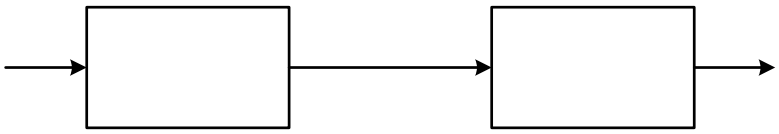
Konversi dari Teks ke Ucapan

Oleh : Arry Akhmad Arman
Peneliti dan Dosen di Departemen Teknik Elektro ITB
Email : aa@lss.ee.itb.ac.id, aa_arman@rocketmail.com

Sistem *Text to Speech* pada prinsipnya terdiri dari dua sub sistem, yaitu :

1) bagian Konverter Teks ke Fonem (*Text to Phoneme*), serta

2) bagian Konverter Fonem to Ucapan (*Phoneme to Speech*).



Bagian Konverter Teks ke Fonem berfungsi untuk mengubah kalimat masukan dalam suatu bahasa tertentu yang berbentuk teks menjadi rangkaian kode-kode bunyi yang biasanya direpresentasikan dengan kode fonem, durasi serta *pitch*-nya. Bagian ini bersifat sangat *language dependant*. Untuk suatu bahasa baru, bagian ini harus dikembangkan secara lengkap khusus untuk bahasa tersebut.

Bagian Konverter Fonem ke Ucapan akan menerima masukan berupa kode-kode fonem serta pitch dan durasi yang dihasilkan oleh bagian sebelumnya. Berdasarkan kode-kode tersebut, bagian Konverter Fonem ke Ucapan akan menghasilkan bunyi atau sinyal ucapan yang sesuai dengan kalimat yang ingin diucapkan. Ada beberapa alternatif teknik yang dapat digunakan untuk implementasi bagian ini. Dua teknik yang banyak digunakan adalah *formant synthesizer*, serta *diphone concatenation*.

Formant synthesizer bekerja berdasarkan suatu model matematis yang akan melakukan komputasi untuk menghasilkan sinyal ucapan yang diinginkan. Synthesizer jenis ini telah lama digunakan pada berbagai aplikasi. Walaupun dapat menghasilkan ucapan dengan

tingkat kemudahan interpretasi yang baik, synthesizer ini tidak dapat menghasilkan ucapan dengan tingkat kealamian yang tinggi.

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 kealamian (*naturalness*) yang tinggi.

Struktur sistem seperti di atas pada prinsipnya merupakan konfigurasi tipikal yang digunakan pada berbagai sistem Text to Speech berbagai bahasa. Namun demikian, pada setiap sub-sistem terdapat sifat-sifat serta proses-proses yang sangat spesifik dan sangat tergantung dari bahasanya.

Konversi dari teks ke fonem sangat dipengaruhi oleh aturan-aturan yang berlaku dalam suatu bahasa. Pada prinsipnya proses ini melakukan konversi dari simbol-simbol tekstual menjadi simbol-simbol fonetik yang merepresentasikan unit bunyi terkecil dalam suatu bahasa. Setiap bahasa memiliki aturan cara pembacaan dan cara pengucapan teks yang sangat spesifik. Hal ini menyebabkan implementasi unit konverter teks ke fonem menjadi sangat spesifik terhadap suatu bahasa.

