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

Muhammad Subali, Djati Kerami

subali@staff.gunadarma.ac.id

***Abstrak***

*Sistem pensintesa suara dengan. teknik diphone concatenative merupakan teknik pensintesa di mana segmen suara dalam bentuk diphone sudah direkam sebelumnya. Suatu teks akan disintesa menjadi suara berdasarkan pada gabungan dari diphone-diphone penyusun pada teks tersebut. Untuk memperoleh hasil suara yang alami, maka perlu diseleksi (dipilih) diphone mana yang tepat untuk digabungkan.Dengan model linier dinamik akan ditentukan diphone mana yang akan dipilih.Eksperimen menggunakan TTS- INDO (MBROLA) untuk beberapa diphone.*

*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 berkembangannya 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 ucapan bahasa manusia menjadi suatu teks atau tulisan, sedangkan dalam teknologi *SS* kebalikannya yaitu bagaimana komputer dapat membangkitkan ucapan manusia dari suatu teks, dan dari kedua teknologi ini yang akan dibahas dalam tulisan ini adalah tentang pensintesa ucapan (*Speech syntesis/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 pesintesa ini menggunakan teknik *diphone concatenation* yang bekerja dengan cara menggabung-gabungkan segmen-segmen bunyi berupa *diphone* (gabungan dua buah fonem) yang telah direkam sebelumnya.

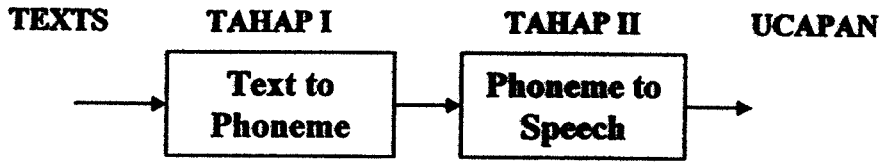
Tantangan teknis utama pada teknik *concatenative* adalah mencari algoritma untuk menggabungkan *diphone* dengan *diphone* lainnya, serta algoritma untuk memanipulasi *diphone*, khususnya untuk mengubah durasi serta *pitch diphone*. Berbagai teknik yang berkembang untuk mendukung pensintesa jenis ini 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].

Sejak tahun 1996 , Proses penyeleksian untuk penggabungan unit-unit bunyi ini dilakukan dengan beberapa metode yang berkembang terus, dan pada tahun 2003 suatu metode baru diperkenalakan oleh [*Jitendra Vepa, Simon King, 2003*] untuk menghitung *join-cost* pada sistem pensintesa ucapan *unit-selection* dengan menggunakan teknik parametrik yaitu melalui pendekatan stokastik seperti model linier dinamik (LDM) diantaranya *Hidden Markov Model (HMM), dan Kalman Filter*.

Dan dalam tulisan ini akan dibahas proses penyeleksian 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 /ay/ yang harus dipilih.

