MODEL LINIER DINAMIK SEBAGAI DASAR PENYELEKSIAN DIPHONE PADA SISTEM PENSITESAN SUARA CONCATENATIVE DALAM BAHASA INDONESIA

Muhammad Subali, Djati Keram

subali@staff.gunadarma.ac.id

Abstrak


Kata Kunci: Model linier dinamik, pensintesa suara, Concatenative, diphone, kalman filter

1. Pendahuluan

Sinyal suara manusia mengandung berbagai macam informasi terutama sekali sinyal suara tersebut mengandung kata ataupun pesan dalam bentuk ucapan yang akan disampaikan. Dengan semakin berkembangnya teknologi komputer yang dapat memicu perkembangan di berbagai bidang, dimana salah satunya adalah pemanfaatan komputer untuk dapat memproses suara manusia. Pada dasarnya teknologi dalam bidang ini dikelompokkan dalam dua kelompok utama yaitu pengenalan ucapan (Speech Recognition/ SR) dan pensintesa ucapan (Speech synthesis/SS). Untuk teknologi SR kegiatan yang dilakukan adalah bagaimana komputer dapat memproses dan mengenali suara bagaimana komputer menjadi suatu teks atau tulisan, sedangkan dalam teknologi SS kebalikannya yaitu bagaimana komputer dapat membangkitkan suara manusia dari suatu teks, dan dari kedua teknologi ini yang akan dibahas dalam tulisan ini adalah tentang pensintesa ucapan (Speech synthesis/SS) khususnya untuk Bahasa Indonesia hasil karya Dr. Arry Akhmad Armand (Elektro, ITB) yang dipublikasikan pada tahun 2000 yang diberi nama TTS INDO. Di mana pensintesa ini menggunakan teknik diphone concatenation yang bekerja dengan cara menggabung-gabungkan segmen-segmen bunyi berupa diphone (gabungan dua buah fonem) yang telah direkam sebelumnya.

Tantangan teknis utama pada teknik concatenative adalah mencari algoritma untuk menggabungkan diphone dengan diphone lainnya, serta algoritma untuk memanipulasi diphone, khususnya untuk mengubah durasi serta pitch diphone. Berbagai teknik yang berkembang untuk mendukung pensintesa jenis ini diantaranya adalah yang tergolong sistem non-paramtrik seperti overlap and add (OLA), pitch synchronouse overlap and add (PSOLA) dikembangkan oleh France Telecom (CNET), multiband resynthesis-OLA (MBROLA), time domain-PSOLA (TD-PSOLA) serta Linear Prediction-PSOLA (LP-PSOLA) [Dut97].


Dan dalam tulisan ini akan dibahas proses penyeleksiaisian diphone dengan menggunakan model linier dinamik dalam hal ini kalman filter pada TTS INDO. Akan dilakukan eksperimen untuk melihat penggabungan dari diphone-diphone dalam suatu kata atau kalimat, sehingga dapat ditentukan diphone mana yang tepat untuk dipilih dalam penggabungan tersebut. Misalnya diphone /ai/ pada kata “pandai”, apakah diphone /ai/ atau /sy/ yang harus dipilih.
2. Sistem Pensintesa Suara

Adapun sistem dari pensintesa suara (SS) ini pada prinsipnya terdiri dari dua sub sistem, yaitu bagian converter teks ke fonem (Text to Phoneme) dan bagian converter fonem ke ucapan (Phoneme to Speech), dimana kedua bagian ini dapat digambarkan pada gambar 1 di bawah.

**TEXTS**

**TAHAP I**

Text to Phoneme

**UCAPAN**

Phoneme to Speech

*Gambar 1. Prinsip dari SS*

Urutan dari tahapan-tahapan proses untuk sistem pensintesa ucapan ini secara detail digambarkan pada gambar 2.

*Gambar 2. Blok Konversi Teks Ke Ucapan*

Proses-proses pada tahap I.

