Perkembangan Teknologi TTS Dari Masa ke Masa

Penelitian di bidang pensintesa ucapan mengalami perjalanan yang sangat panjang dan telah dimulai sejak lama. Salah satu catatan literatur awal yang berhubungan dengan sintesa ucapan adalah pernyataan seorang ahli matematika dan *engineer* terkenal yang bernama Leonhard Euler pada tahun 1761. Euler menyatakan "*It would be a considerable invention indeed, that of a machine able to mimic speech, with its sounds and articulations. I think it is not imposible"*.

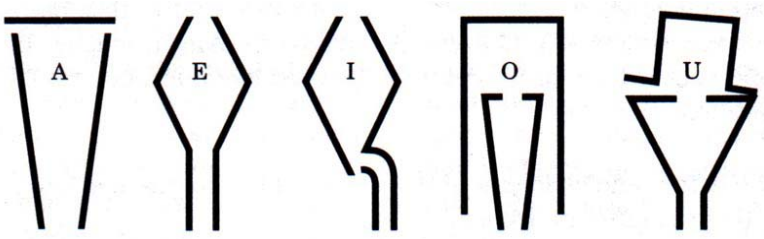
Berdasarkan hasil studi literatur dari berbagai sumber bacaan, perkembangan teknologi pensintesa ucapan dapat dibagi menjadi tiga kurun waktu. Kurun waktu pertama adalah sebelum 1930. Pada masa ini penelitian-penelitian lebih banyak ditujukan untuk memahami karakteristik sinyal ucapan serta pengembangan pensintesa ucapan berbasis mekanik atau elektromekanik. Kurun waktu kedua dimulai sejak tahun 1930-an sampai dengan ditemukannya komputer digital. Masa ini ditandai dengan pengembangan berbagai alat pensintesa ucapan menggunakan teknologi elektronik analog. Kurun waktu ketiga dimulai sejak ditemukannya komputer digital hingga sekarang. Pada masa ini, sintesa ucapan dilakukan menggunakan pendekatan-pendekatan pemrosesan sinyal digital.

Kurun Waktu Pertama

Penelitian tentang ucapan dimulai dengan penelitian-penelitian untuk melakukan pemahaman tentang sinyal ucapan. Pada tahun 1779, *Imperial Academy of St. Petersburg* menyelenggarakan suatu kompetisi dengan tujuan untuk mengetahui hal-hal berikut [Pel93].

1. "*What is the nature and character of the sounds of the vowels a, e, i, o, u that make them different from one another*?"
2. "*Can an instrument be constructed like the vox humana pipes of an organ, which shall accurate express the sounds of the vowels*?"

Seorang peneliti dari Rusia yang bernama Christian Gottlieb Kratzenstein memenangkan kompetisi tersebut dengan membuat satu set resonator akustik yang mensimulasikan mulut manusia. Resonator Kratzenstein terdiri dari 5 bentuk tabung, masing-masing untuk mensimulasikan satu bunyi vokal.

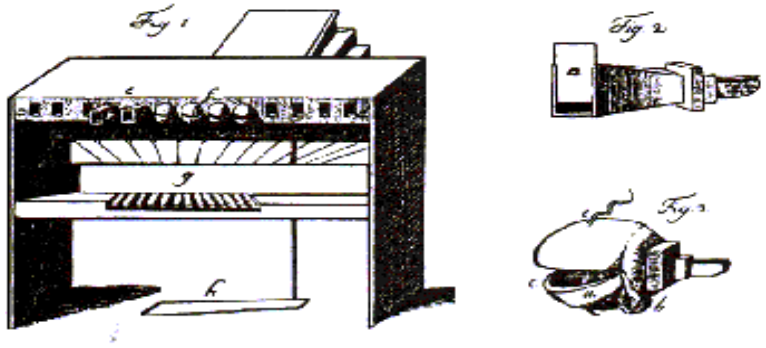


Gambar 2.1. Resonator Kratzenstein [Pel92]

Robert Willis, pada tahun 1829 melakukan penelitian yang berhasil memperlihatkan bahwa sintesa ucapan yang dihasilkan oleh Kratzenstein dapat pula dicapai dengan hasil yang sama menggunakan tabung tunggal yang dapat diatur panjangnya.

