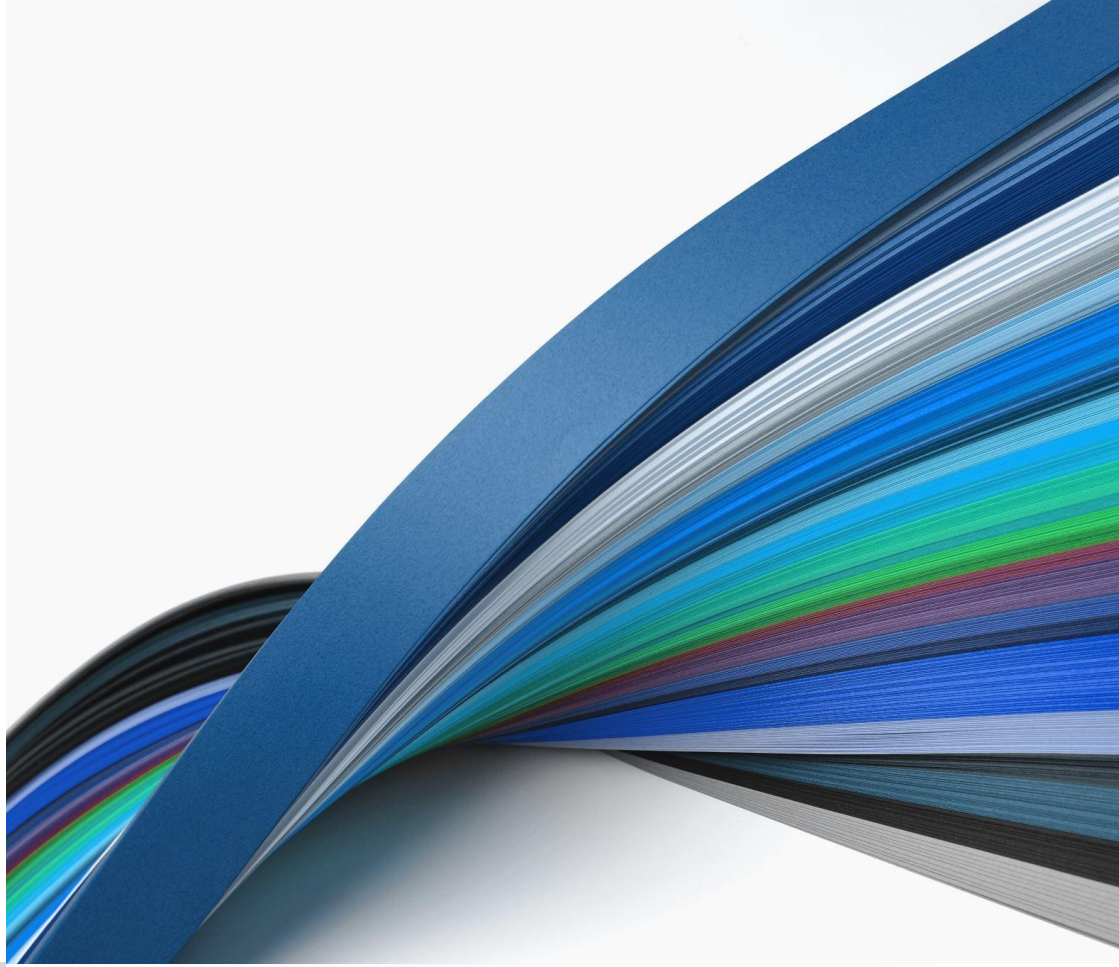
Sintesis Bentuk Gelombang



Pemrosesan Bahasa Lisan

Fakultas Ilmu Komputer Universitas Indonesia

Semester Gasal 2024/2025

Referensi

- TTS Waveform Synthesis – Andrew Maas
- Text-to-Speech Synthesis – Ondřej Dušek
- Speech Synthesis – Preethi Jyothi
- https://speechprocessingbook.aalto.fi/

Sintesis bentuk gelombang: ikhtisar

- Membangun sistem teks-ke-ucapan
- Sintesis berbasis forman
- Sintesis konkatenatif
  - Sintesis Difone
  - Sintesis seleksi unit
- Sintesis parametrik

Dua Tahap dalam Sintesis

TEXT: PG&E will file schedules on April 20th

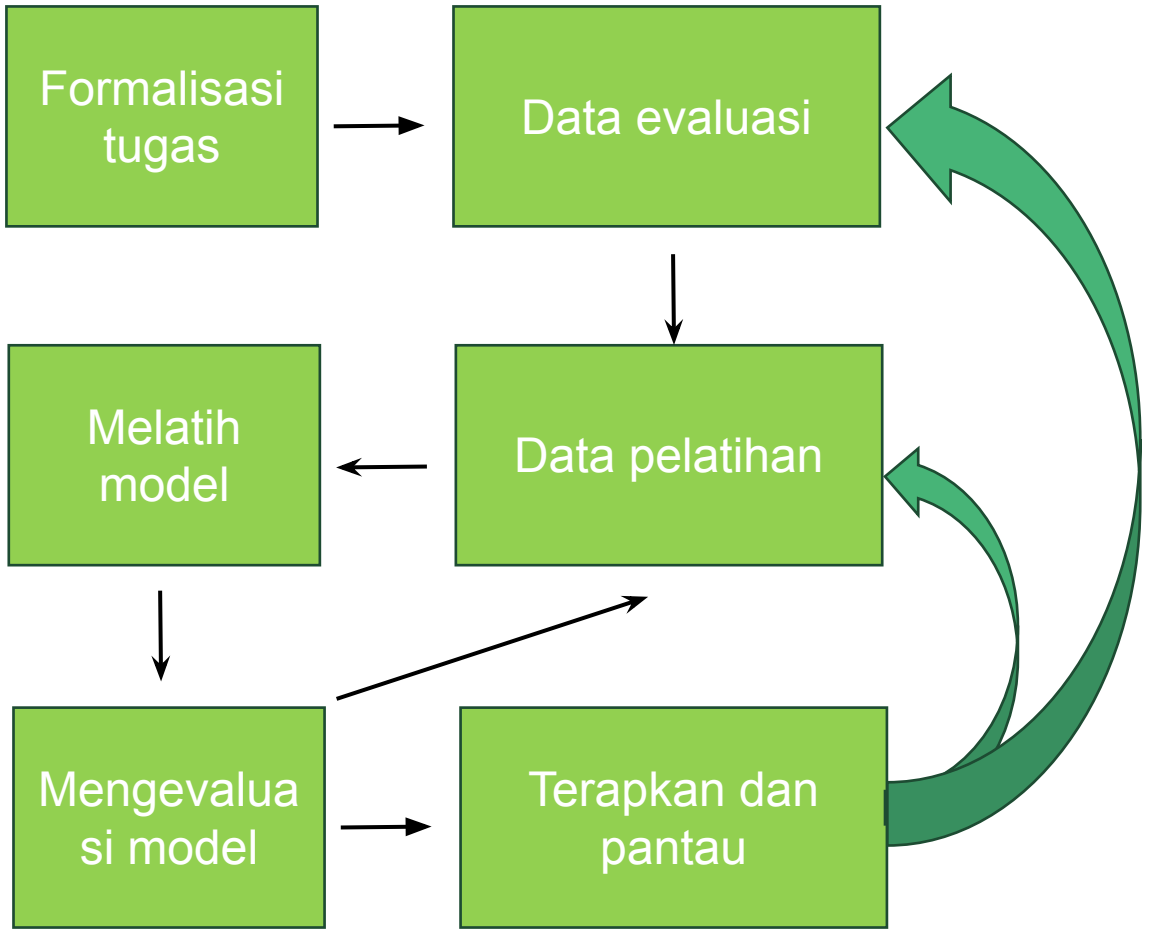
1. Text analysis: Apa yang ingin dikatakan?
   - Teks -> representasi perantara

2. Waveform synthesis: Bagaimana cara mengatakannya?
   - Representasi perantara -> gelombang suara audio

Text-to-speech modern bergantung pada pembelajaran mesin

- Mencocokkan data pelatihan + arsitektur sistem dengan penggunaan yang direncanakan
- Kumpulan data pelatihan berkualitas tinggi dengan cakupan luas untuk mencapai suara yang diinginkan
- Memanfaatkan model yang ada sebagai titik awal jika berlaku
- Jadi, bagaimana kita membangun sistem ML hebat yang menggunakan audio dari satu orang atau lebih?