- **Tahap normalisasi teks** berfungsi untuk mengubah semua teks kalimat yang ingin diucapkan menjadi teks yang secara lengkap memperlihatkan cara pengucapanannya (Misalnya 1 $ setara Rp 1000; maka di ucapan satu dollar setara dengan seribu rupiah)
- **Tahap berikutnya Exception Dictionary Lookup dan Letter-to-Phoneme Conversion** adalah melakukan konversi dari teks yang sudah secara lengkap merepresentasikan kalimat yang ingin diucapkan menjadi kode-kode fonem dengan aturan tertentu misalnya,
Left-context [letter-set] right-context = phoneme string

Proses Pada Tahap II.

- Bagian prosodi generator akan melengkapi setiap unit fonem yang dihasilkan dengan data durasi pengucapannya serta pitchnya. Data durasi serta pitch diperoleh berdasarkan kombinasi antara tabel atau database serta model prosodi (misalnya /s/ [40ms] [90Hz]). Secara simbolik, hasil dari bagian ini sudah menghasilkan informasi yang cukup untuk menghasilkan ucapan yang diinginkan.
- Tahap berikutnya yang masih sering dilakukan adalah Phonetic Analysis. Tahap ini dapat dikatakan sebagai tahap penyempurnaan, yaitu melakukan perbaikan di tingkat bunyi. Sebagai contoh, dalam bahasa Indonesia, fonem /k/ dalam kata bapak tidak pernah diucapkan secara tegas, atau adanya sisipan fonem /y/ dalam pengucapan kata alamiah antara fonem /i/ dan /a/.
- Proses Speech Synthesis
- Sistem ini untuk memodifikasi prosodi dari segmen suara pada database, dimana teknik yang dapat digunakan dalam proses ini, diantaranya adalah teknik formant synthesizer dan concatenative (diphone concatenation dan unit-selection) serta sistem yang menggunakan pendekatan stokastik (model linier dinamik).

3. Teknik Concatenative

Pensintesa suara dengan teknik concatenative merupakan sistem pensintesa yang melakukan penggabungan segmen-segmen ucapan yang direkam sebelumnya, teknik ini terdiri dari i dari diphone concatenation dan Unit-Selection.


Gambar 3. Proses Diphone Concatenation

Gambar 3 di atas menjelaskan teknik diphone dalam penggabungan kata “dog”. Dimana tanda (titik 3 kali “...”) menandakan awal diphone dan akhir diphone. Sehingga penggabungannya adalah segmen [...d] +{do}+{og}+{g...} yang mana signal untuk masing-masing segmen ini sudah tersimpan
dalam database. Pada teknik ini variasi kumpulan segment-segment diphone yang tersimpan dalam database terbatas, sehingga hasil penggabungsannya kurang optimal. Salah satu pendekatan untuk menghasilkan sintesa gelombang suara dengan bunyi yang lebih alami adalah bagaimana menyeleksi dan menggabungkan unit-unit (fonem-fonem) yang terdapat dalam data base (Unit-Selection). Gambar 4 menunjukkan proses dari unit-selection.

4. Penyeleksian Unit Bunyi

Seperti telah diuraikan di atas bahwa untuk menghasilkan sintesa gelombang suara dengan bunyi yang lebih alami adalah bagaimana menyeleksi dan menggabungkan unit-unit (fonem-fonem) yang terdapat dalam data base (Unit-Selection). Dalam Teknik unit selection, unit-unit dalam data base dapat dipandang sebagai jaringan transisi keadaan, dan dalam proses penyeleksianya digunakan viterbi Search [Andrew J. Hunt dan Alan W. Black1996]. Bentuk dari jaringan transisi diilustrasikan pada gambar 5 di bawah.

![Diagram](image)

Gambar 6. Target Cost dan Concatenation Cost

Dimana target cost adalah estimasi dari perbedaan unit dalam database \( u_i \) dan target \( t_i \), \( C(u_i, t_i) \), yang diperkirakan akan ditampilkan. Sedangkan concatenation cost \( C(u_{t,i}, u_i) \) adalah estimasi dari kualitas gabungan antara unit-unit yang direnetikan \( (u_{t,i} \text{ dan } u_i) \). Tugas dari pensintesa ini adalah mencari jalur lintas melalui jaringan transit keadaan sehingga diperoleh deretan unit-unit dalam database dengan total cost yang minimum, dimana total cost ini merupakan jumlah dari target cost dan concatenation cost.