Selama dua dekade, antara tahun 1769 sampai dengan 1790, Wolfgang Ritter von Kempelen telah menghasilkan *speaking machine* yang lengkap. Pada prakteknya, Wolfgang telah membuat 3 model yang berbeda, semuanya dioperasikan dengan tangan. Penemuannya dipublikasikan dalam bentuk buku pada tahun 1791.

Wolfgang von Kempelen berpendapat bahwa untuk membuat mesin yang dapat berbicara, pertama-tama harus dapat menghasilkan suara vokal. Wolfgang mulai dengan mencari sumber bunyi yang sesuai, yaitu suatu substitusi mekanik dari suara vokal. Dia mencoba menggunakan *reed* bergetar yang biasa digunakan dalam instrumen musik, walaupun hasilnya kurang memuaskan. Suara dari *reed* disalurkan melalui suatu alat berbentuk bel yang dilengkapi *baffle* pada mulut yang dapat digerakan untuk menghasilkan bunyi vokal yang berbeda. Tidak puas dengan hasil percobaannya yang pertama, von Kempelen menggunakan tangannya untuk menggantikan *baffle*. Meskipun hasilnya menjadi lebih baik, tetapi suara yang dihasilkan masih belum memuaskan.



Gambar 2.2. Model Kedua Pensintesa Ucapan
Buatan Wolfgang von Kempelen [Pel92]

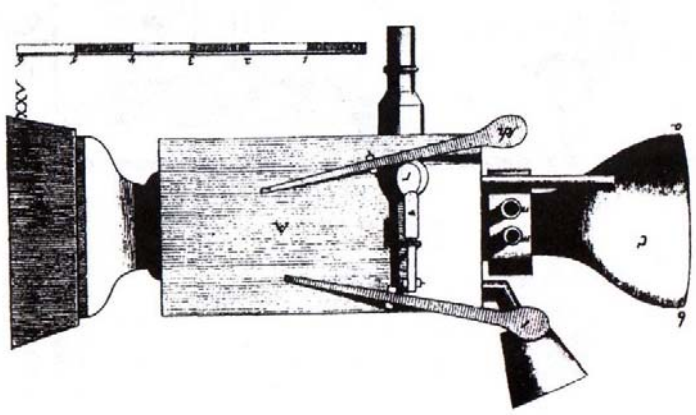
Model yang kedua dirancang untuk memenuhi kebutuhan akan perlunya beberapa resonansi pada beberapa frekuensi yang berbeda untuk mencapai berbagai suara berlainan yang diinginkan. Versi ini bersifat modular, berupa tiga belas buah resonator yang masing-masing dilengkapi dengan *reed* dan bersifat dapat dibongkar pasang, sehingga dapat saling dipertukarkan. Gambar 2.2 memperlihatkan model tersebut.

Dengan mesin tersebut, von Kempelen mengklaim bahwa dia telah mampu menghasilkan suara vokal a, o dan u serta suara p, m dan l yang dapat diterima. Secara monotonik, mesin buatannya dapat mengucapkan suara seperti “mama” dan “papa”, tetapi masih menghadapi dua masalah utama. Pertama, suara vokal yang dihasilkan mengandung bunyi yang sifatnya eksplosif yang mirip bunyi “k”. Masalah lain yang dihadapi adalah transisi antara dua bunyi yang berdekatan yang tidak *smooth* seperti suara alami. Satu bunyi dengan bunyi berikutnya masih terasa sebagai dua bunyi yang terpisah. Untuk mengatasi masalah tersebut, dia menambahkan kulit halus pada *reed*, juga menggunakan *reed* tunggal sebagai pengganti dari sejumlah *reed* yang sebelumnya digunakan pada setiap resonator.