Mengembangkan sistem text-to-speech berbasis ML



Membangun sistem text-to-speech yang hebat

- Evaluasi & pengukuran:
  - Pilih kriteria (alami, emosional?)
  - Siapkan tes mendengarkan evaluasi manusia
- Pengumpulan data:
  - Kualitas akustik dibatasi oleh data yang dikumpulkan
  - Memerlukan rentang emosi, ekspresi
- Modeling:
  - Sistem pembelajaran mendalam bekerja paling baik
  - Sistem konkatenatif lebih mudah dibangun dengan cepat
  - Desain antarmuka yang dapat dikontrol untuk pengembang

Evaluasi TTS (Text-to-Speech)

- Evaluasi TTS umumnya memerlukan manusia!
  - Paradigma Tes Mendengarkan: Dengarkan contoh ucapan, lalu nilai berbagai aspek (naturalness, intelligibility, friendliness, expressiveness, dll.) pada skala 1-5.
  - Mean Opinion Score (MOS): Rata-rata dari semua penilaian.
  - Tes AB (Preferensi A/B): Memilih preferensi antara A atau B.
- Intelligibility Tests
  - Apakah pendengar mendengar dengan benar? Ini bisa dilakukan dengan tugas *completion*, menulis apa yang didengar, atau sekadar memberi penilaian.
- Overall Quality Tests
  - Tes preferensi A/B dibandingkan dengan narator manusia sebagai "batas atas" kualitas

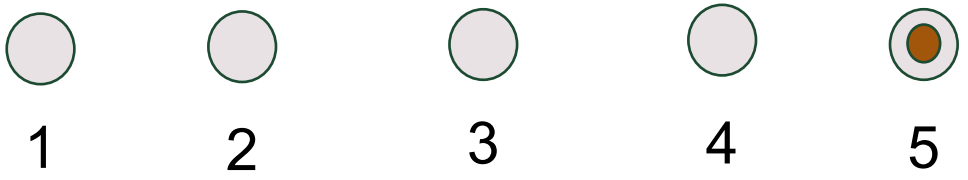
Mean opinion score (MOS)

Menggunakan pemeringkatan hasil sintesis yang dilakukan secara crowdsourced berdasarkan:

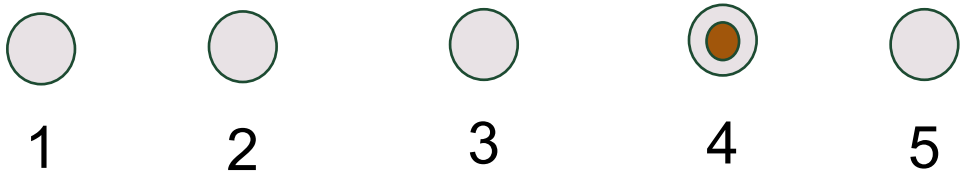
- Kejelasan
  - Biasanya diukur secara objektif melalui transkripsi
- Keterpahaman
  - Seberapa mudahkah memahami suatu ucapan tertentu?
- Kealamian
  - Seberapa alami ucapan itu terdengar?
- Ekspresi
  - Seberapa baik intonasi sesuai dengan substansi ucapan?

Transkripsikan ucapan tersebut:
"The golden sun dipped below the horizon, casting a warm glow over the tranquil sea."

dapat dipahami



kealamian



A/B testing

- Menggunakan pilihan yang bersumber dari banyak orang untuk memperoleh preferensi langsung

Ucapan 1

Ucapan 2

Ucapan manakah yang lebih Anda sukai?
(Yang mana yang lebih mudah dipahami dan terdengar lebih alami)

1 2

Diagnostic rhyme test (DRT)

- Manusia melakukan identifikasi pilihan mendengarkan antara dua kata yang berbeda dengan satu fitur fonetik
  - Suara, nasalitas, sutensi, desisan
- 96 pasangan rima
- % jawaban yang benar adalah skor kejelasan

Karakteristik	Keterangan	Contoh
Suara	bersuara - tak bersuara	veal - feel, dense - tense
Nasalitas	nasal - oral	reed - deed
Sustensi	berkelanjutan - terputus-putus	vee - bee, sheat - cheat
Sibilasi	bersuara - tak bersuara	sing - thing
Ketegasan	berat - tajam	weed - reed
Kekompakan	padat - menyebar	key - tea, show - sow

Pengumpulan data

- Kualitas akustik yang bagus
  - Minimal 16 kHZ
  - Mikrofon yang bagus
  - Kebisingan latar belakang minimal (termasuk membalik halaman dan bernapas!)
- Rentang emosional dan fonetik sesuai dengan aplikasi
  - Sistem akan mengkloning aksen pembicara tunggal
  - Harus mengumpulkan ucapan emosional jika diperlukan
  - Membaca vs berbicara berbeda
  - Simulasikan percakapan antarmanusia dengan permainan peran mungkin
- Data yang cukup
  - ~10 jam mungkin cukup untuk membaca ucapan
  - Transfer leaning memungkinkan beberapa pembagian data

Contoh: Rekaman untuk Google Assistant

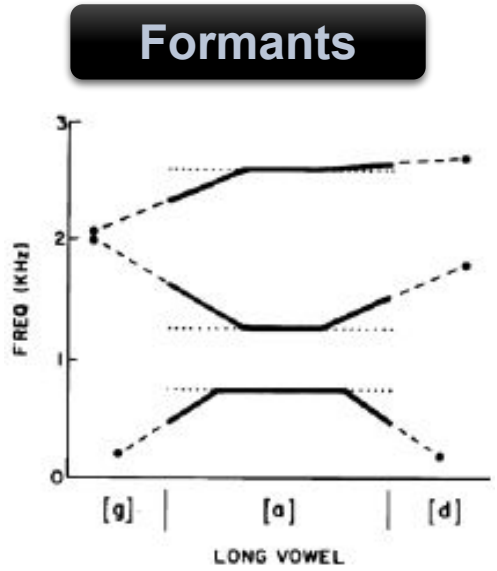
- Kondisi perekaman yang bagus, perhatian pada prosodi, unit untuk frasa umum
- Lebih penting untuk sistem pemilihan unit, tetapi kualitas data dapat menjadi faktor pembatas untuk text-to-speech modern
- Tonton: https://drive.google.com/file/d/1vayhixbgUypP3gICN_xmYJEghh4PcAng/view?resourcekey

Sintesis bentuk gelombang: ikhtisar

- Membangun sistem teks-ke-ucapan
- Sintesis berbasis forman
- Sintesis konkatenatif
  - Sintesis Difone
  - Sintesis seleksi unit
- Sintesis parametrik

Sintesis berbasis forman

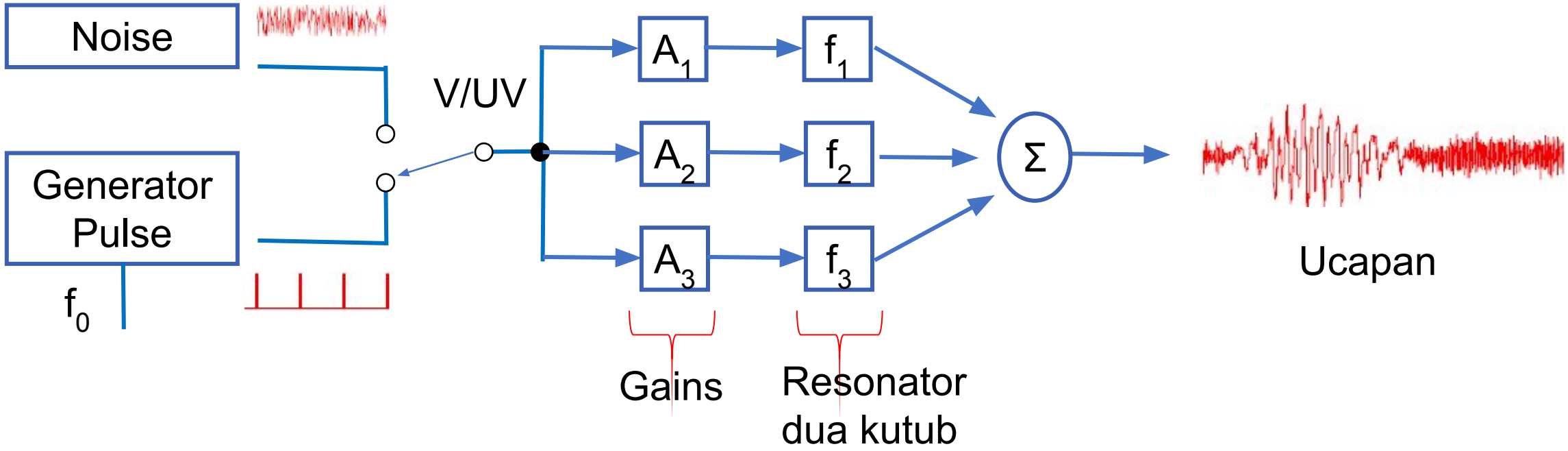
- Sistem awal
- Aturan untuk menyusun gelombang suara keluaran
  - Berdasarkan resonator forman + komponen tambahan (filter sumber)
  - Aturan untuk kombinasi suara (misalnya "b sebelum vokal bulat belakang")
  - Aturan untuk suprasegmental – nada, kenyaringan, dll.
- Hasilnya tidak terlalu alami, tetapi sangat mudah dipahami pada akhirnya
- Jejak perangkat keras sangat rendah
- Contoh: Speak & Spell, DECTalk



Sintesis berbasis forman

Pada contoh ini synthesizer dikontrol oleh 8 parameter:

f_0 V/UV A_1 A_2 A_3 f_1 f_2 f_3



Perhatikan bahwa dalam kasus ini amplitudo dan frekuensi dikontrol tetapi bandwidthnya tetap. Parameter akan diperbarui kira-kira setiap 10 ms.