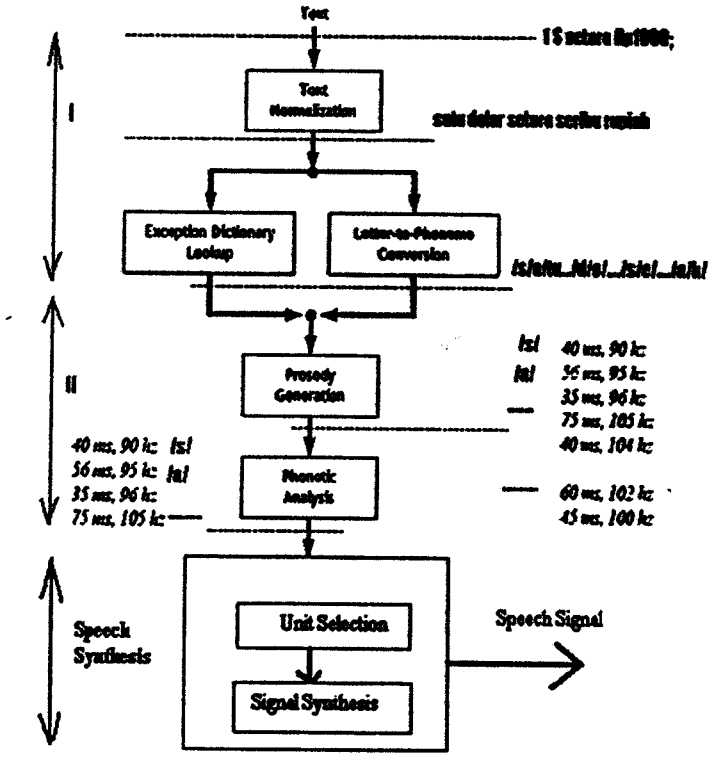
## 2. Sistem Pensintesa Suara

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.




*Gambar 1. Prinsip dari SS*

Urutan dari tahapan-tahapan proses untuk sistem pesintesa 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 pengucapannya (Misalnya 1 $ setara Rp 1000; maka di ucapkan *satu dollar setara dengan seribu rupiah* )
- Tahap berikutnya *Exception Dictionary Lookup* dan *Letter –to-Phonemme 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/.

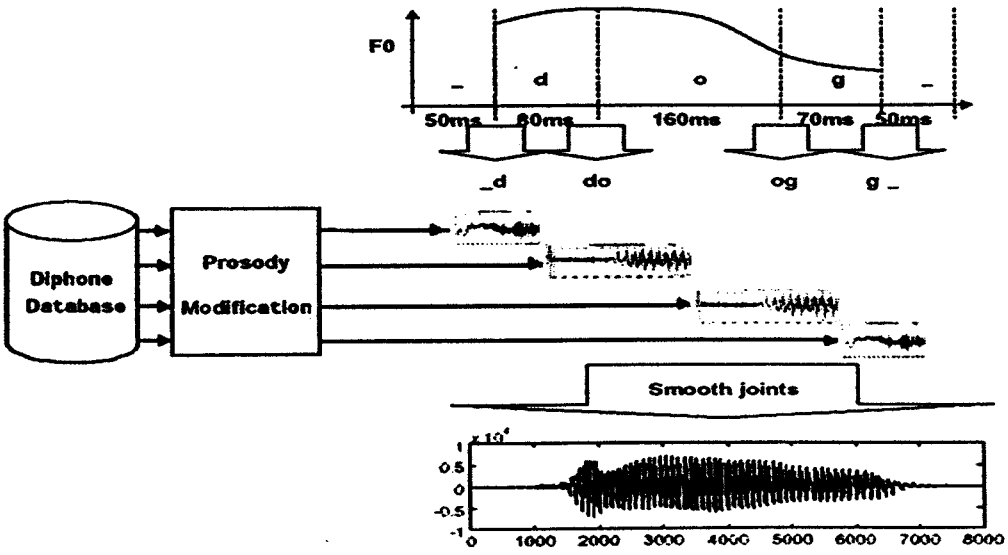
**Proses Speech Synthesis**

- 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 pensintsa yang melakukan penggabungan segmen-segmen ucapan yang direkam sebelumnya, teknik ini terdiri dari i dari *diphone concatenation* dan *Unit -Selection.*

