Speech Recognition Menggunakan Algoritma Program Dinamis

Ananti Selaras Sunny

Program Studi Teknik Informatika
Sekolah Teknik Elektro dan Informatika, Institut Teknologi Bandung
e-mail: if17009@students.if.itb.ac.id

ABSTRAK

***Speech recognition* yang dikenal sebagai *automatic speech recognition* atau *computer speech recognition* menerjemahkan perkataan yang diucapkan menjadi text. Teknologi *speech recognition* ini sudah ada sejak lama dan sekarang banyak sekali jenis aplikasi yang dikembangkan menggunakan teknologi ini. Teknologi ini dikembangkan menggunakan algoritma program dinamis dalam hal ini adalah DTW singkatan dari *Dynamic Time Warping*. Di dalam sistem *speech recognition* mengandung kata-kata dan pengenalan kata-kata tersebut membutuhkan perbandingan antara sinyal masuk dari kata dan bermacam-macam kata yang ada di dalam kamus. Dan permasalahan seperti ini akan mudah diselesaikan secara efisien dengan menggunakan DTW dengan mengukur kesamaan antara dua sekuensial pada waktu yang berbeda baik dari segi kecepatannya. Agoritma DTW ini diimplementasikan pada video, audio, dan grafik dan tentu saja data-data bisa diubah ke dalam bentuk representasi linear yang bisa dianalisis oleh DTW. Maka, dalam makalah ini akan dibahas tentang algoritma program dinamis yang diterapkan dalam *Speech recognition*, yaitu algoritma *Dynamic Time Warping*.**

Kata kunci: *speech recognition*, *dynamic time warping*, program dinamis.

1. PENDAHULUAN

Speech recognizer pertama kali muncul di tahun 1952 dan terdiri dari *device* untuk pengenalan satu digit yang diucapkan. Kemudian pada tahun 1964, muncul IBM Shoebox.

Salah satu teknologi yang cukup terkenal di Amerika dalam bidang kesehatan adalah *Medical Transcriptionist* (MT) merupakan aplikasi komersial yang menggunakan *speech recognizer.* Dan sampai sekarang banyak aplikasi yang dikembangkan menggunakan *speech recognizer*, antara lain di bidang kesehatan terdapat MT, di bidang militer terdapat High-performance fighter aircraft, Training air traffic controllers, sampai pada alat yang membantu orang-orang yang memiliki kesulitan dalam menggunakan tangan, maka diciptakannya komputer yang dapat dioperasikan menggunakan deteksi pengucapan user.

Sebenarnya ada dua pemodelan dasar untuk *speech recognition* ini yaitu *Hidden Markov model (HMM)-based speech recognition* dan *Dynamic time warping (DTW)-based speech recognition*.

Modern general-purpose speech recognition system umumnya menggunakan model *Hidden Markov*. Model ini merupakan model yang statistikal dimana output adalah sekuens dari simbol atau kuantitas. Satu alasan yang mengapa model *Hidden Markov* digunakan, karena sebuah sinyal dari pengucapan bisa dilihat seperti *piecewise stationary signal* atau *short-time stationary signal*. Alasan lainnya mengapa metode ini populer, sederhana dan secara komputasional bisa digunakan.

Dynamic time warping adalah pendekatan yang pernah sejarahnya digunakan untuk *speech recognition* yang sekarang sudah digantikan oleh model *Hidden Markov.*

Pada pengembangannya maka alat *speech recognizer* diimplementasikan menggunakan *Dynamic Time Wraping Algorithm* (DTW). DTW pertama kali dikenalkan pada tahun 60an dan dieksplorasi sampai tahun 70an yang menghasilkan alat *speech recognizer.* DTW sering digunakan dalam area: *handwriting and online signature matching*, *sign language recognition and gestures recognition, data mining and time series clustering*, *computer vision and computer animation*, *surveillance*, *protein sequence alignment and chemical engineering*, dan *music and signal processing.*

Dan pada makalah kali ini hanya akan membahas implementasi algoritma DTW pada speech recognition.