Sintesis bentuk gelombang: ikhtisar

- Membangun sistem teks-ke-ucapan
- Sintesis berbasis forman
- Sintesis konkatenatif
  - Sintesis Difone
  - Sintesis seleksi unit
- Sintesis parametrik

Membangun Skema Diphone

- Temukan daftar fonem dalam bahasa:
  - Tambahkan alofon menarik
  - Perhatikan stress, ton, kluster, onset/coda, dll.
  - Tambahkan fonem asing (langka)
- Bangun pembawa untuk:
  - Konsonan-vokal, vokal-konsonan
  - Vokal-vokal, konsonan-konsonan
  - Hening-fonem, fonem-hening
  - Kasus khusus lainnya
- Periksa hasilnya:
  - Daftar semua diphone dan jelaskan yang hilang
  - Setiap daftar diphone memiliki kesalahan

Recording Conditions

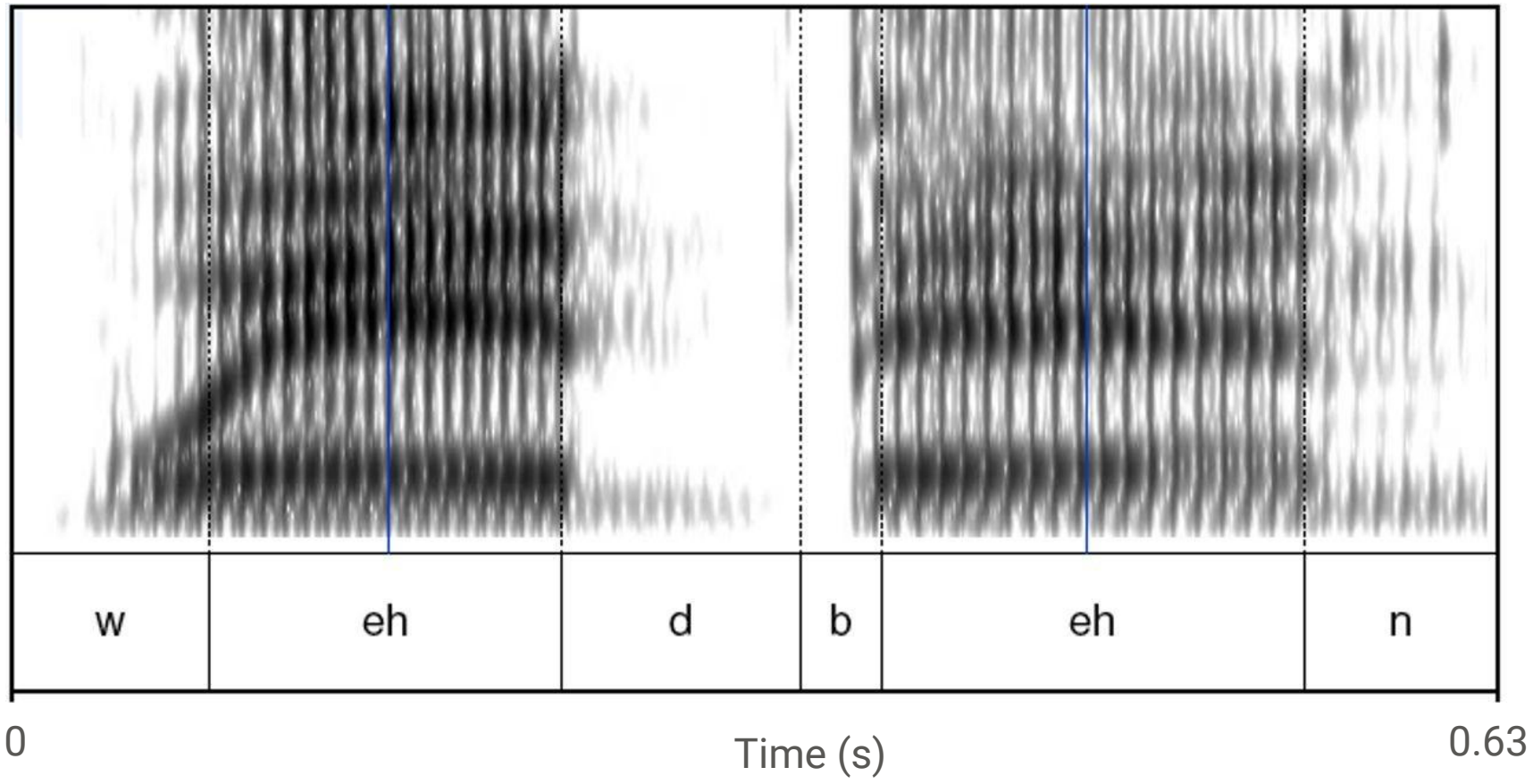
- Ideal:
  - Ruangan anechoic (tanpa gema)
  - Rekaman kualitas studio
  - Sinyal EGG (Electroglottography)
- More likely:
  - Ruangan tenang
  - Mikrofon murah/sound blaster
  - Tanpa EGG
  - Mikrofon yang dipasang di kepala
- Yang bisa dilakukan:
  - Kondisi yang dapat diulang
  - Pengaturan tingkat audio dengan cermat

Diphones

- Bagian tengah fonem lebih stabil daripada pinggirnya.
- Membutuhkan sekitar~|phones|^2 unit.
  - Beberapa kombinasi tidak ada (semoga).
  - Sistem ATT (Olive et al., 1998) memiliki 43 fonem
    - 1849 kemungkinan diphone.
    - Fonotaktik: Fonem [h] hanya muncul sebelum vokal, tidak perlu menyimpan diphone melintasi hening.
    - Hanya 1172 diphone yang benar-benar ada.
  - Mungkin memasukkan stress, kluster konsonan, sehingga bisa lebih banyak.
  - Banyak pengetahuan fonetik dalam desainnya.
- Database relatif kecil (standar saat ini)
  - Sekitar 8 megabytes untuk English (16 KHz 16 bit)

Diphones

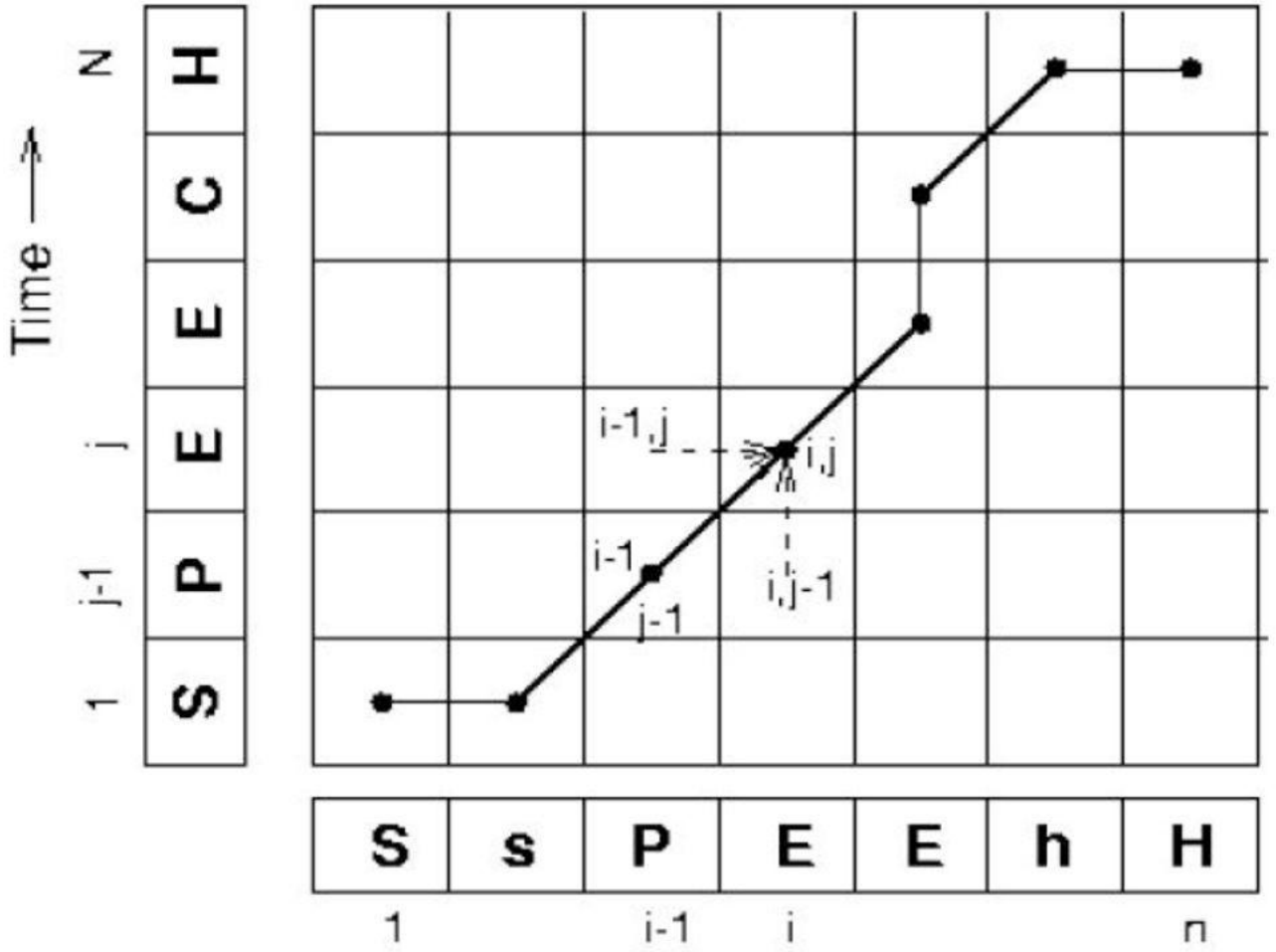
- Bagian tengah fonem lebih stabil daripada bagian tepi:



Labeling Diphones

- Jalankan pengenal suara dalam mode forced alignment.
  - Forced alignment:
    - Diberikan: Sistem ASR yang telah dilatih, file wav, dan transkrip.
    - Menghasilkan: Alignment fonem ke file wav
- Keuntungan dibandingkan pelabelan fonetik manual:
  - Urutan kata dan fonem sudah didefinisikan.
  - Artikulasi jelas.
  - Namun terkadang pembicara masih mengucapkan dengan salah, jadi perlu dicek.
- Batas fonem kurang penting
  - +- 10 ms sudah cukup baik
- Batas midphone penting
  - karena bagian tengah lebih stabil.
  - Dapatkah bagian stabil ditemukan secara otomatis?