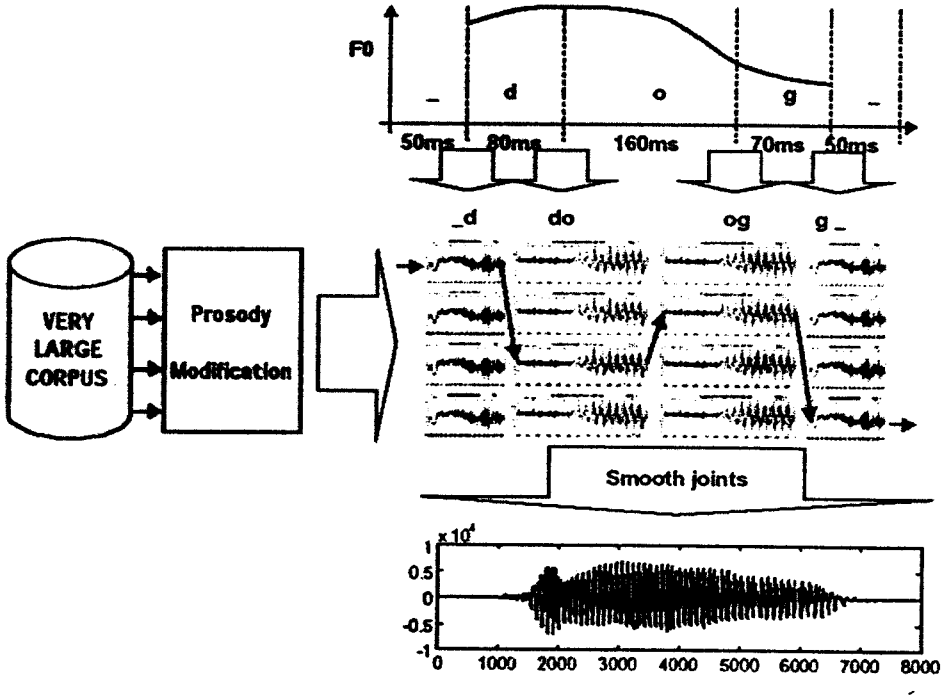
Synthesizer yang menggunakan teknik *diphone concatenation* bekerja dengan cara menggabung-gabungkan segmen-segmen bunyi yang telah direkam sebelumnya. Setiap segmen berupa *diphone* (gabungan dua buah fonem). Synthesizer jenis ini dapat menghasilkan bunyi ucapan dengan tingkat kealamiahan (*naturalness*) yang tinggi. Teknik ini diperkenalkan sejak tahun 1977 , proses penggabungan dua buah fonem dalam model ini, diilustrasikan pada gambar 3 dibawah




*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

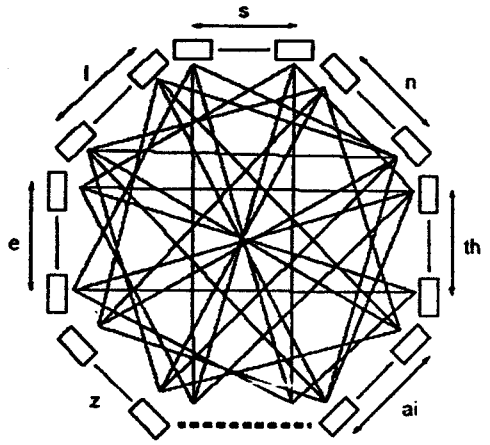
dalam database. Pada teknik ini variasi kumpulan segmen-segmen diphone yang tersimpan dalam database terbatas, sehingga hasil penggabungannya kurang optimal. Salah satu pendekatan untuk menghasilkan sintesa gelombang suara dengan bunyi yang lebih alami adalah bagaimana menyeleksi dan menggabungan unit-unit (fonem-fonem) yang terdapat dalam data base (*Unit-Selection*). Gambar 4 menunjukkan proses dari unit-selection.




*Gambar 4. Proses Unit Selection*

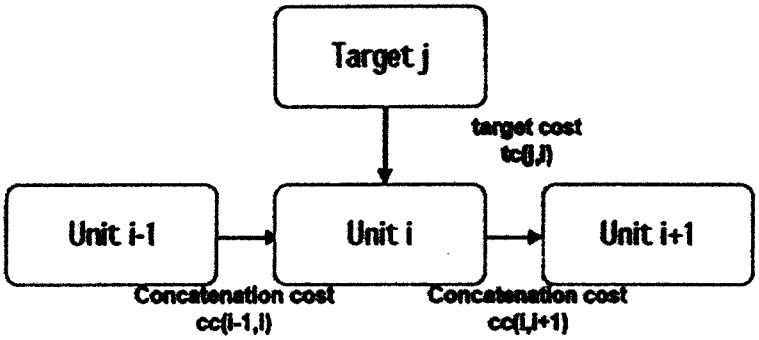
## 4. Penyeleksian Unit Bunyi

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 transisis keadaan, dan dalam proses penyeleksiannya digunakan *viterbi Search* [*Andrew J Hunt dan Alan W Black1996*]. Bentuk dari jaringan transisi diilustrasikan pada gambar 5 di bawah.