2. METODE

2.1 Program Dinamis

Program dinamis adalah metode pemecahan maslah dengan cara menguraikan solusi menjadi sekumpulan langkah (*step*) atau tahapan (*stage*) sedemikian sehingga

solusi dari persoalan dapat dipandang dari serangkain keputusan yang saling berkaitan.
Pada penyelesaian persoalan dengan metode ini:

1. Terdapat sejumlah berhingga pilihan yang mungkin,
2. Solusi pada setiap tahap dibangun dari hasil solusi tahap sebelumnya,
3. Menggunakan pesyaratan optimasi dan kendala untuk membatasi sejumlah pilihan yang harus dipertimbakan dalam satu tahap.

Pada kasus *speech recognition* ini memenuhi persoalan pada poin ketiga, yaitu menggunakan pesyaratan optimasi dan kendala untuk membatasi sejumlah pilihan yang harus dipertimbakan dalam satu tahap.

Pada program dinamis, rangkaian keputusan yang optimal dibut dengan menggunakan **prinsip optimalitas**. Prinsip optimalitas: *jika solusi total optimal, maka bagian solusi sampai tahap ke-k juga optimal.* Prinsip optimalitas berarti bahwa jika kita bekerja dari tahap ke *k* ke tahap *k*+1, dapat menggunakan hasil optimal dari tahap *k* tanpa harus kembali ke tahap awal. Ongkos pada tahap *k*+1 sama dengan ongkos yang dihasilkan pada tahap *k* ditambahkan dengan ongkos dari tahap *k* ke tahap *k*+1.

Dengan prinsip optimalitas ini dijamin bahwa pengambilan keputusan pada suatu tahap adalah keputusan yang benar untuk tahap-tahap selanjutnya. Pada metode greedy hanya satu rangkaian keputusan yang pernah dihasilkan, sedangkan pada metode program dinamis lebih dari satu rangkaian keputusan. Hanya rangkaian keputusan yang memenuhi prinsip optimalitas yang akan dihasilkan.

Ada dua pendekatan program dinamis, yaitu maju (*forward* atau *up-down*) dan mundur (*backward* atau *bottom-up*). Misalkan a_1, a_2, ..., a_m menyatakan peubah (*variable*) keputusan yang harus dibuat masing-masing untuk tahap 1, 2, ..., n. Maka:

1. Program dinamis maju. Program dinamis bergerak mulai tahap 1, terus maju ke tahap 2,3, dah seterusnya sampai tahap *n*. Runtunan peubah keputusan adalah a_1, a_2, ..., a_n.
2. Program dinamis mundur. Program dinamis bergerak mulai tahap $n - 1$, $n - 2$, dan seterusnya sampai tahap 1. Runtunan peubah keputusan adalah a_n, a_{n-1}, ..., a_1

Langkah-langkah pengembangan algoritma program dinamis:

1. Karakteristikkan struktur solusi optimal
2. Definisikan secara rekursif nilai solusi optimal.
3. Hitung nilai solusi optimal secara maju atau mundur.
4. Konstruksi solusi optimal.

2.2 *Dynamic Time Warping Algorithm* (DTW)

Dynamic Time Warping algorithm (DTW) [Sakoe , H. & S. Chiba-8] adalah algortima yang menghitung *optimal warping path* antara dua waktu. Algoritma ini menhitung baik antara nilai *warping path* dari dua waktu dan jaraknya. Misalnya, kita memiliki dua sekuens numerik (a_1, a_2, ..., a_m) dan (b_1, b_2, ..., b_m). Dengan pemisalan ini, maka dapat dikatakan bahwa panjang dua sekuens ini bisa saja berbeda. Algoritma ini memulain dengan penghitungan jarak lokal antara elemen dari sekuens menggunakan tipe jarak yang berbeda. Frekuensi yang paling banyak menggunakan method untuk penghitungan jarak adalah jarak absolut antar nilai dua elemen. Jika dalam matriks maka dapat ditulis dengan memiliki *n* garis dan *m* kolom, secara umum:

$$d_{ij} = \left| a_i - b_j \right|,\ i = \overline{1,n},\ j = \overline{1,m}.$$

Mulai dengan matrik jarak lokal, kemudian minimum jarak matriks antar sekuens ditentukan menggunakan algoritma program dinamis dan mengikuti kriteria optimasi berikut:

$$a_{ij} = d_{ij} + \min(a_{i-1,j-1}, a_{i-1,j}, a_{i,j-1})$$

Dimana a_{ij} merupakan jarak minimal antara subsekuens. *Warping path* adalah sebuah *path* yang melewti jarak matrik minimum dari elemen a_{11} ke a_{nm} . ongkos *warping path* secara global dari dua sekuens:

$$\mathrm{GC} = \frac{1}{p}\sum_{i=1}^{p} w_i$$

Dimana W_i adalah elemen yang dimiliki *warping path* san p adalah jumlahnya. Penghitungannya dibuat untuk dua sekuens diperlihatkan pada gambar dibawah dan *warping path* diberi *highlight*.

	-2	10	-10	15	-13	20	-5	14	2
3	5	12	25	37	53	70	78	89	90
-13	16	28	15	43	37	70	78	105	104
14	32	20	39	16	43	43	62	62	74
-7	37	37	23	38	22	49	45	66	71
9	48	38	42	29	44	33	47	50	57
-2	48	50	46	46	40	55	36	52	**54**

Gambar 1. ***Warping path***

Ada tiga kondisi yang menentukan pada DTW algorithm yang meyakinkan konvergensi cepat:

1. Monotony – *path* yang tidak pernah ada kembalian, yang berarti antara index i dan j digunakan untuk menyebrang sekuens tidak pernah berkurang.
2. Continuity – *path* berkembang yang secara berangsur-angsur, tahap per tahap, yang berarti index i dan j naik dengan maksimum kenaikan 1 unit setiap langkahnya.

3. Boundary – *path* mulai dari pojok kiri bawah dan berakhir pada pojok kanan atas.

Karena prinsip optimasi dalam program dinamis diimplementasikan pada teknik "backward", mengidentifikasi *warp path* menggunakan tipe struktur dinamis yang disebut stack. Seperti algoritma program dinamis lainnya. DTW memiliki kompleksitas polinomial. Ketika sekuens memiliki banyak elemen, minimal ada dua ketidaknyamanan:

1. Mengingat matriks yang besar
2. Menampilkan banyak perhitungan jarak

Ada perbaikan dalam standar DTW algorithm yang merangkum dua masalah diatas dengan nama: FastDTW (Fast Dynamic Time Warping) [Stan Salvador, Philip Chan - 6]. Solusi yang ditawarkan berisi pembagian jarak matriks ke dalam 2,4,8,16,dst. Dengan cara ini, perhitungan jarak diperlihatkan pada matriks yang lebih kecil dan *warp path* digunakan saat menggabungkan dari matriks kecil tadi.

2.3 Menggunakan DTW Algorithm dalam *Speech Recognition*

Vocal Signal Analysis. Suara merambat melalu udara sebagai gelombang longitudinal dengan kecepatan yang tergantung densitas udara. Cara yang paling mudah untuk merepresentasikan suara adalah dengan grafik sinusoidal. Grafik tersebut merepresentasikan variasi dari tekan udara tergantung waktunya.

Ada tiga hal yang membentuk gelombang suara, yaitu amplitudo, frekuensi, dan fase. Amplitudo diukur menggunakan satuan decibels (DB), pengukuran dilakukan dengan mengikuti fungsi logaritma sebagai standar suara. Pengukuran amplitudo menggunakan DB sangat penting karena ini representasi langsung bagaimana suara dirasakan oleh orang. Frekuensi adalah banyaknya gelombang per detik, biasa diukur menggunakan skala Hertz (Hz). Kemudian, fase mengukur posisi dari awal gelombang sinus

Untuk membuat suara menjadi kurva sinusoidal, digunakanlah teorema Fourier.