Dynamic Time Warping



Menggabungkan Diphone: Junctures

- Jika gelombang suara sangat berbeda, akan terdengar klik di sambungan.
  - Solusi: Gunakan windowing.
- Jika kedua diphone bersuara (voiced), perlu disambung secara sinkron dengan pitch.
- Artinya, kita perlu tahu di mana setiap periode pitch dimulai, agar dapat menyambungkan pada titik yang sama di setiap periode pitch.
  - Penandaan Pitch atau Deteksi Epoch: Tandai tempat setiap pulsa atau epoch pitch terjadi.
    - Menemukan Instant of Glottal Closure (IGC).
    - (Perhatikan perbedaannya dengan pelacakan pitch).

Sintesis Waveform/Gelombang Suara

Diberikan:

- Rangkaian fonem
- Prosodi
  - F0 yang diinginkan untuk keseluruhan ucapan
  - Durasi untuk setiap fonem
  - Nilai stress untuk setiap fonem, mungkin juga nilai aksen

Hasilkan:

- Waveform

F0 Generation

- Berdasarkan aturan
- Menggunakan regresi linear atau machine learning
- Beberapa batasan:
  - Berdasarkan aksen dan batas-batas kalimat
  - F0 menurun secara bertahap selama suatu ucapan (“declination”)

F0 Generation by Rule

- Hasilkan daftar titik target F0 untuk setiap suku kata. Misalnya:
- Hasilkan aksen sederhana H* "hat" (nilai F0 tetap spesifik untuk pembicara) dengan 3 titik pitch: [110, 140, 100]
  - Modified by
  - gender,
  - declination,
  - end of sentence,
  - etc.

F0 Generation by Regression

- Menggunakan pembelajaran mesin terawasi
- Prediksi: Nilai F0 di 3 posisi dalam setiap suku kata
- Fitur Prediktor:
  - Accent of current word, next word, previous
  - Boundaries
  - Syllable type, phonetic information
  - Stress information
- Dibutuhkan: Set pelatihan dengan aksen pitch yang telah dilabeli
- Definisi F0: Biasanya didefinisikan relatif terhadap rentang pitch
- Rentang antara frekuensi baseline dan topline dalam sebuah ucapan
- Sistem modern menggunakan ML untuk mempelajari generasi F0

Speech as Short Term Signals

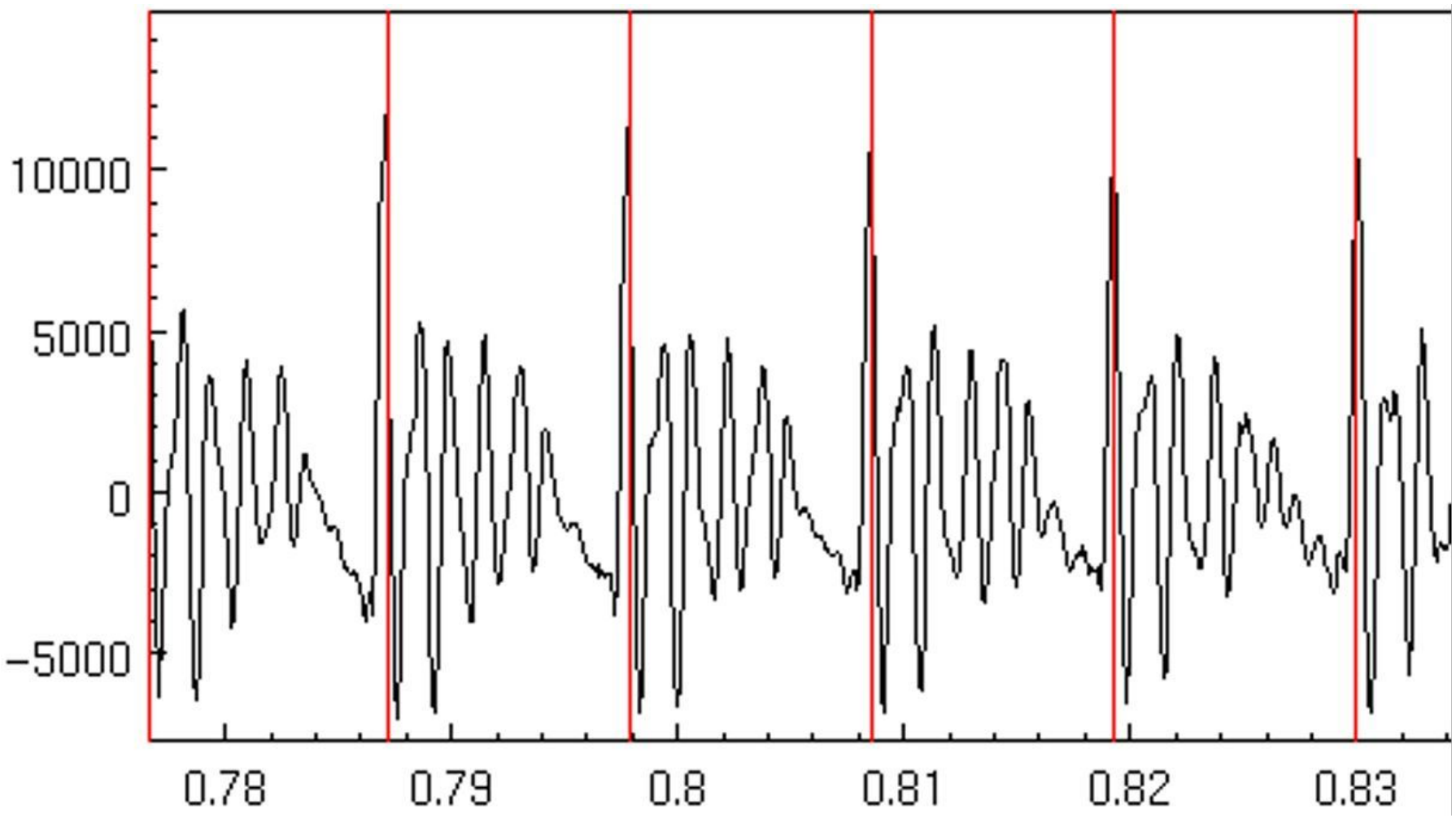
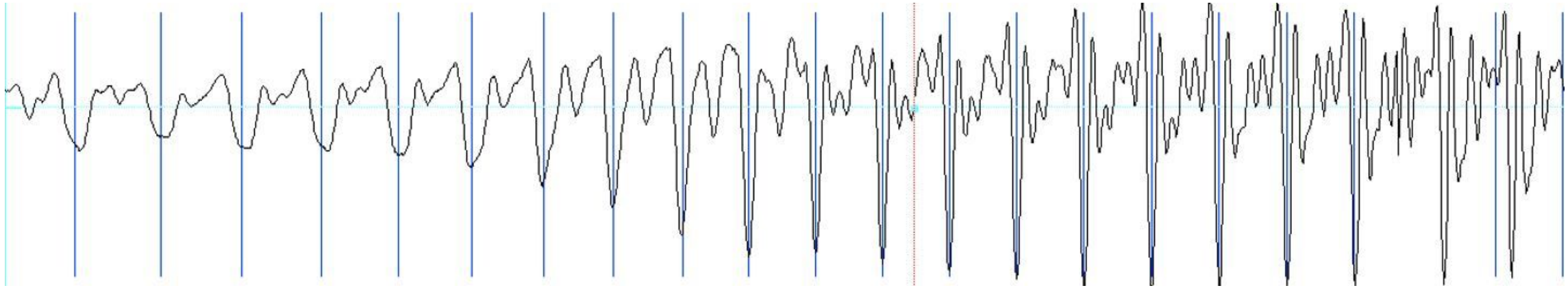


Figure: Alan Black

Epoch-labeling

- An example of epoch-labeling using “SHOW PULSES” in Praat:



Epoch-labeling: Electroglottograph (EGG) = Laryngograph, Lx

- Dipasang di leher pembicara dekat laring
- Mengirimkan arus frekuensi tinggi melalui jakun
- Jaringan manusia menghantarkan arus dengan baik; udara tidak sebaik itu
- Transduser mendeteksi seberapa terbuka glotis (yaitu jumlah udara di antara pita suara) dengan mengukur impedansi.