*Gambar 5. Jaringan Fonem pada Data Base*

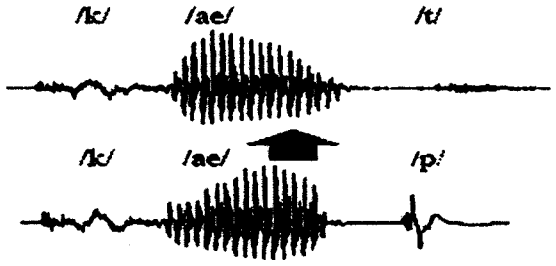
Penyeleksian fonem – fonem ini dilakukan berdasarkan dua fungsi *cost*, yang terdiri dari target *cost* dan concatenation *cost*. Yang dapat diilustrasikan pada gambar 6 di bawah.




*Gambar 6. Target Cost dan Concatenation Cost*

Dimana target *cost* adalah estimasi dari perbedaan unit dalam database $u_i$ dan target $t_i$; $C^t(u_i,t_i)$, yang diperkirakan akan ditampilkan.. Sedangkan *concatenation cost* $C^c(u_{i-1}, u_i)$ adalah estimasi dari kualitas gabungan antara unit-unit yang direntetkan ($u_{i-1}$ dan $u_i$). Tugas dari pensintesa ini adalah mencari jalan/lintasan melalui jaringan transisi 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*.

Beberapa penelitian untuk dapat mengatasi kekurangan-kekurangan pada pensintesa suara ini terus dilakukan. [*John Wouters dan Michael W Macon* ,JW,1998] melakukan penelitian tentang evaluasi persepsi dari ukuran jarak pada pensintesa suara concatenative. Menurutnya setiap fonem dalam suatu bahasa terdiri dari variasi-variasi fonetik yang disebut dengan allophone. Misalnya /ae/ dalam "cat" akan berbeda dengan /ae/ dalam "cap".seperti pada gambar 7 dibawah




*Gambar 7. Perbedaan signal /ae/ dalam "cat" dan "cap"*

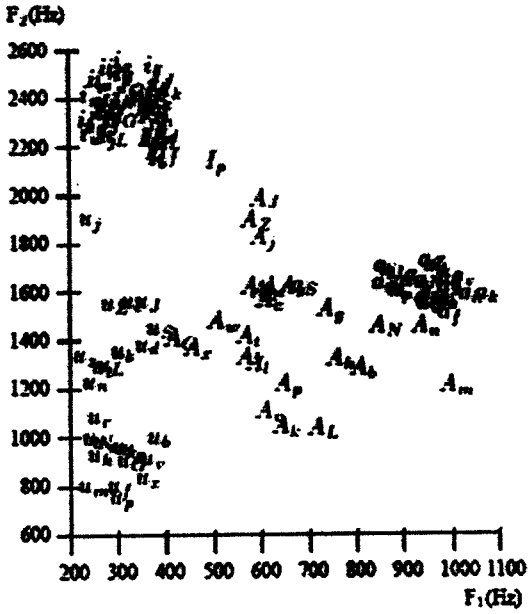
[*Esther Klabbers* dan *Raymond Veldhuis*, EK 1998], melakukan penelitian untuk dapat menangani masalah diskontinu dalam pensintesaan suara, menurut dia masalah utamanya adalah pada spectrum alami. Beberapa pendekatan ditawarkan untuk memecahkan masalah ini:

1. Suara yang terdengar diskontinu dari diphone dapat direduksi dengan menggunakan unit yang lebih besar seperti triphone. Namun hal ini akan menyebabkan terjadinya lonjakan yang drastis pada inventory (database)
2. Ketidaksesuaian 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 /I/ 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*

Untuk menentukan ukuran jarak objektif dalam masalah diskontinu spectral pada pensintesa suara concatenative, *[Jitendra Vepa,Simon King dan Paul taylor JV,2002]* melakukan experimen persepsi untuk mengukur korelasi antara persepsi manusia sebagai subjek dan variasi objective berdasarkan pengukuran spectral (experiment menggunakan *r-Voice* dari *Rhetorical system Ltd*).