Mesin ketiga buatan von Kempelen secara fisik sangat berbeda dari mesin-mesin sebelumnya (lihat Gambar 2.3). Paru-paru disimulasikan dengan pompa yang digerakan dengan bahu yang secara kontinyu dapat menghembuskan udara. Vokal dapat dihasilkan dengan cara menutup “nostrils” mesin tersebut dengan tangan kanan sambil menghembuskan udara dari simulator paru-paru. Sementara itu, tangan kiri harus mengatur resonansi melalui alat berbentuk bel. Hanya orang yang terlatih memainkannya

yang dapat menghasilkan bunyi-bunyi yang diharapkan. Suara seperti F, H, V, W dan beberapa lainnya adalah suara-suara yang juga dapat dihasilkan dengan mesin tersebut. Wolfgang mengklaim bahwa mesin ketiga buatannya dapat menghasilkan semua suara vokal serta sembilan belas konsonan. Meskipun mesin tersebut memiliki kapasitas menghasilkan udara sekitar enam kali lebih besar dari kapasitas paru-paru manusia, tetapi mesin ini hanya mampu mengucapkan kalimat yang pendek sebelum kehabisan udara. Pada tahun 1791 von Kempelen mempublikasikan hasil penelitiannya dalam bahasa Jerman dan Perancis dengan judul "*Mechanismus der menschlichen Sprache nebst der Berschreibung seimer sprechenden Maschine"*.

Di Perancis, pada waktu yang hampir bersamaan dengan von Kempelen, Abbe' Mical mengembangkan mesin lain yang dikenal sebagai "*two talking head*". Mesin ini terdiri dari dua silinder yang mirip dengan silinder yang biasa kita lihat pada instrumen musik. Satu silinder disediakan untuk memainkan sejumlah ucapan tertentu dengan *prosody*-nya. Silinder lainnya digunakan untuk menghasilkan semua bunyi dalam bahasa Perancis. Tidak diketahui dengan pasti otentikasi mesin buatannya tersebut.

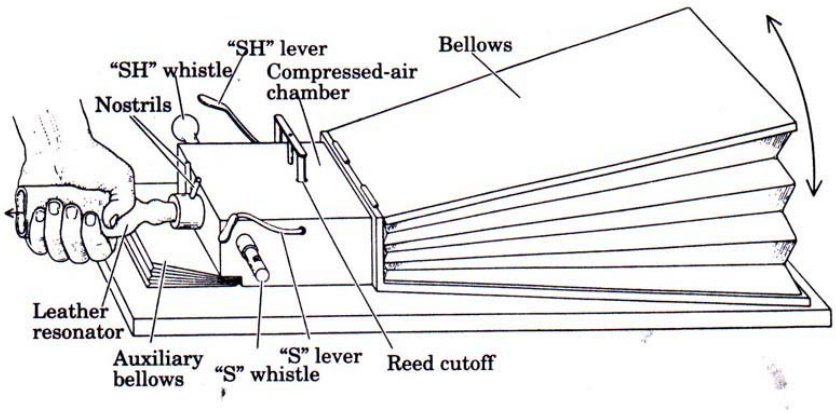


Gambar 2.3. Model Ketiga Pensintesa Ucapan
Buatan Wolfgang von Kempelen [Pel92]

Hermann Helmholtz, seorang perintis peneliti akustik, pada pertengahan abad ke-19 membuat perangkat elektro-mekanik yang terdiri dari sejumlah garpu yang dapat ditala, kumparan elektrik, dan sejumlah resonator yang dapat mensintesa suara komposit yang sangat mirip suara vokal manusia. Perangkat ini mungkin tidak memperlihatkan

hubungan langsung dengan berbagai penemuan alat-alat lainnya yang berhubungan dengan aplikasi suara, tetapi keberadaan mesin tersebut memberikan ilham bagi Alexander Graham Bell yang menghasilkan beberapa penemuan di bidang aplikasi mesin yang berhubungan dengan suara manusia. Pada saat yang bersamaan juga, Hermmann Helmholtz telah melakukan berbagai penelitian yang memberikan pemahaman yang lebih mendalam tentang akustik.