![Gambar 7. Perbedaan signal /æ/ dalam "cat" dan "cap"](image)

[Esther Klabbers dan Raymond Veldhuis, EK, 1998], melakukan penelitian untuk dapat menangani masalah diskontinu dalam pensintesa suara, menurut dia masalah utamanya adalah pada spectrum alami. Beberapa pendekatan ditawarkan untuk memecahkan masalah ini:

1. Suara yang terdengar diskontinu dalam diphone dapat direduksi dengan menggunakan unit yang lebih besar seperti triphone. Namun hal ini akan menyebabkan terjadinya lonjakan yang drastis pada inventory (database)
2. Ketidaksensiaian spectrum dapat diminimisasi dengan memvariasikan lokasi diphone yang bergantung dengan konteks.
3. Context-sensitive diphone atau specialized units dalam database.

Dari hasil eksperimen ini dicari suatu korelasi dengan metode pengukuran jarak spectral dengan metode Kullback-Leiber measure (KL) dan Mel-Frequency Cepstral Coefficient (MFCC). Gambar 8 dibawah menunjukkan bahwa vocal /a/, /i/, dan /u/ mempunyai variasi yang kecil sedangkan /A/ dan /u/ mempunyai variasi yang besar. Sehingga dari kedua metode diatas dapat ditentukan...
diphone mana yang harus dibuat menjadi triphone (Diphone clustering) sehingga dapat membatasi penambahan diphone.

\[ D(i,j) = \sqrt{(F_{1,i} - F_{1,j})^2 + (F_{2,i} - F_{2,j})^2} \]

Gambar 8. Variasi frekuensi untuk vocal


Persepsi manusia dari uji coba persepsi pendengaran dilakukan terhadap 17 orang partisipan untuk menyeleksi adanya diphthong dalam suatu kalimat alami dalam hal ini diphthong American English (ey, ow, ey, aw dan oy). Pengukuran jarak dilakukan melalui parameter signal suara seperti Mel frequency Cepstral Coefficients (MFCCs), Line Spectral Frequency (LSFs) dan Multiple Centroid Analysis (MCA).


<table>
<thead>
<tr>
<th>Diphthong</th>
<th>MFCC</th>
<th>LSF</th>
<th>MCA</th>
<th>MCA wgts</th>
<th>LDM</th>
</tr>
</thead>
<tbody>
<tr>
<td>ey</td>
<td>0.21</td>
<td>0.37</td>
<td>0.36</td>
<td>0.44</td>
<td>0.58</td>
</tr>
<tr>
<td></td>
<td>0.66</td>
<td>0.59</td>
<td>0.46</td>
<td>0.60</td>
<td>0.17</td>
</tr>
<tr>
<td>ow</td>
<td>0.31</td>
<td>0.21</td>
<td>0.19</td>
<td>0.19</td>
<td>0.26</td>
</tr>
<tr>
<td></td>
<td>0.56</td>
<td>0.40</td>
<td>0.46</td>
<td>0.52</td>
<td>0.34</td>
</tr>
<tr>
<td>ay</td>
<td>0.39</td>
<td>0.01</td>
<td>0.03</td>
<td>-0.02</td>
<td>0.56</td>
</tr>
<tr>
<td></td>
<td>0.66</td>
<td>0.61</td>
<td>0.45</td>
<td>0.49</td>
<td>0.59</td>
</tr>
<tr>
<td>aw</td>
<td>0.34</td>
<td>0.66</td>
<td>0.35</td>
<td>0.49</td>
<td>-0.02</td>
</tr>
</tbody>
</table>

Model Linier Dinamik Sebagai Dasar Penyeleksian Diphone pada Sistem Pensintesa Suara Concatenative dalam Bahasa Indonesia
5. Model Linier Dinamik (Kalman Filter) [10, 11, 13, 20, 21]