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,ay,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) .

*Jitendra Vepa,Simon King dan Paul Taylor [JV,2002]*, pada tahun yang sama meneliti dengan metode baru untuk mengukur jarak objektif, yang disebut dengan *weight distance.*Suatu metode baru diperkenalakan oleh *[Jitendra Vepa, Simon King JV, 2003]* untuk menghitung *joincost* pada system pensintesa ucapan *unit-selection* dengan menggunakan model linier dinamik (LDM) yang dikenal sebagai Kalman Filter. Hasil yang diperoleh dari metode-metode di atas dalam bentuk faktor korelasi pada tabel 1 di bawah.

*Tabel 1. Nilai faktor korelasi dengan beberapa metode*

| Diphtong | MFCC | LSF | MCA | MCA wgts | LDM |
|---|---|---|---|---|---|
| ey | 0.21 | 0.37 | 0.36 | **0.44** | **0.58** |
| | **0.66** | **0.59** | **0.46** | **0.60** | 0.17 |
| ow | 0.31 | 0.21 | 0.19 | 0.19 | 0.26 |
| | **0.56** | 0.40 | **0.46** | **0.52** | 0.34 |
| ay | 0.39 | 0.01 | 0.03 | -0.02 | **0.56** |
| | **0.66** | **0.61** | **0.45** | **0.49** | **0.59** |
| aw | 0.34 | **0.66** | 0.35 | **0.49** | -0.02 |

| | | | | | |
|---|---|---|---|---|---|
| | **0.77** | **0.78** | **0.57** | **0.62** | **0.50** |
| oy | 0.17<br>-0.01 | 0.28<br>0.17 | **0.53**<br>0.30 | **0.55**<br>0.39 | **0.45**<br>-0.14 |

## 5. Model Linier Dinamik (Kalman Filter).[10,11,13,20,21]

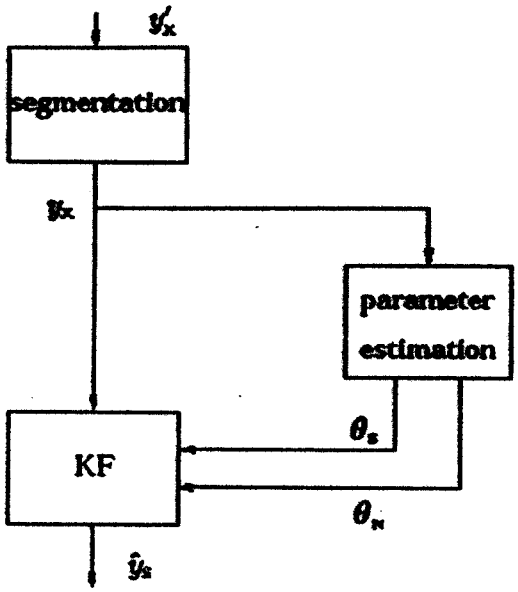
Model linier dinamik akan digunakan dalam penelitian ini untuk mengatur proses penggabungan unit-unit bunyi yang akan direntetkan. Model ini merupakan *Linear time-invariant systems* yang dikenal sebagai model ruang keadaan (*state-space models*) atau dikenal juga sebagai Kalman Filter, yang akan mengestimasi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

$$X(t+1) = A.X(t) + B.U(t) + V(t) \text{ dan } y(t) = C.X(t) + D.U(t) + W(t). \quad (1)$$

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 dilustrasikan pada persamaan 2.