Peranan Sir Charles Wheatstone yang lebih dikenal dengan "Jembatan Wheatstone"-nya tidak dapat diabaikan dalam perkembangan alat pensintesa ucapan manusia. Wheatstone tumbuh sambil membantu bisnis penjualan perangkat musik keluarganya di London. Tahun 1821, pada usia sembilan belas tahun ia mendemonstrasikan alat ciptaannya yang dapat menggetarkan batang logam yang dieksitasi oleh suatu sumber yang vibrasinya dirambatkan melalui konduktor yang padat. Pada tahun 1835, Wheatstone mendemonstrasikan ciptaannya kepada Dublin Association.



Gambar 2.4. Versi Wheatstone dari Model Ketiga Pensintesa Ucapan Buatan Wolfgang von Kempelen [Pel92]

Alexander Graham Bell yang lahir di Edinburg pada tahun 1846 dikenal sebagai penemu telpon. Berdasarkan buku yang ditulis oleh Kempelen, Bell beserta dua saudaranya (Melly dan Ted) pernah melakukan pengembangan mesin yang dapat menirukan ucapan-ucapan manusia. Pengembangan tersebut dilakukan di Edinburg sekitar tahun 1863. Pada usia 19 tahun, Bell mencoba mengulangi penelitian akustik Helmholtz. Bell mengira

bahwa garpu tala dapat mentransmisikan bunyi vokal secara elektrik. Untuk memperbaiki kesalahan dugaan tersebut, akhirnya dia menemukan suatu keyakinan bahwa suara apapun dapat ditransmisikan secara elektrik. Pada akhirnya, Bell berhasil menemukan telpon.

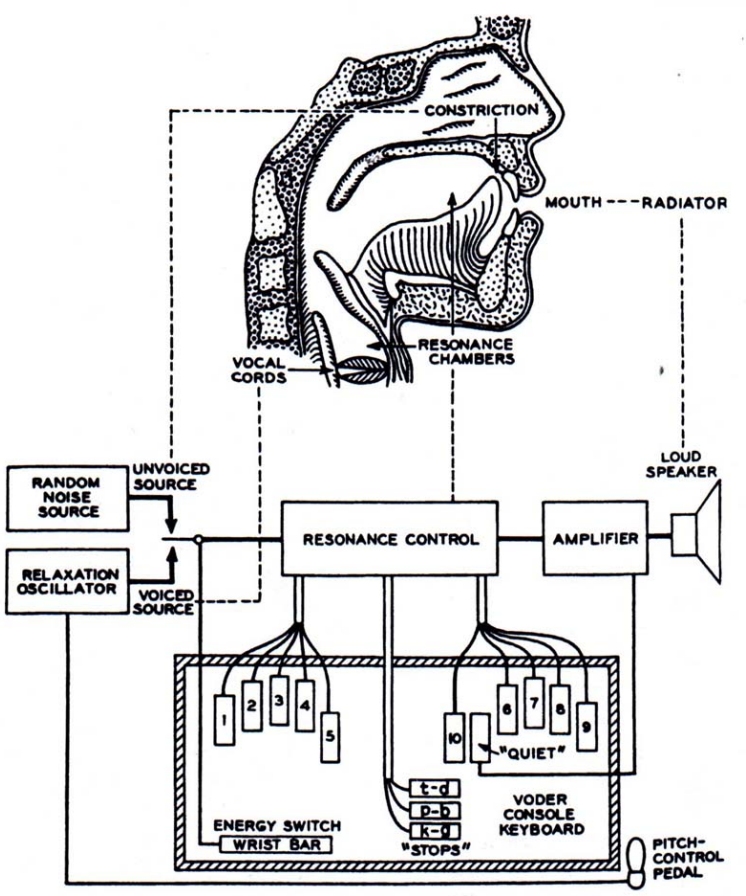
Pada awal tahun 1990-an, J. L. Flanagan melaporkan hasil kerjanya yang merupakan kelanjutan dari pemikiran Helmholtz dan menguji berbagai alat yang dapat melakukan sintesa suara vokal. Penelitian ini meliputi penggunaan pipa organ, multiple sirens, garpu vibrasi yang dapat ditala, serta ide R. R. Riesz yang pada tahun 1937 mengusulkan alat bicara mekanik yang dapat dioperasikan dengan jari-jari tangan.