Model linier dinamik akan digunakan dalam penelitian ini untuk mengatur proses penggabungan unit-unit bunyi yang akan direndetkan. Model ini merupakan Linear time-invariant systems yang dikenal sebagai model ruang keadaan (state-space models) atau dikenal juga sebagai Kalman Filter, yang akan mengestimasi parameter-parameter dari unit bunyi tersebut sehingga dapat diprediksi unit-unit bunyi yang mana yang sesuai untuk digabungkan. Model untuk sistem ini mempunyai persamaan seperti di bawah yang dikenal sebagai one-step prediction pada persamaan 1.

\[ X(t+1) = AX(t) + BU(t) + V(t) \quad \text{dan} \quad y(t) = CX(t) + DU(t) + W(t) \]

Di mana, \( X(t) \) dan \( Y(t) \) variabel keadaan dan variabel keluaran; \( W(t) \) dan \( V(t) \) variabel gangguan masukan dan keluaran; \( A \), \( B \), \( C \) dan \( D \) merupakan matriks yang mencirikan dinamika dari sistem. Data step dari kalman filter dapat digunakan untuk mengukur \( u(t) \) dan \( y(t) \), seperti diilustrasikan pada persamaan 2.

\[
\hat{y}(t|t-1) = C \hat{x}(t|t-1) + Du(t)
\]

\[
e(t) = y(t) - \hat{y}(t|t-1)
\]

\[
L(t) = P(t|t-1)C^T(CP(t|t-1)C^T + R)^{-1}
\]

\[
\hat{x}(t + t) = \hat{x}(t|t-1) + L(t)e(t)
\]

\[
P(t|t) = P(t|t-1) - L(t)CP(t|t-1)
\]

Algoritma dari model linier dinamik (kalman filter) dari persamaan diatas prosesnya dapat di gambarkan seperti pada gambar 9 sebagai berikut: signal suara plus noise \( y_x = (y_1 + y_n) \) disegmentasi, lalu dilakukan estimasi dari parameter-parameter signal tersebut dan dilakukan proses pemfilteran dengan Kalman Filter, sehingga diperoleh hasil estimasi dari signal suara \( \hat{y} \), tersebut.
Gambar 9. Proses Estimasi Signal Suara

Adapun proses dari estimasi parameter signal dapat digambarkan pada gambar 10. Power spektrum dari signal $y_x$ diestimasi ($\hat{P}_x$), lalu power spektrum dari signal ini mengestimasi power spektrum noise ($\hat{P}_n$). Estimasi power spektrum signal suara ($\hat{P}_x$) dihitung dari $\hat{P}_x$ dan $\hat{P}_n$. Dan akhirnya parameter-parameter dari signal suara dan noise dapat di peroleh.

Gambar 10. Estimasi parameter signal suara dan noise
6. Hasil Percobaan dan Diskusi


Tabel 2. Nilai Kesalahan rata-rata

<table>
<thead>
<tr>
<th>Diphone /ai/ pada kata</th>
<th>Diphone /ay/</th>
<th>Diphone /ai/</th>
</tr>
</thead>
<tbody>
<tr>
<td>pandai</td>
<td>0,0267</td>
<td>0,0462</td>
</tr>
<tr>
<td>ramai</td>
<td>0,0238</td>
<td>0,0444</td>
</tr>
<tr>
<td>tunai</td>
<td>0,0242</td>
<td>0,0445</td>
</tr>
<tr>
<td>perisai</td>
<td>0,0301</td>
<td>0,0501</td>
</tr>
<tr>
<td>tangkai</td>
<td>0,0259</td>
<td>0,0461</td>
</tr>
<tr>
<td>lunglai</td>
<td>0,0268</td>
<td>0,0457</td>
</tr>
<tr>
<td>gapai</td>
<td>0,0243</td>
<td>0,0439</td>
</tr>
</tbody>
</table>


Model linier dinamik dalam hal ini kalman filter dapat digunakan dalam pemakaian pada pensintesa suara, khususnya untuk melihat tingkat kesalahan dalam memilih diphone-diphone mana yang tepat untuk digabungkan dalam membentuk suatu kata sehingga kata yang terdengar lebih alami.

8. Daftar Pustaka