Gambar: UCLA Phonetics Lab

Modifikasi Prosodik

- Memodifikasi pitch dan durasi secara independen
- Jika mengubah laju sampel, kedua aspek berubah:
  - Chipmunk speech
- Durasi: Gandakan/hapus bagian sinyal untuk menyesuaikan durasi
- Pitch: Ulangi sampling untuk mengubah pitch

Modifikasi Durasi

- Duplikasi/hapus short term signals

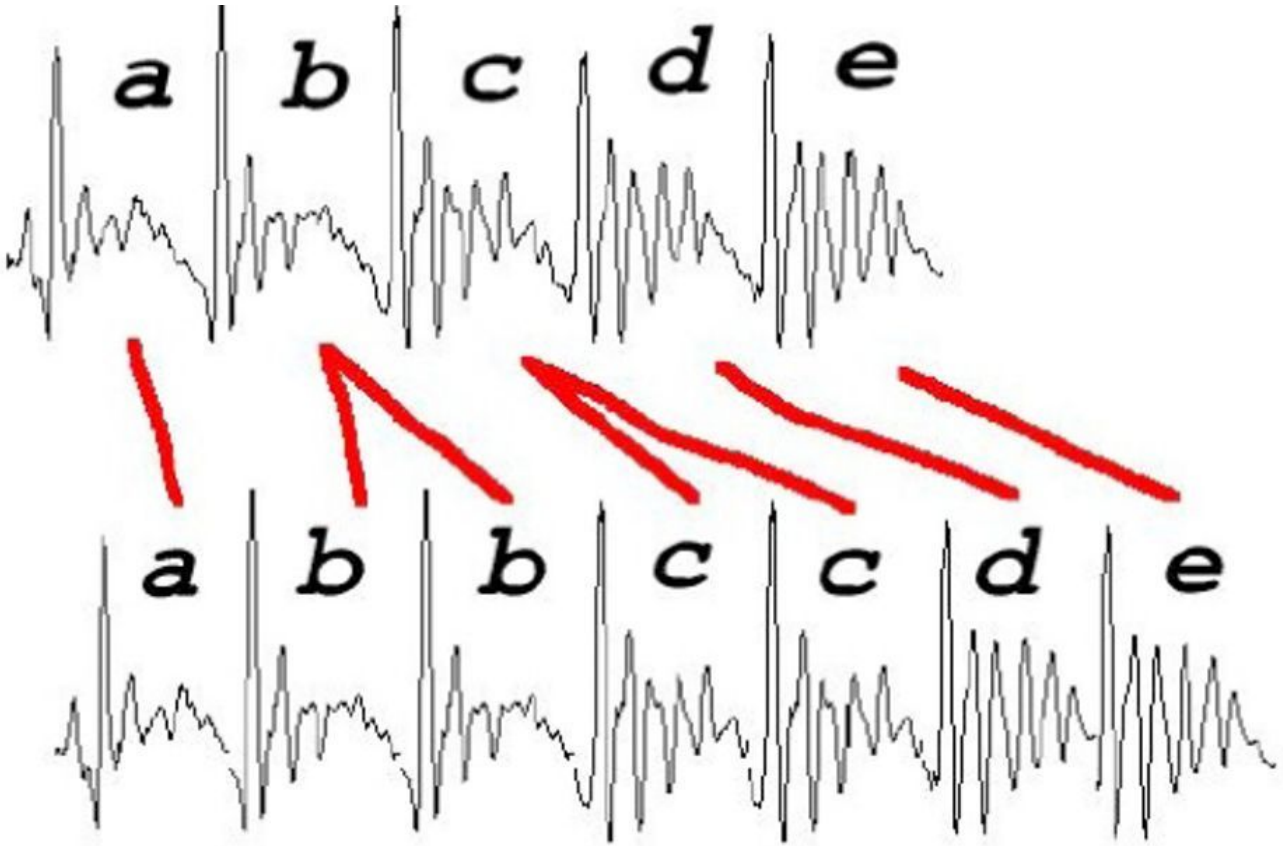


Figure: Richard Sproat

Modifikasi Pitch

- Memindahkan short-term signals untuk saling mendekati atau menjauhi satu dengan lainnya

