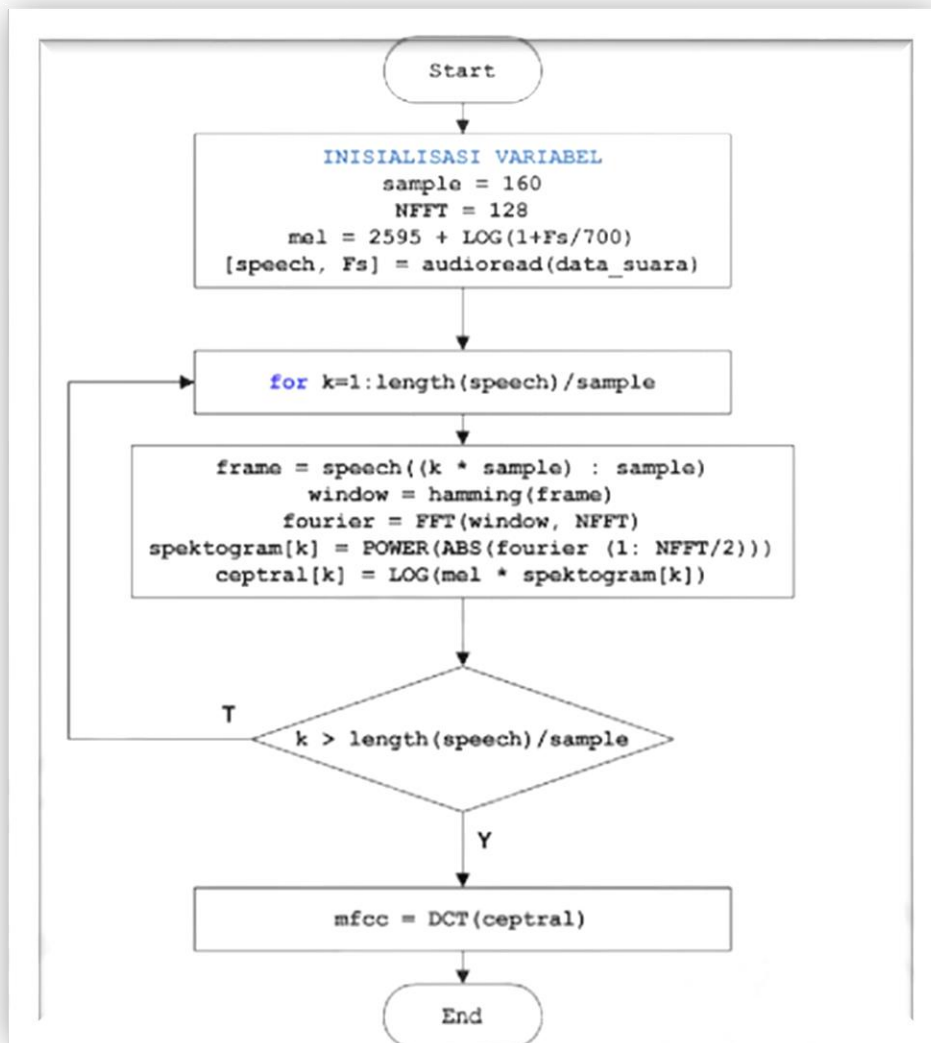
tinggi pada sinyal yang terdapat *noise* (Davis dan Mermelstein. 1990). Pada penelitian ini MFCC juga digunakan sebagai fitur ciri.

Langkah awal untuk mencari koefisien MFCC mirip dengan spektrogram, namun setelah mendapatkan nilai koefisien fourier, ditambahkan menghitung log mel-frekuensi dan mencari ceptral koefisien dengan menghitung *Discrete Cosinus Transform* (DCT) dari log mel-frekuensi. Langkah lengkap dapat dilihat pada gambar 3.5.



Gambar 3.5 Alur MFCC

Pada gambar 3.5, menunjukkan plot hasil 64 koefisien FFT kemudian dilakukan perhitungan mel-frekuensi dengan 12 parameter. Flowchart mengekstraksi ciri suara dengan MFCC pada Matlab dapat dilihat pada gambar 3.6.