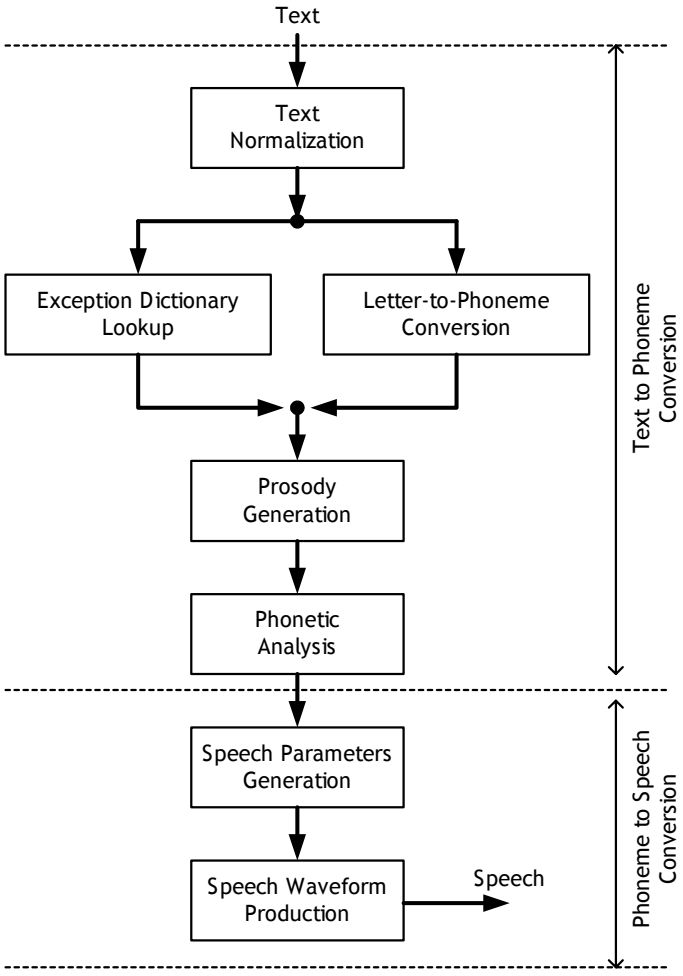
Untuk mendapatkan ucapan yang lebih alami, ucapan yang dihasilkan harus memiliki intonasi (*prosody*). Secara kuantisasi, prosodi adalah perubahan nilai *pitch* (frekuensi dasar) selama pengucapan kalimat dilakukan atau pitch sebagai fungsi waktu. Pada prakteknya, informasi pembentuk prosodi berupa data-data pitch serta durasi pengucapannya untuk setiap fonem yang dibangkitkan. Nilai-nilai yang dihasilkan diperoleh dari suatu model prosodi. Prosodi bersifat sangat spesifik untuk setiap bahasa, sehingga model yang diperlukan untuk membangkitkan data-data prosodi menjadi sangat spesifik juga untuk suatu bahasa. Beberapa model umum prosodi pernah dikembangkan, tetapi untuk digunakan pada suatu bahasa masih perlu banyak penyesuaian yang harus dilakukan.

Konverter fonem ke ucapan berfungsi untuk membangkitkan sinyal ucapan berdasarkan kode-kode fonem yang dihasilkan dari proses sebelumnya. Sub sistem ini harus memiliki pustaka setiap unit ucapan dari suatu bahasa. Pada sistem yang menggunakan teknik *diphone concatenation*, sistem harus didukung oleh suatu *diphone database* yang berisi

rekaman segmen-segmen ucapan yang berupa diphone. Ucapan dalam suatu bahasa dibentuk dari satu set bunyi yang mungkin berbeda untuk setiap bahasa, oleh karena itu setiap bahasa harus dilengkapi dengan diphone database yang berbeda.

Tahapan-tahapan utama konversi dari teks menjadi ucapan dapat dinyatakan dengan diagram seperti terlihat pada Gambar 2.7.

Tahap normalisasi teks berfungsi untuk mengubah semua teks kalimat yang ingin diucapkan menjadi teks yang secara lengkap memperlihatkan cara pengucapannya. Lihat contoh kalimat dan hasil normalisasinya pada Gambar 2.8.



Gambar 2.7. Urutan Proses Konversi dari Teks ke Ucapan
(dimodifikasi dari Pelton, 1992)

Tahap berikutnya adalah melakukan konversi dari teks yang sudah secara lengkap merepresentasikan kalimat yang ingin diucapkan menjadi kode-kode fonem. Konversi teks menjadi fonem biasanya dilakukan dengan dua cara. Sebagian proses konversi dapat dilakukan dengan aturan konversi yang sederhana dan berlaku umum untuk berbagai kondisi. Sebagian proses lainnya bersifat kondisional, tergantung dari huruf-huruf atau fonem-fonem tetangganya, bahkan terdapat bentuk-bentuk translasi yang tidak dapat ditemukan keteraturannya.

Konversi yang teratur dapat diimplementasikan dengan tabel konversi yang berisi pasangan antara urutan huruf dan urutan fonem, bahkan mungkin hanya berisi satu huruf dan satu fonem. Aturan yang lebih sulit biasanya diimplementasikan dengan tabel konversi yang akan diterapkan jika kondisi rangkaian huruf tetangga kiri dan kanannya terpenuhi. Contoh bentuk aturan konversi huruf ke fonem yang memenuhi teknik tersebut adalah sebagai berikut.