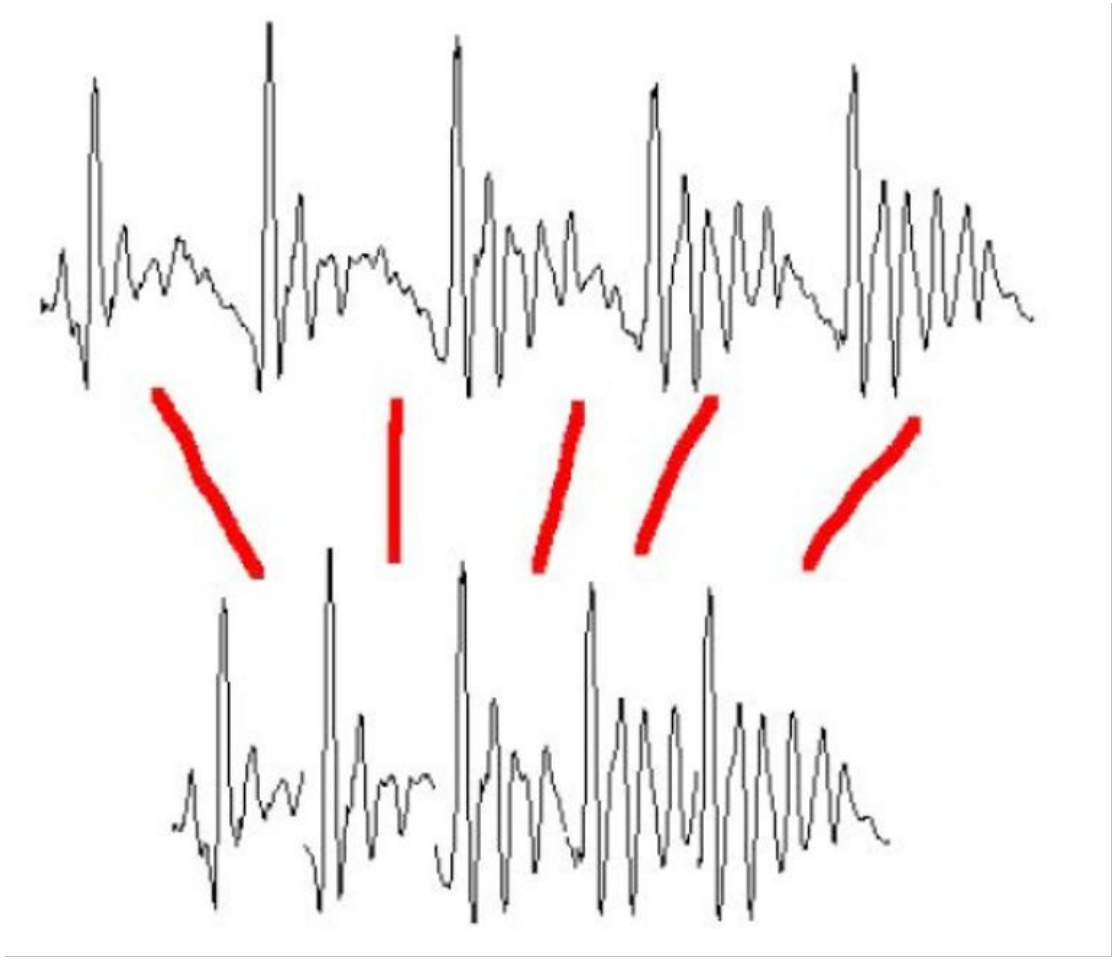


Figure: Richard Sproat

Windowing

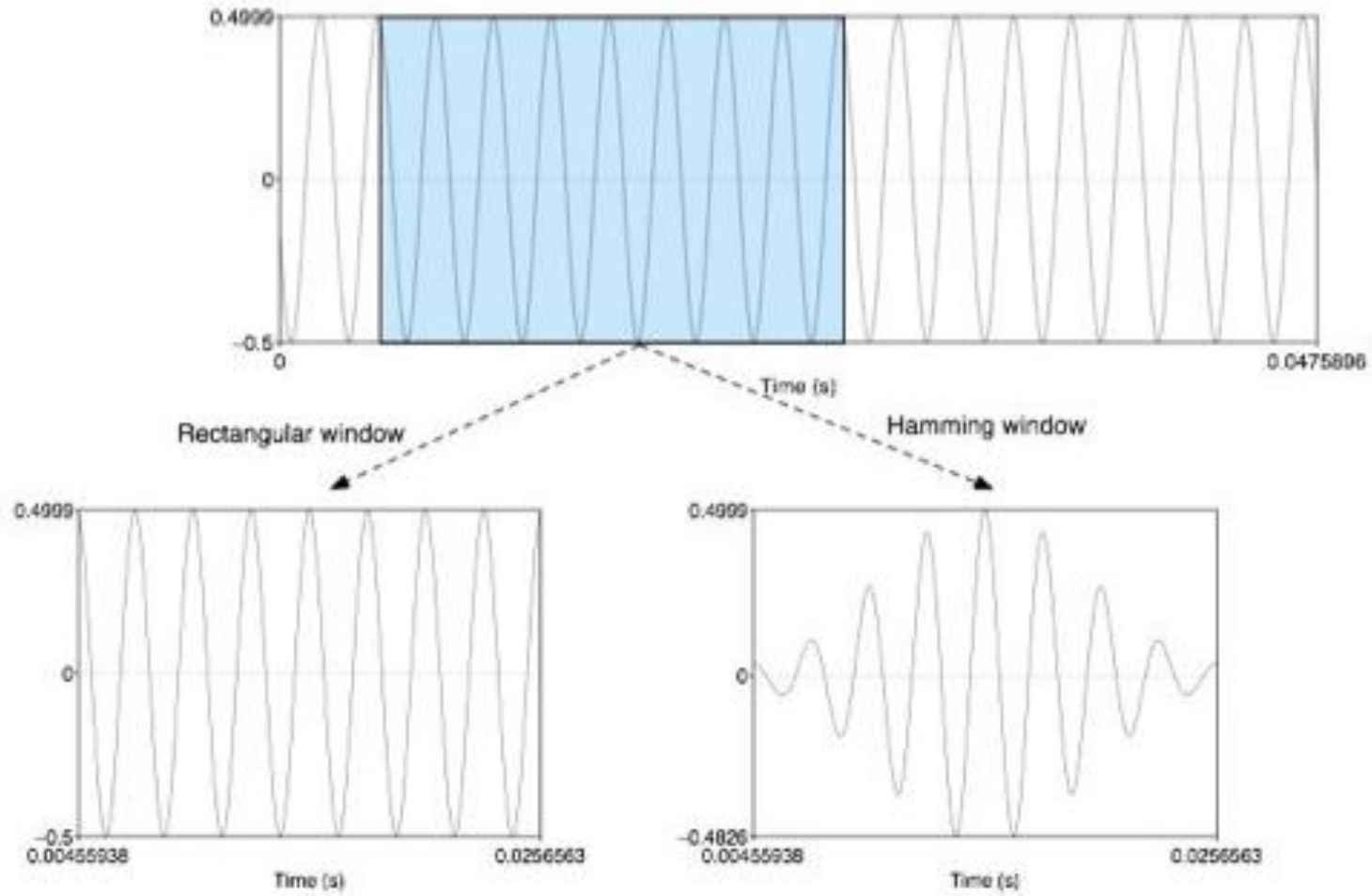
- Kalikan nilai sinyal pada nomor sampel n dengan nilai dari fungsi windowing
- y[n] = w[n]s[n]

$$rectangular \qquad w[n] = \begin{cases} 1 & 0 \le n \le L-1 \\ 0 & \text{otherwise} \end{cases}$$

$$hamming \qquad w[n] = \begin{cases} 0.54 - 0.46\cos(\frac{2\pi n}{L}) & 0 \le n \le L-1 \\ 0 & \text{otherwise} \end{cases}$$

Windowing

- y[n] = w[n]s[n]

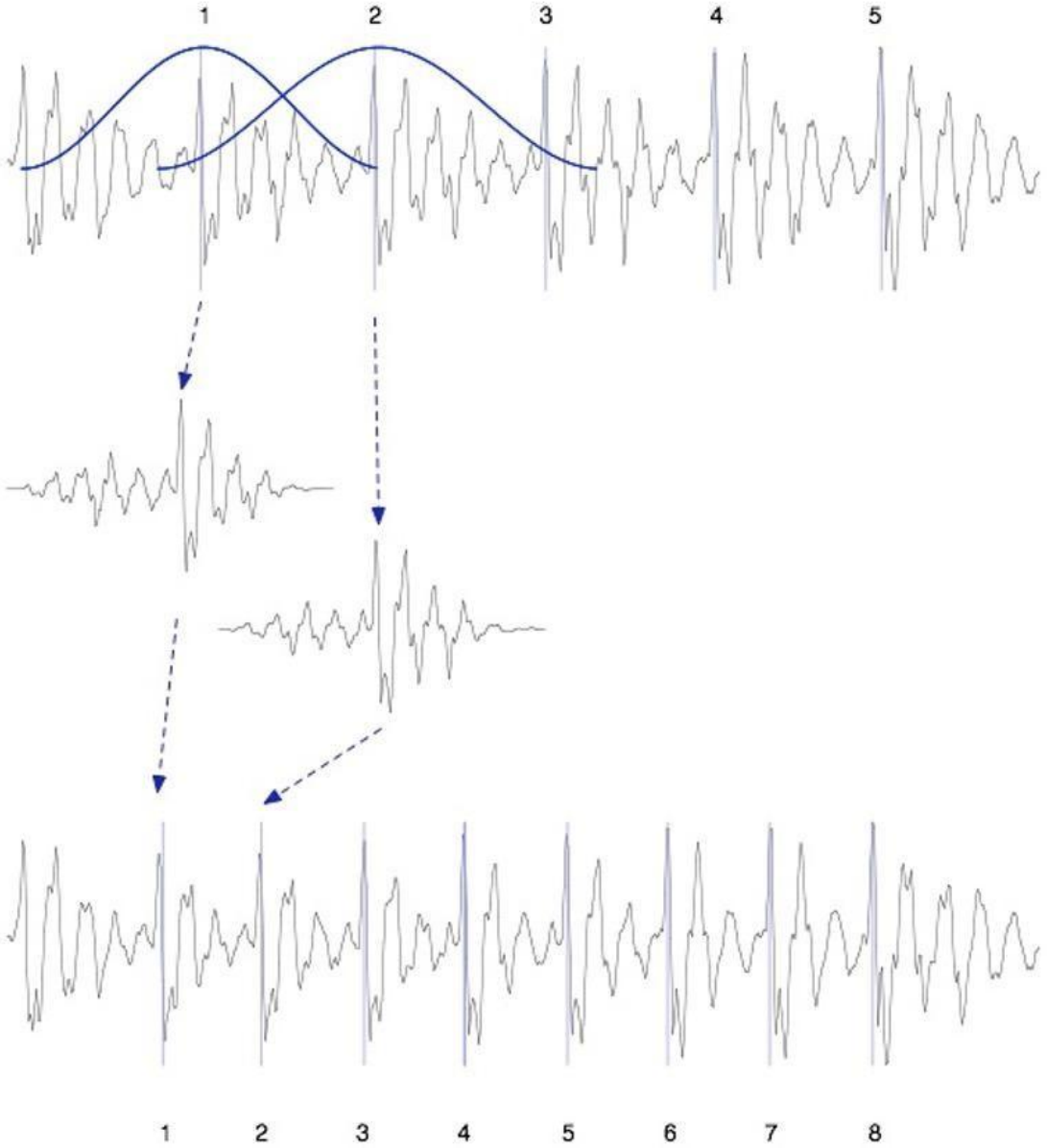


TD-PSOLA ™

Time-**D**omain (Windowed)
Pitch-**S**ynchronous

Over**l**ap-and-**A**dd

- Efficient
- Wide range of Hz
- Join units of any size



Diphone TTS architecture

- Training:
  - Pilih unit (jenis-jenis diphone).
  - Rekam 1 pembicara yang mengucapkan setiap contoh diphone.
  - Tandai batas setiap diphone.
    - Potong setiap diphone dan buat database diphone.
- Sintesis Ucapan:
  - Ambil urutan diphone yang relevan dari database.
  - Gabungkan diphone-diphone, lakukan pemrosesan sinyal ringan di batas-batasnya.
  - Gunakan pemrosesan sinyal untuk mengubah prosodi (F0, energi, durasi) dari urutan diphone.

Ringkasan: Sintesis Diphone

- Teknologi yang sudah matang dan dipahami dengan baik
- Augmentasi:
  - Stress
  - Onset/coda
  - Demi-syllables
- Masalah:
  - Pemrosesan sinyal tetap diperlukan untuk memodifikasi durasi
  - Data sumber masih kurang alami
  - Unit terlalu kecil; tidak dapat menangani efek spesifik kata, dll.

Problems with Diphone Synthesis

- Metode pemrosesan sinyal seperti TD-PSOLA meninggalkan artefak, membuat suara terdengar tidak alami.
- Sintesis diphone hanya menangkap efek lokal.
  - Namun, ada banyak efek global lainnya (struktur suku kata, pola stress, efek di level kata).

Sintesis bentuk gelombang: ikhtisar

- Membangun sistem teks-ke-ucapan
- Sintesis berbasis forman
- Sintesis konkatenatif
  - Sintesis Difone
  - Sintesis seleksi unit
- Sintesis parametrik

Unit Selection Synthesis

Generalisasi dari konsep diphone

- Menggunakan unit yang lebih besar
  - Dari diphone hingga kalimat
- Banyak salinan dari setiap unit
  - 10 jam rekaman ucapan, bukan hanya 1500 diphone (beberapa menit ucapan)
- Sedikit atau tanpa pemrosesan sinyal yang diterapkan pada setiap unit, berbeda dengan sintesis diphone

Unit Selection Synthesis

Data alami menyelesaikan masalah yang ada pada diphone.