Word detection. Teknologi sekarang ini bisa mengidentifikasi secara akurat awal dan akhir satu kata diucapkan dalam audio stream, tergantung pada proses sinyal yang berbeda dengan waktu. Dengan mengevaluasi energi dan rata-rata magnitud dalam waktu yang singkat dan menghitung rata-rata *zero-crossing rate*. Menetapkan poin awal dan akhir merupakan masalah sederhana jika rekaman audio dilakukan dalam kondisi yang ideal. Dalam kasus ini, rasio *signal-noise*-nya tinggi karena mudah untuk menentukan lokasi dalam stream yang terdiri dari sinyal valid dengan analisis sampel. Dalam kondisi sebenarnya tidak lah sesederhana itu, *background-noise* memiliki intensitas yang signifikan dan dapat mengganggu proses isolasi kata dalam stream.

Algoritma yang paling baik untuk mengisolasi kata adalah Rabiner-Lamel Algorithm. Jika kita mempertimbangkan *signal-window* $\{s_1, s_2, ..., s_n\}$ dimana n adalah jumlah sampel dari *window* dan s_i, $i = 1, n$ adalah ekspresi numerik dari sampel, energi yang berasosiasi dengan *signal-window*:

$$E(n) = \frac{1}{n}\sum_{i=1}^{n} s_i^{\,2}$$

Rata-rata *zero-crossing rate*:

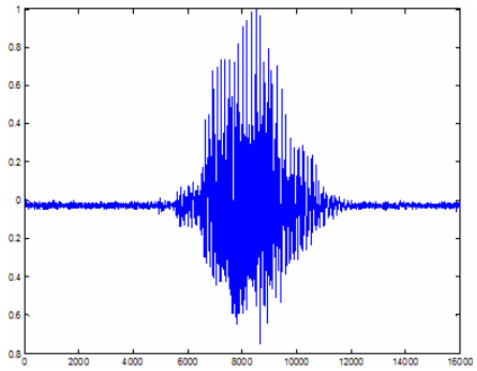
$$ZCR(n) = \sum_{i=1}^{n-1} sign(s_i)\cdot sign(s_{i+1})$$

$$\text{where } sign(s_i) = \begin{cases} 1 & \text{if } s_i > 0 \\ 0 & \text{if } s_i < 0 \end{cases}.$$

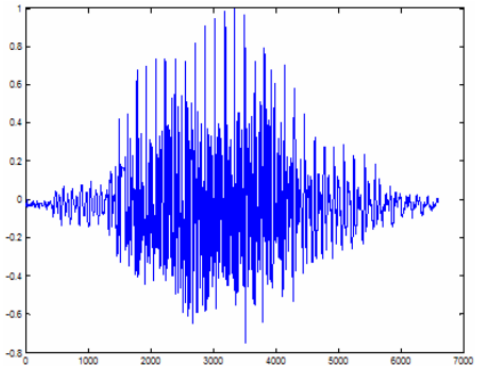
Metode menggunakan tiga numerik level: dua untuk energi (superior, inferior) dan satu untuk Rata-rata *zero-crossing rate*.

2.4 Menggunakan DTW Algorithm dalam *Word Identification*

Identifikasi kata bisa dilakukan dengan perbandingan langusng numerik form dari sinyal atau dari spectogram sinyal. Proses perbandingan dalam dua kasus di atas harus dijamin bahwa pada keduanya memiliki perbedaan panjang dari sekuens dan non-linear. DTW algorithm sukses dalam mengurutkan penyelesaian masalah dengan menemukan hubungan *warp path* ke jarak optimal antara dua sekuens yang berbeda panjang. Ada beberapa kekhususan ketika algoritma ini digunakan dalam dua kasus:



Gambar 2. ***Sinyal suara untuk kata "noua" (kasus 1)***



Gambar 3. ***Sinyal suara untuk kata "noua" (kasus 2)***

1. Perbandingan langsung dari bentuk numerik atau sinyal. Dalam kasus ini, untuk setiap numerik sekuens, sebuah sekuens baru dibuat, maka sekuens ada yang memiliki dimensi yang lebih kecil. Numerik sekuens bisa memiliki ribuan nilai numerik, ketika subsekuens bisa memiliki ratusan. Mengurangi jumlah nilai numerik bisa dilakukan dengan menghilangkan yang ada diantara ekstrim poin. Proses pengurangan numerik sekuens ini tidak diperkenankan menggabungkan bentuknya. Dan proses ini bisa membawa ke pengurangan pengenalan presisi. Tetapi, membuat nilainya naik dalam hal kecepatan, presisi pada faktanya meningkat dengan memperluas jumlah kata dalam kamus.
2. Representasi sinyal spektrogram dan mengaplikasikannya pada DTW algorithm untuk perbandingan spektogram. Method ini terdiri dari pembagian numerik sinyal dalam banyak *window* (interval) yang akan *overlap*. Setiap window, angka real di interval, *quick fourier's transform* akan dihitung dan disimpan dalam matriks: spektrogram suara. Parameter yang digunakan akan sama untuk semua operasi penghitungan, yaitu panjang *window*, panjang *fourier's transform*, dan panjang *overlap window* untuk dua *window* berturut-turut. Fourier's transform secara simstris berhubungan dengan pada pusat dan bilangan kompleks dari setengah konjugat bilangan kompleksdari bilangan simetris pada setengah pertama. Dalam kenyataan ini, hanya nilai dari setengah pertama yang disimpan, maka spektrogram akan menjadi matriks bilangan kompleks, banyaknya baris sama dengan setangah dari panjang *fourier's transform* dan banyaknya kolom tergantung dari panjangnya suara. DTW akan diaplikasikan dalam matriks bilangan real yang dihasilkan dari konjugasi nilai spektrogram.

IV. KESIMPULAN

Dynamic Time Warping Algorithm sangan berguna untuk mengisolasi pengenalan kata yang diucapkan dalam kamus yang terbatas. Untuk pengenalan pengucapan secara fasih, Hidden Markov Chains digunakan. Dan penggunaan program dinamis meyakinkan sebual kompleksitas polinomial pada algoritma: $O(n^2v)$, dimana n adalah panjang sekuens dan v adalah jumlah kata dalam kamus.

Ada beberapa kelemahan dari *Dynamic Time Warping Algorithm*. Pertama, $O(n^2v)$ kompleksitasnya tidak memuaskan untuk kamus lebih besar yang bisa meyakinkan kenaikan dalam kesuksesan proses *recognition*. Dua, sulit untuk mengevaluasi dua elemen dari dua sekuens yang berbeda, mengambil dari nilai yang ada banyak jalur yang memiliki fitur khusus. Bagaimanapun juga, *Dynamic Time Warping Algorithm* tetap sebuah algoritma yang mudah untuk diimplementasikan, sangat cocok untuk aplikasi yang membutuhkan *word recognition* yang sederhana.

REFERENSI

[1] Munir, Rinaldi, Strategi Algoritma, Program Studi Informatika, Institut Teknologi Bandung. 2007.

[2] English Wikipedia 2008 http://en.wikipedia.org/wiki/Speech_recognition

[3] English Wikipedia 2008 http://en.wikipedia.org/wiki/Dynamic_time_warping

[4] Titus, Felix, Dynamic Programming Algorithms in Speech Recognition, Academy of Economic Studies, Bucharest.

[5] Pavel, Senin, Dynamic Time Warping Algorithm Review, Information and Computer Science Department, University of Hawaii at Manoa, 2008.

[6] Cory Myers, Lawrence R. Rabiner, Aaron E.Rosenberg, Performance Tradeoffs in Dynamic Time Warping Algorithms for Isolated Word Recognition, Ieee Transactions On Acoustics, Speech, And Signal Processing, Vol. Assp-28, No. 6, December 1980 http://www.caip.rutgers.edu/~lrr/publications.html