Kurun Waktu Kedua

Sejak 1930 para peneliti mulai menggunakan model elektrik untuk analisis dan menirukan ucapan. Pensintesa elektrik pertama yang berfungsi untuk menghasilkan ucapan adalah Dudley's voder. VODER (*Voice Operated DEmonstratoR*) dikembangkan oleh Bell Laboratories. VODER merupakan sistem elektronik analog yang mensimulasikan bagian-bagian alat ucap manusia. VODER pertama kali diperkenalkan kepada publik dalam suatu pameran di New York pada tahun 1939. Pada saat tersebut berhasil didemonstrasikan bagaimana manusia dapat berdialog dengan mesin VODER yang dimainkan oleh seorang operator.

Gambar 2.5. VODER dalam New York World's Fair
pada Tahun 1939 [Pel92]

Gambar 2.6 memperlihatkan blok diagram VODER serta ekivalensinya dengan alat-alat ucap manusia. Suara bersumber dari dua buah sumber bunyi, yaitu : noise dan osilator. Sumber noise disediakan untuk mensintesa ucapan yang menyerupai noise, sedangkan osilator untuk ucapan lainnya. Frekuensi osilator dikendalikan oleh pedal. Frekuensi yang dihasilkan akan menentukan *pitch* dari bagian ucapan yang dihasilkan. Sumber yang dihasilkan akan dilewatkan pada sepuluh bandpass filter yang dihubungkan secara paralel dan masing-masing frekuensinya dapat diatur. Tiga pengatur lainnya disediakan untuk mengatur proses transien, yaitu untuk reproduksi konsonan stop, yaitu t, d, p, b, k, g. Mesin ini berhasil membangkitkan suara yang *intelligible*. Mesin ini harus dimainkan oleh seorang operator yang sangat terlatih.



Gambar 2.6. Ekivalensi VODER dengan
Alat Ucap Manusia [Pel92]

Kurun Waktu Ketiga

Selama 50 tahunan, teknologi pensintesa ucapan mengalami banyak perubahan. Penemuan komputer digital telah memungkinkan untuk melakukan simulasi sebelum melakukan pengembangan perangkat keras. Sekitar tahun 1960-an, teknik analisis dan sintesa ucapan terbagi menjadi dua paradigma. Pendekatan pertama disebut *articulatory synthesis*. Dalam pendekatan ini, mekanisme produksi ucapan dimodelkan secara fisiologi dengan cukup rinci. Pendekatan lainnya disebut *terminal-analogue synthesis*. Pada pendekatan kedua ini, ucapan dimodelkan dengan model apapun. Orientasinya lebih ditekankan pada usaha untuk memodelkan sinyal ucapan, bukan pada bagaimana cara membangkitkannya.

Sebelum adanya komputer digital, sebenarnya belum ada sistem seperti yang sekarang kita kenal sebagai sistem TTS. Pengembangan yang ada saat itu hanya terbatas pada bagian untuk membangkitkan atau mensintesa ucapannya saja. TTS yang melakukan konversi secara otomatis dari mulai teks berkembang setelah adanya komputer digital.