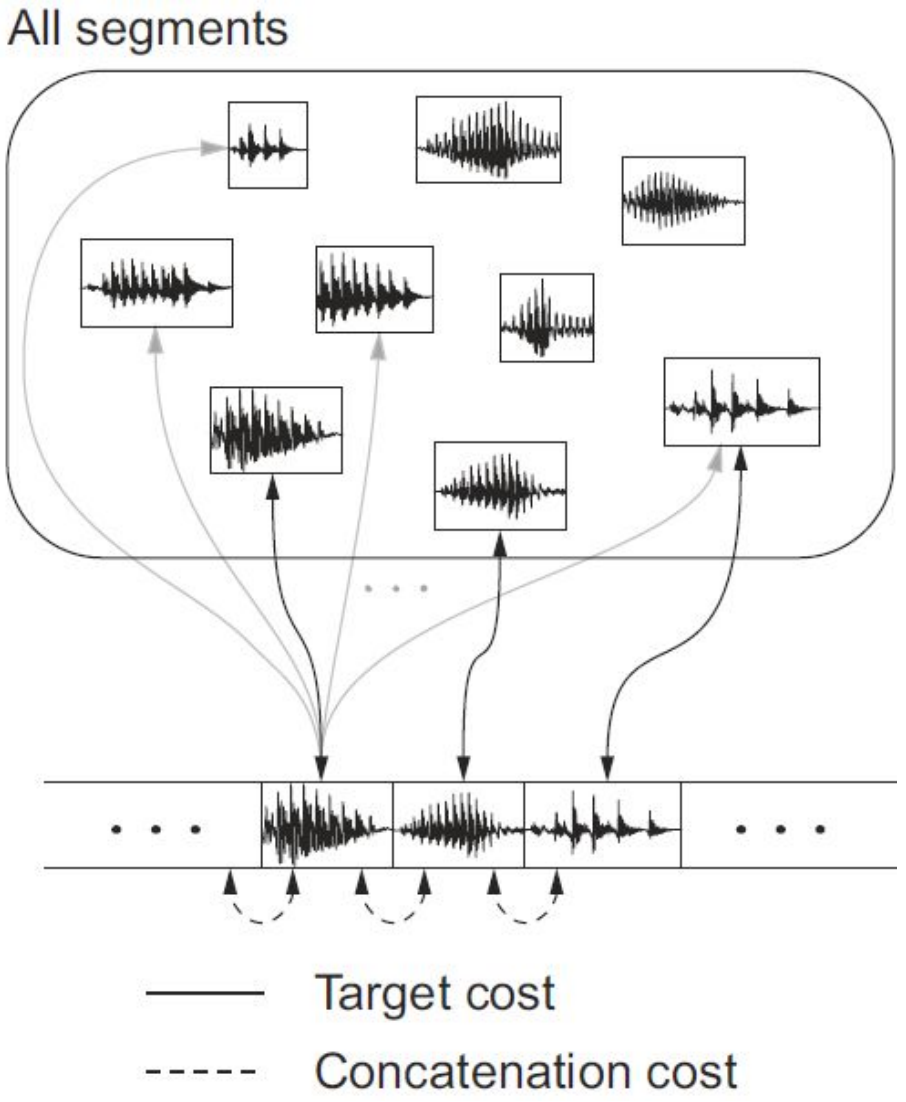
- Database diphone dirancang dengan hati-hati, tetapi:
  - Pembicara mungkin membuat kesalahan.
  - Pembicara mungkin tidak berbicara dalam dialek yang diinginkan.
  - Desain database harus benar.
- Jika otomatis:
  - Dilabeli sesuai dengan apa yang sebenarnya dikatakan oleh pembicara.
  - Koartikulasi, schwa, dan flaps muncul secara alami.
- “There’s no data like more data”
  - Banyak salinan setiap unit memungkinkan pemilihan yang tepat sesuai konteks.
  - Unit yang lebih besar memungkinkan menangkap efek yang lebih luas.

Sintesis konkatenatif

- Langkah-langkah:
  - Buat basis data difon dari satu penutur
  - Pencarian pemilihan unit
  - Penggabungan unit
- Difon = bagian kedua dari satu fonem & bagian pertama dari fonem lainnya
  Misalnya, dalam [daɪfəʊn], difonnya adalah [da], [aɪ], [ɪf], [fə], [əʊ], [ʊn]
  - Dapat dikelola: Inggris 1500, Spanyol 800, Jerman 2500
  - Representasi: forman, LPC, bentuk gelombang
- Perlu banyak rekaman dari satu orang
- Masih canggih untuk beberapa bahasa
- Demo (Festival): https://www.cs.cmu.edu/~awb/festival_demos/general.html

Pencarian pemilihan unit

- Menyintesiskan kalimat baru dengan memilih unit subkata dari basis data ujaran
  - Menghasilkan gabungan hal-hal yang direkam bersama-sama
- Apa arti unit “terbaik”?
  - Mencocokkan target nada, kenyaringan, dll. (spesifikasi s_t) – target cost $T(u_t, s_t)$
  - Mencocokkan unit tetangga – join cost $J(u_t, u_{t+1})$



Pencarian pemilihan unit

- Target cost antara kandidat, u_i, dan unit target t_i:

$$C^{(t)}(t_i, u_i) = \sum_{j=1}^{p} w_j^{(t)} C_j^{(t)}(t_i, u_i)$$

- Join cost antara unit kandidat:

$$C^{(c)}(u_{i-1}, u_i) = \sum_{k=1}^{q} w_k^{(c)} C_k^{(c)}(u_{i-1}, u_i)$$

- Temukan rangkaian unit yang meminimalkan cost keseluruhan:

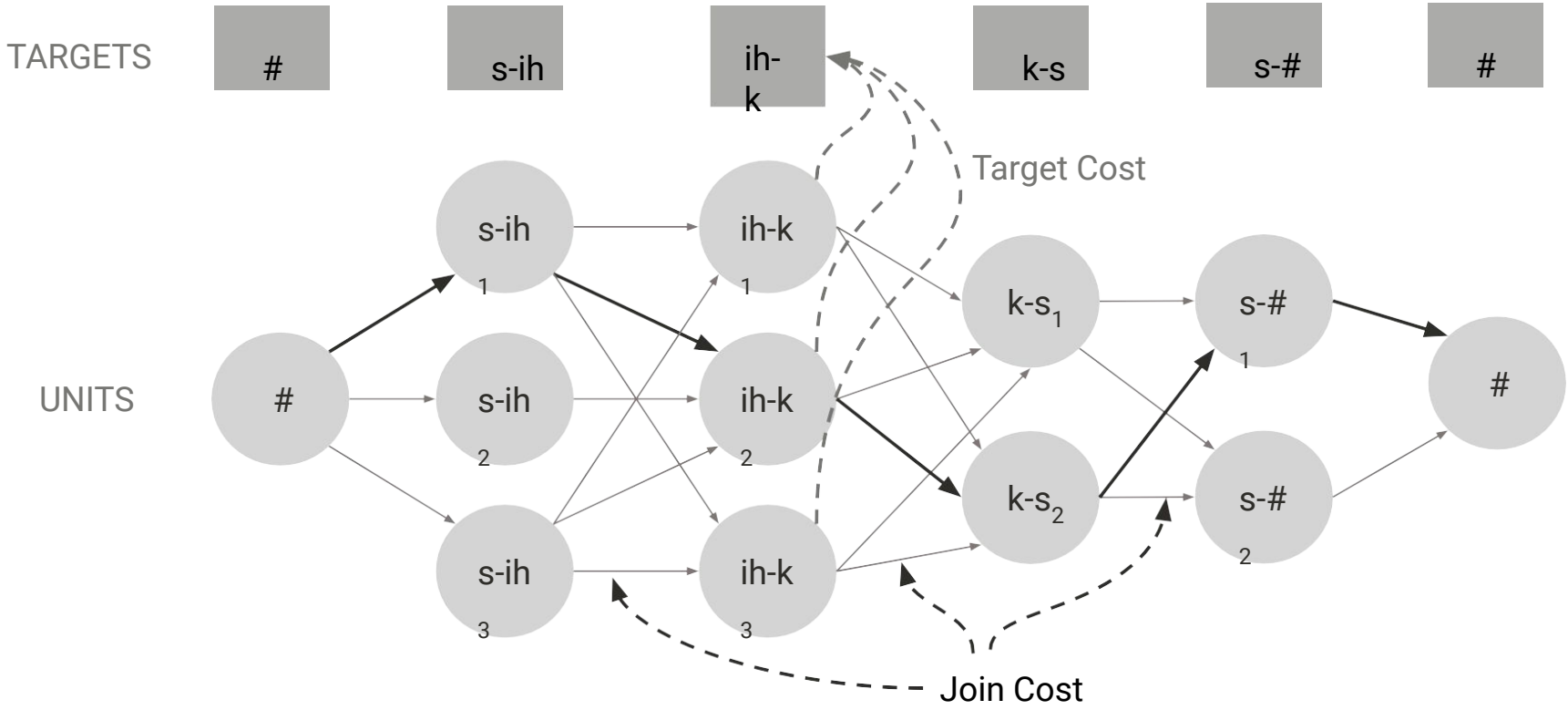
$$\hat{u}_{1:n} = \arg\min_{u_{1:n}} \{C(t_{1:n}, u_{1:n})\}$$

$$C(t_{1:n}, u_{1:n}) = \sum_{i=1}^{n} C^{(t)}(t_i, u_i) + \sum_{i=2}^{n} C^{(c)}(u_{i-1}, u_i)$$

Pencarian pemilihan unit

- Dapat dilakukan dengan pencarian Viterbi
- Setiap sisi dikaitkan dengan total biaya

Contoh: pencarian pemilihan unit untuk kata “six” [s ih k s]