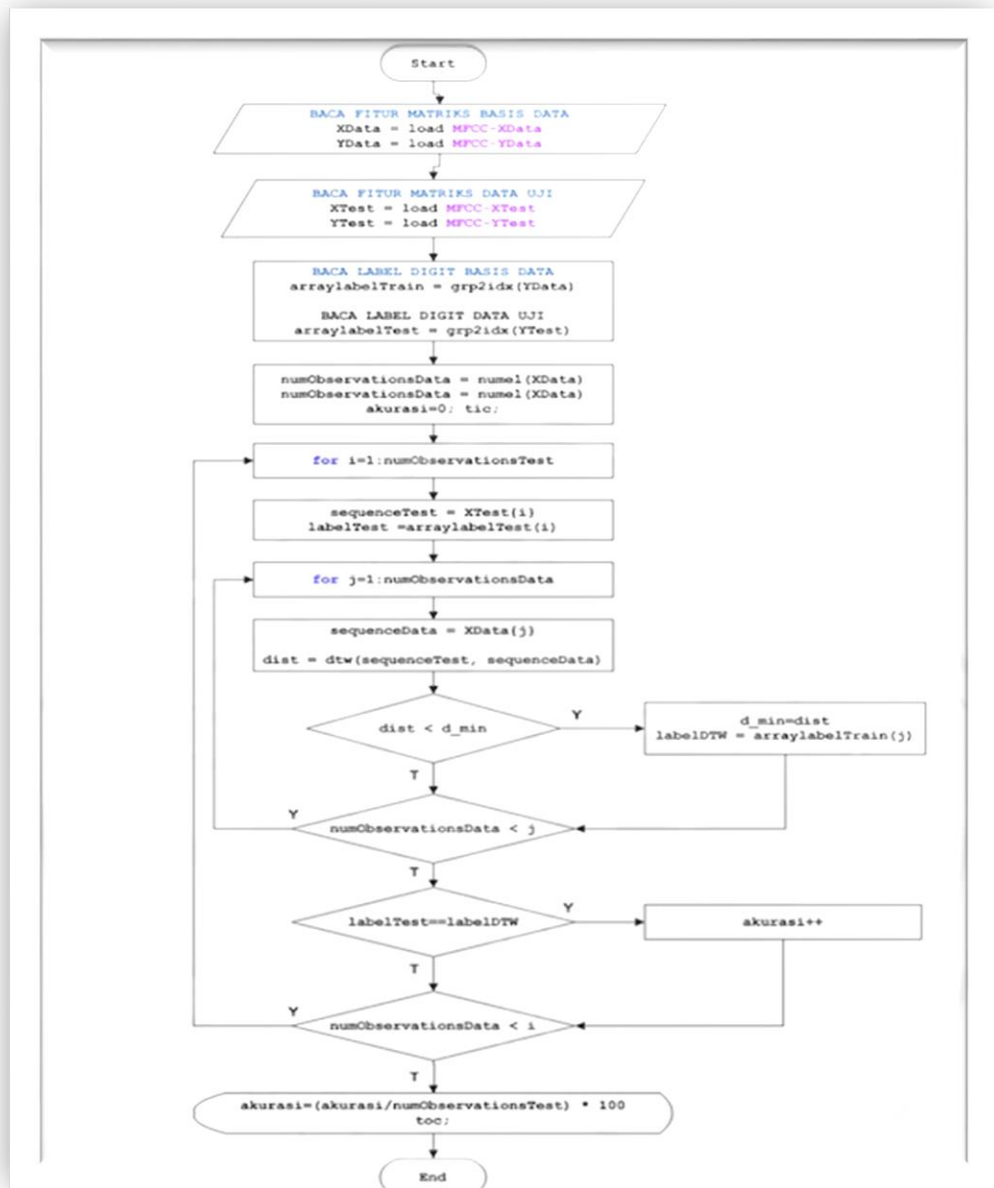


Gambar 3.6 Flowchart MFCC

3.4 Dynamic Time Warping (DTW)

Untuk melakukan pengenalan suara dengan algoritma DTW, sebelumnya dilakukan pengambilan fitur ciri spektrogram atau MFCC dari 7990 data latih, begitu juga dengan 790 data uji. Koefisien spektrogram atau MFCC tersebut kemudian yang dijadikan data matriks untuk pengenalan dengan DTW. Terdapat 12 koefisien yang dihasilkan jika menggunakan MFCC, dan 64 koefisien jika menggunakan spektrogram.