Pada tahun 1931, perusahaan Audichron membuat mesin pertama yang secara otomatis dapat mengucapkan waktu dan temperatur melalui saluran telpon. Sejak itu, banyak dikembangkan perangkat elektrik yang berhubungan dengan aplikasi ucapan, diantaranya adalah spektograf suara yang dapat menampilkan pola ucapan pada layar CRT.

Salah satu sistem komersial yang menerapkan teknologi komputer digital untuk aplikasi pemrosesan ucapan adalah IBM 7770 Audio Response Unit yang menggunakan *drum* berputar untuk menyimpan data-data ucapan. Pada awal tahun 1980-an berkembang beberapa sistem lainnya yang menggunakan komputer mainframe atau komputer mini. Dengan sistem ini, sejumlah institusi finansial saat itu dapat memberikan layanan sistem otomatis melalui pesawat telpon. Keadaan tersebut berubah semakin cepat setelah teknologi IC serta komputer mikro berkembang dengan pesat.

Berkembangnya komputer digital tidak hanya menyebabkan berkembangnya sistem TTS, tetapi juga melahirkan alternatif-alternatif baru untuk mengimplementasikan bagian pembangkit ucapannya. Pada era komputer digital, pembangkitan ucapan dilakukan menggunakan algoritma-algoritma pemrosesan sinyal digital yang diimplementasikan menggunakan perangkat lunak.

Bentuk pensintesa digital yang berkembang pada awalnya adalah pensintesa yang dikenal dengan istilah *formant synthesizer*, bekerja dengan cara mensimulasikan komponen-komponen frekuensi utama pembentuk ucapan yang disebut formant. Salah satu pensintesa ucapan jenis ini yang populer dan banyak digunakan pada berbagai aplikasi adalah *cascade-parallel formant synthesizer* yang pertama kali diusulkan oleh Dennis Klatt pada tahun 1990. Synthesizer tersebut merupakan pengembangan dari generasi sebelumnya yang juga dirancang oleh Klatt pada tahun 1980.

Pensintesa formant tidak dapat menghasilkan suara dengan tingkat kealamian yang tinggi, sehingga perkembangan TTS mengarah pada pencarian alternatif untuk mencari pendekatan yang dapat menghasilkan ucapan yang lebih alami. Seiring dengan kecepatan

prosesor serta media penyimpanan komputer yang semakin tinggi, pendekatan tersebut mengarah pada sistem yang melakukan penggabungan segmen-segmen ucapan yang direkam sebelumnya. Berdasarkan berbagai pertimbangan teknis dan kualitas yang ingin dicapai, bentuk segmen yang dianggap paling optimum dan banyak digunakan adalah *diphone* atau dua fonem yang berurutan. Pendekatan dengan cara penyusunan ucapan dari diphone ini disebut *diphone concatenation*.

Tantangan teknis utama pada teknik *diphone concatenation* 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 autoregressive (AR), Glottal AR, hybrid harmonic/stocastic, time domain PSOLA (TD-PSOLA), multiband resynthesis-PSOLA (MBR-PSOLA), serta Linear Prediction-PSOLA (LP-PSOLA) [Dut97].

Kini, speech synthesizer berkualitas tinggi telah tersedia untuk sejumlah bahasa, misalnya Bahasa Inggris, Perancis, Belanda, Jerman dan beberapa bahasa lainnya. Namun demikian, speech synthesizer untuk bahasa Indonesia sampai saat ini belum tersedia. Salah satu perusahaan yang telah menghasilkan TTS berkualitas baik adalah perusahaan *Lernout and Hauspie* di Belgia. Perusahaan tersebut sudah memproduksi sistem TTS berkualitas tinggi untuk bahasa Inggris, Jerman, Perancis, Belanda, Spanyol dan Portugis.