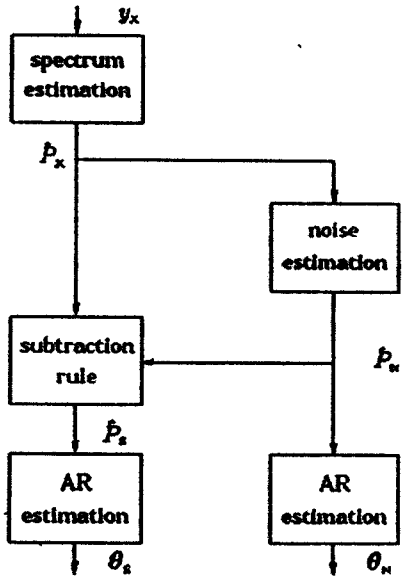
$$\begin{aligned}
&\hat{y}(t|t-1) = C\hat{x}(t|t-1) + Du(t) \\
&e(t) = y(t) - \hat{y}(t|t-1) \\
&L(t) = P(t|t-1)C^T(CP(t|t-1)C^T + R)^{-1} \\
&\hat{x}(t+t) = \hat{x}(t|t-1) + L(t)e(t) \\
&P(t|t) = P(t|t-1) - L(t)CP(t|t-1)
\end{aligned} \quad (2)$$

Algoritma dari model linier dinamik (kalman filter) dari persamaan diatas prosesnya dapat di gambarkan seperti pada gambar 9 sebagai berikut: signal suara plus nois $y'_x = (y_s+y_n)$ disegmentasi, lalu dilakukan estimasi dari parameter-parameter signal tersebut dan dilakukan proses pemfiteran dengan Kalman Filter, sehingga diperoleh hasil estimasi dari signal suara $\hat{y}_s$ tersebut.




*Gambar 9. Proses Estimasi Signal Suara*

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




*Gambar 10. Estimasi parameter signal suara dan noise*

## 6. Hasil Percobaan dan Diskusi

Dalam penelitian ini dilakukan percobaan dengan menggunakan Kalman filter untuk mengestimasi signal suara pada pensintesa suara concatenative dalam Bahasa Indonesia (TTS-INDO). Karena dalam pensintesa suara ini, suara yang akan dibangkitkan merupakan penggabungan dari segmen-segmen suara (diphone) dari suatu teks, maka agar dapat menghasilkan suara yang alami perlu dipilih diphone-diphone mana yang tepat untuk digabungkan.Misalnya diphone/ai/ dan /ay/, mana yang harus dipilih untuk membangkitkan kata /pandai/,/ramai/,/lambai/ dsb.Mengingat dalam bahasa indonesia pada kata-kata tersebut selalu menggunakan diphone /ay/, maka akan dilihat seberapa besar kesalahan (error) yang terjadi antara kedua diphone tersebut terhadap kata yang akan dibangkitkan. Dalam percobaan ini, segmen /ai/ dari kata /pandai/,/ramai/,/lambai/ dsb itu diasumsikan sebagai signal (suara + noise) atau $y'_x$ sedangkan diphone /ay/ dan /ai/ yang tersimpan dalam data base diasumsikan sebagai signal suara $y_x$. Percobaan dilakukan dengan MATLAB, dimana hasilnya seperti pada tabel 2.

*Tabel 2. Nilai Kesalahan rata-rata*

| Diphone /ai/ pada kata | Diphone /ay/ | Diphone /ai/ |
|---|---|---|
| pandai | 0,0267 | 0,0462 |
| ramai | 0,0238 | 0,0444 |
| tunai | 0,0242 | 0,0445 |
| perisai | 0,0301 | 0,0501 |
| tangkai | 0,0259 | 0,0461 |
| lunglai | 0,0268 | 0,0457 |
| gapai | 0,0243 | 0,0439 |

Dari hasil percobaan di atas, terlihat bahwa diphone /ai/ pada kata pandai, ramai, tunai, perisai, tangkai, lunglai dan gapai mempunyai tingkat kesalahan lebih kecil jika menggunakan diphone /ay/ dibandingkan diphone /ai/. Ini berati bahwa diphone yang harus dipilih adalah diphone /ay/ untuk menggabungkan diphone tersebut pada kata-kata di atas.