DTW membuat *distance* matriks untuk me-normalisasi perbedaan ukuran antara matriks yang akan dibandingkan, sehingga jika matriks data uji (A) menggunakan MFCC berukuran 83x12 dan matriks data latih (B) berukuran 52x12, maka *distance* matriks DTW akan berukuran 83x52. *Distance* matriks DTW menggunakan spektrogram juga berukuran yang sama, jika matriks data uji (A) berukuran 83x64 dan matriks data latih (B) berukuran 52x64, maka *distance* matriks DTW akan berukuran 83x52. Flowchart pengujian dengan DTW dapat dilihat pada gambar 3.7.



Gambar 3.10 Flowchart DTW

Setiap sel *distance* matriks (D) berisi kombinasi antara jarak euclidean dan bobot antara sel A dan sel B. Proses ini dihitung hingga semua sel matriks D terisi. Jarak DTW antara matriks A dan matriks B kemudian didapatkan dari sel matriks D yang berada di sel terakhir dari *distance* matriks atau di sel pada baris 83 dan kolom 52.

DAFTAR PUSTAKA

- McLoughlin, I. (2009). *Applied Speech and Audio Processing: with Matlab Examples*. Cambridge University Press.
- Saksono, Muh. Widyanto Tri, Achmad Hidayatno, dan Ajub Ajulian Z. 2008. *Aplikasi Pengenalan Ucapan Sebagai Pengatur Mobil Dengan Pengendali Jarak Jauh*. http://eprints.undip.ac.id/4310/1/mar08_t05_ucapan_ayub.pdf.
- Jurafsky, D., dan Martin, J. H. (2008). *Speech and Language Processing (Prentice Hall Series in Artificial Intelligence)*. Prentice Hall.
- Chu, W. C. (2003). *Speech Coding Algorithms: Foundation And Evolution of Standardized Coders*. A John Wiley & Sons Inc.
- Chavan, M. R. S. dan Sable, G. S. (2013). An Overview Of Speech Recognition Using HMM. *International Journal of Advanced Research in Electrical, Electronics and Instrumentation Engineering (IJAREEIE)*, 2(6):233,238.
- Davis, S. B. dan Mermelstein, P. (1990). Comparison of Parametric Representations for Monosyllabic Word Recognition In Continuously Spoken Sentences. *Readings in Speech Recognition*, pages 65–74. Elsevier.
- Stevens, K. N. (2000). *Acoustic phonetics, volume-30. MIT press*.
- Ningthoujam N., dan Prathima V. R. (2016). A Survey On Feature Extraction Algorithm for The Speech Recognition System. *International Journal of Computer Science and Mobile Computing*, 5(4).
- Rabiner, L. R. & Juang, B. H. (1986). *An Introduction to Hidden Markov Model*. IEEE ASSP Magazine 0740-7467/86/0100-0004\$01.00©1986 IEEE
- Aggarwal, A., Sahay, T., dan Chandra, M. (2015). *Performance Evaluation of Artificial Neural Networks for Isolated Hindi Digit Recognition with LPC*

And MFCC. International Conference on Advanced Computing and Communication Systems, 2015, pages 1-6. IEEE.

Al-Haddad, S. A. R., Samad, S. A., Hussain, A., Ishak, K. A., dan Mirvaziri, H. (2007). Decision Fusion for Isolated Malay Digit Recognition Using Dynamic Time Warping (DTW) And Hidden Markov Model (HMM). SCORED 2007. 5th Student Conference on Research and Development, 2007., pages 1-6. IEEE.

Ali, H., Jianwei, A., dan Iqbal, K. (2015). Automatic Speech Recognition of Urdu Digits with Optimal Classification Approach. International Journal of Computer Applications, 118(9).

Chapaneri, S. V. dan Jayaswal, D. J. (2013). Efficient Speech Recognition System for Isolated Digits. International Journal Computer Science and Engineering Technologies, 4(3):228–236.

Cucu, H., Caranica, A., Buzo, A., dan Burileanu, C. (2015). On Transcribing Informally-Pronounced Numbers In Romanian Speech. 38th International Conference on Telecommunications and Signal Processing (TSP) 2015, pages 372–376. IEEE.