Left-context [letter-set] right-context = phoneme string

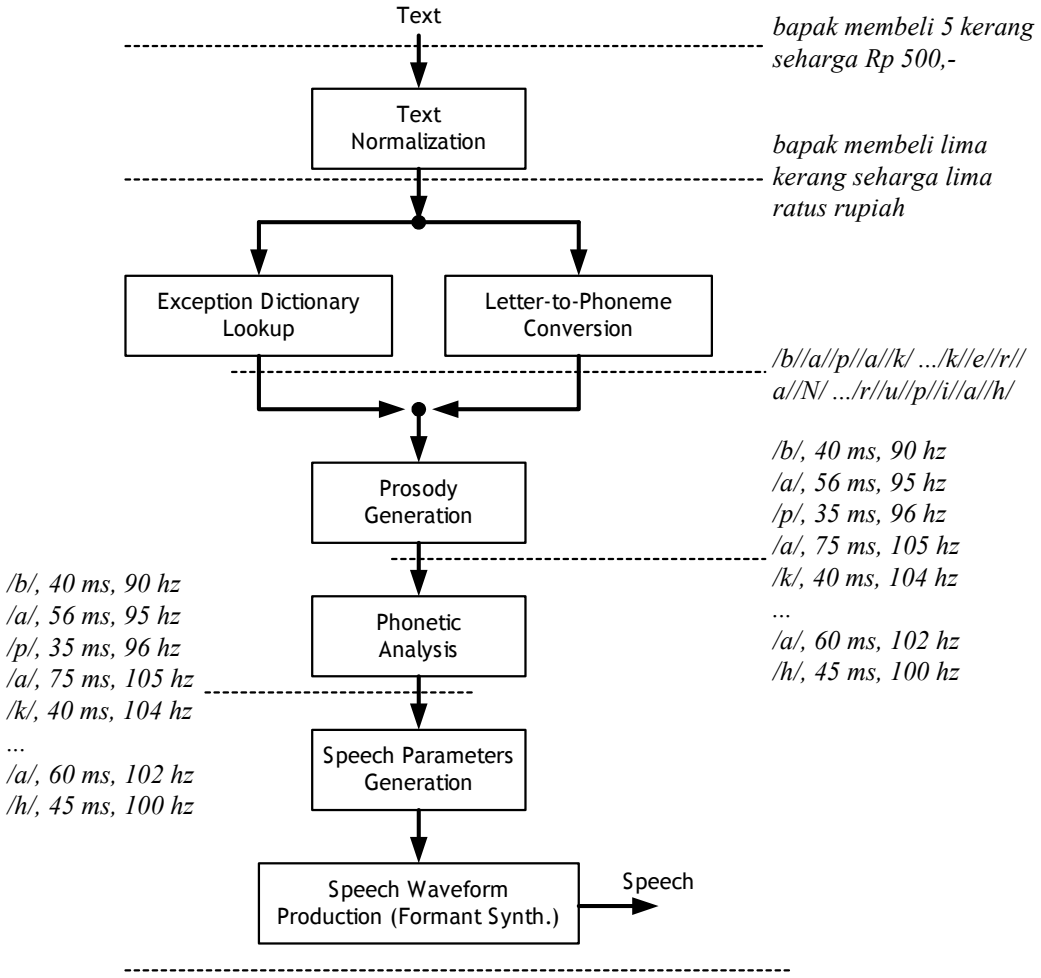
Huruf tertentu yang ditunjuk dalam posisi ***[letter-set]*** akan dikonversikan menjadi suatu fonem dalam "***phoneme string***" jika ***left-context*** dan ***right context*** terpenuhi.

Bahasa Inggris termasuk bahasa yang mempunyai keteraturan yang rendah untuk proses konversi teks ke fonem. Suatu TTS bahasa Inggris biasanya dilengkapi dengan suatu basis data yang berisi ribuan kata serta konversi padanan urutan fonemnya. Bahasa Indonesia termasuk bahasa yang jelas aturan konversinya. Sebagian besar kata dalam Bahasa Indonesia dapat dikonversikan menjadi fonem dengan aturan yang jelas dan sederhana, walaupun tetap ada kondisi-kondisi yang tidak dapat ditemukan keteraturannya. Sebagai contoh, simbol huruf e dapat diucapkan sebagai *e pepet* atau *e taling*, artinya harus dikonversikan menjadi fonem yang berbeda untuk kondisi yang berbeda. Dalam blok diagram di atas, kondisi yang masih dapat ditangani oleh aturan diimplementasikan dengan blok *Letter to Phoneme Conversion.* Konversi yang tidak teratur ditangani oleh bagian *Exception Dictionary Lookup*.

Hasil dari tahap tersebut adalah rangkaian fonem yang merepresentasikan bunyi kalimat yang ingin diucapkan. Bagian *prosody 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. Secara simbolik, hasil dari bagian ini sudah menghasilkan informasi yang cukup untuk menghasilkan ucapan yang diinginkan.

Satu 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/.



Gambar 2.8. Besaran-besaran Dalam Setiap Tahap Proses Konversi dari Teks ke Ucapan (dimodifikasi dari Pelton, 1992)

Daftar Pustaka

1. [Dut97] Dutoit. Thierry. (1997). "***An Introduction to Text-to-Speech Synthesis***", Kluwer Academic Publisher, Dordrecht.
2. [Par86] Parsons. Thomas W. (1986). "***Voice and Speech Processing***", McGraw-Hill, New York.
3. [Pel93] Pelton. Gordon E. (1993). "***Voice Processing***", McGraw-Hill, New York.