## 7. Kesimpulan.

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

[1] A. Hunt and A. Black, "Unit selection in a concatenative speech synthesis system using a large speech database," *Proc. ICASSP*, pp. 373–376, 1996.
[2] E. Klabbers and R. Veldhuis, "On the reduction of concatenation artefacts in diphone synthesis," *Proc.ICSLP98*, pp. 1983–1986, 1998.
[3] J. Wouters and M. Macon, "Perceptual evaluation of distance measures for concatenative speech synthesis,"*Proc. ICSLP98*, pp. 2747–2750, 1998.
[4] Y. Stylianou andAnn K. Syrdal, "Perceptual and objective detection of discontinuities in concatenative speech synthesis," *Proc. ICASSP*, 2001.
[5] Robert E. Donovan, "A new distance measure for costing spectral discontinuities in concatenative speech synthesisers," *The 4th ISCA Tutorial and Research Workshop on Speech Synthesis*, 2001.
[6] J.Vepa,S.king, and P.Taylor,"Objective distance measures for spectral discontinuitas in concatenative speech synthesis," in proc.ICLSP.Denver USA, September 2002.
[7] J.Vepa,S.king, and P.Taylor," N ew Objective distance measures for spectral discontinuitas in

concatenative speech synthesis," in proc.IEEE.workshop on Speech Synthesis. Santa Monica, USA, September 2002.

[8] J.Frankel and S.king,"ASR-articulation speech recognition,"in Proc. Eurospeech, Aalborg, Denmark, September 20001,pp.599-602.

[9] G.Smith,J de Frietes,T.Robinson,and M. Niranjan,"Speech modeling using subspace and EM techniques,"Advances in Neural Information Processing systems," vol. 12,pp796-802,1999.

[10] J.McKenna and S.Isard, "Tailoring Kalman Filtering towards Speakers Characterisation," in Proc.Eurospeech,Budapst,hungary,Septembaer 1999, pp 2793-2796.

[11] Z.Grahramani and G.Hinton,"Parameter Estimation for linear dynamical system,"in tech.rep.CRG-TR-96-2. Dept of Computer Sience. Univ of Toronto, 1996

[12] Joe frankel,Linear dynamic models for automatic speech recognition,Ph.D.thesis, University of Edinburgs, April 2003.

[13] Jithendra vepa,Simon King,"Kalman-Filter based Join Cost For Unit-selection speech Synthesis, Centre for Technology Researc,University of Edinburgh,2003.

[14] Arry Akhmad Arman,"Proses pembentukan dan karakteristik signal ucapan", Teknik Elektro ITB, Juni 2003.

[15] Arry Akhmad Arman,"Perkembangan eknologi TTS Dari masa ke Masa", Teknik Elektro ITB, 2003.

[16] L. Rabiner and B. Juang, *Fundamentals of SpeechRecognition*, Prentice Hall, 1993.

[17] Dutoit. Thierry. (1997). "*An Introduction to Text-to-Speech Synthesis*",Kluwer Academic Publisher, Dordrecht.

[18] Parsons. Thomas W. (1986). "*Voice and Speech Processing*", McGraw-Hill, New York.

[19] Pelton. Gordon E. (1993). "*Voice Processing*", McGraw-Hill, New York.

[20] Yong How Tong(2001)."Speech Processing Using Kalman Filtering" Thesis, Dep of Electrical Engineering University of Queensland.

[21] Bc.Jan Kybic (1998)," Kalman Filtering and Speech Enhancement"Diploma work,Ecole polytechnique federale De Lausanne.

[23] Thierry Dutoit, Henri Leich; 93,96 ,MBR-PSOLA:Text-To Speech Synthesis Based On an MBE-Re-Synthesis of The segments DataBase. *Faculté Polytechnique de Mons, TCTS-Multitel, WWW : http://tcts.fpms..ac.be*