Darabkh, K. A., Khalifeh, A. F., Bathech, B. A., dan Sabah, S. W. (2013). Efficient DTW-Based Speech Recognition System for Isolated Words of Arabic Language. Proceedings of International Conference on Electrical and Computer Systems Engineering (ICECSE 2013), pages 689–692. Citeseer.

Dewi, I. N., Firdausillah, F., dan Supriyanto, C. (2013). Sphinx-4 Indonesian Isolated Digit Speech Recognition. Journal of Theoretical & Applied Information Technology, 53(1).

Dhandhanian, V., Hansen, J. K., Kandi, S. J., dan Ramesh, A. (2012). A Robust Speaker Independent Speech Recognizer for Isolated Hindi Digits. International Journal of Computer and Communication Engineering, 1(4):483.

Dixit, A., Vidwans, A., dan Sharma, P. (2016). Improved MFCC And LPC Algorithm for Bundelkhandi Isolated Digit Speech Recognition. International Conference on Electrical, Electronics, and Optimization Techniques (ICEEOT), pages 3755–3759. IEEE.

Graves, A., Mohamed, A. R., dan Hinton, G. (2013). Speech Recognition with Deep Recurrent Neural Networks. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2013, pages 6645–6649. IEEE.

Gulić, M., Lučanin, D., dan Šimić, A. (2011). A Digit And Spelling Speech Recognition System for The Croatian Language. Proceedings of the 34th International Convention MIPRO, 2011, pages 1673–1678. IEEE.

Hachkar, Z., Farchi, A., Mounir, B., dan El-Abbadi, J. (2011). A Comparison Of

- DHMM And DTW for Isolated Digits Recognition System of Arabic Language *International Journal on Computer Science and Engineering*, 3(3):1002–1008.
- Hochreiter, S., dan Schmidhuber, J. (1997). Long Short-Term Memory. *Neural Computation*, 9(8):1735–1780.
- Lamere, P., Kwok, P., Gouvea, E., Raj, B., Singh, R., Walker, W., Warmuth, M., dan Wolf, P. (2003). The CMU Sphinx-4 Speech Recognition System. *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP 2003)*, Hong Kong, volume-1, pages 2–5.
- Limkar, M., Rao, R., dan Sagvekar, V. (2012). Isolated Digit Recognition Using MFCC And DTW. *Mumbai University, India*, 1:59–64.
- Mukhedkar, A. S., dan Alex, J. S. R. (2014). Robust Feature Extraction Methods for Speech Recognition In Noisy Environments. *First International Conference on Networks & Soft Computing (ICNSC), 2014*, pages 295–299. IEEE.
- Sakoe, H., dan Chiba, S. (1978). Dynamic Programming Algorithm Optimization for Spoken Word Recognition. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 26(1):43–49.
- Sakoe, H., Isotani, R., Yoshida, K., Iso, K., dan Watanabe, T. (1990). Speaker Independent Word Recognition Using Dynamic Programming Neural Networks. Readings in Speech Recognition, pages 439–442. Elsevier.
- Terissi, L. D., dan Gómez, J. C. (2005). Template-Based And Hmm-Based Approaches For Isolated Spanish Digit Recognition. *Inteligencia Artificial. Revista Iberoamericana de Inteligencia Artificial*, 9(26).
- Pandit, P., dan Bhatt, S. (2014). Automatic Speech Recognition of Gujarati Digits Using Dynamic Time Warping. *International Journal of Engineering and Innovative Technology*, 3(12).
- Prakoso, H., Ferdiana, R., dan Hartanto, R. (2016). Indonesian Automatic Speech Recognition System Using CMU-Sphinx Toolkit And Limited Dataset. *International Symposium on Electronics and Smart Devices (ISESD)*, pages 283–286. IEEE.
- Yi, J., Ni, H., Wen, Z., Liu, B., dan Tao, J. (2016). CTC Regularized Model Adaptation For Improving LSTM RNN Based Multi-Accent Mandarin Speech Recognition. *2016 10th International Symposium on Chinese Spoken Language Processing (ISCSLP)*, pages 1–5. IEEE.