Pelatihan sintesis konkatenatif

- Menggunakan estimasi bobot otomatis [Hunt dan Black, 1996]
  - Menggunakan pencarian grid di berbagai nilai bobot
  - Menggunakan model regresi untuk memprediksi nilai terbaik untuk bobot

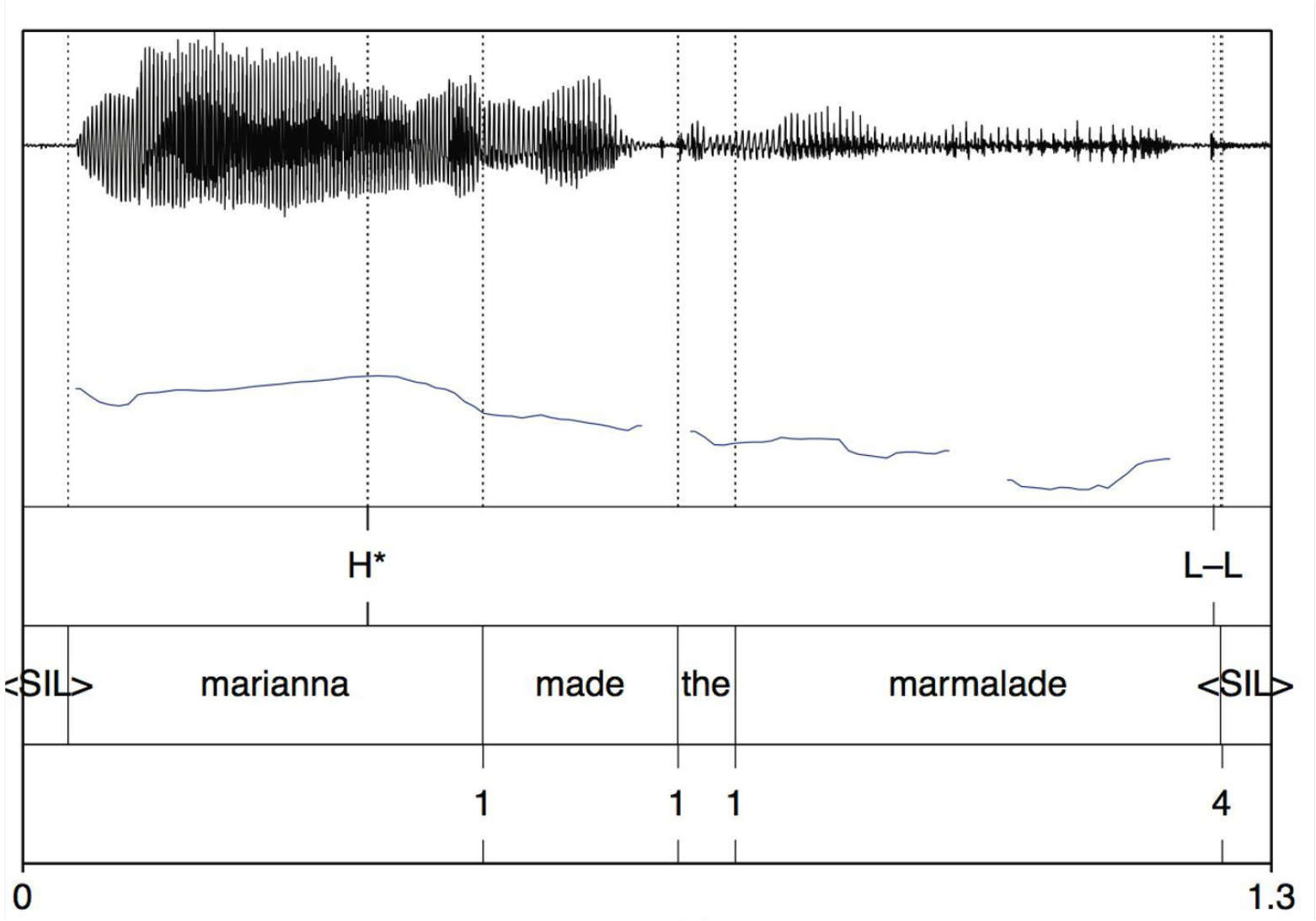
Predicting Intonation in TTS

- **Prominensi/Aksen:** Tentukan kata mana yang memiliki aksen, suku kata mana yang memiliki aksen, serta jenis aksen.
- **Batasan:** Tentukan di mana batas intonasi berada.
- **Durasi:** Tentukan panjang setiap segmen.
- **F0:** Hasilkan kontur F0 berdasarkan elemen-elemen ini.

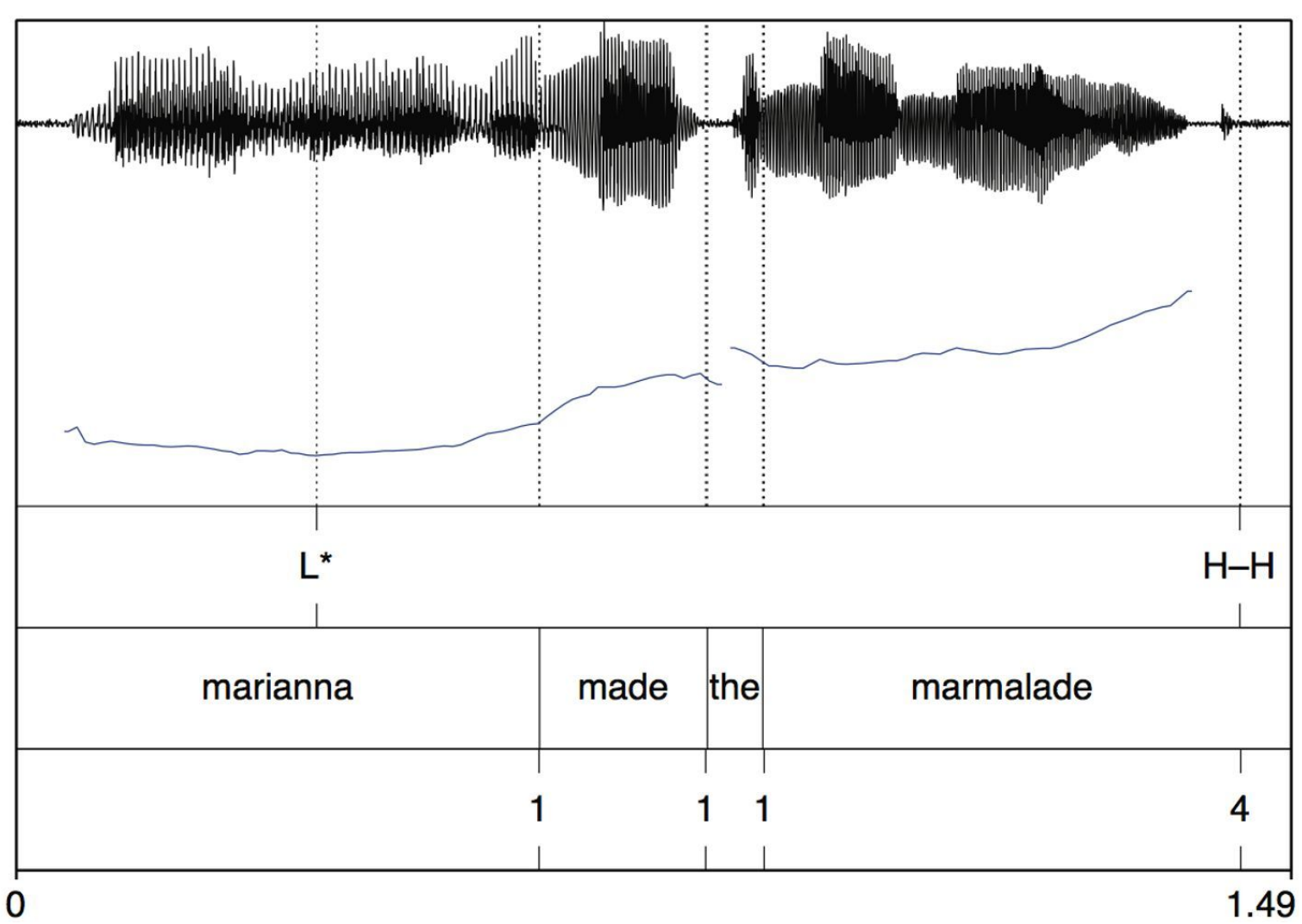
ToBI: Nada dan Indeks Pemisahan

- Nada Aksen Pitch:
  - H*: "peak accent" (aksen puncak)
  - L*: "low accent" (aksen rendah)
  - L+H*: "rising peak accent" (aksen puncak naik, kontras)
  - L*+H: "scooped accent" (aksen cekung)
  - H+!H*: high downstepped (aksen tinggi turun)
- Nada Batas:
  - L-L%: rendah-final (kontur deklaratif bahasa Inggris Amerika)
  - L-H%: kenaikan untuk kelanjutan
  - H-H%: pertanyaan ya-tidak
- Indeks Pemisahan (Break Indices):
  - 0: klitik
  - 1: batas kata
  - 2: jeda pendek
  - 3: frasa intonasi menengah
  - 4: frasa intonasi penuh/batas akhir

ToBI: Tones and Break Indices



ToBI: Tones and Break Indices



ToBI: Tones and Break Indices

- Silverman, K., Beckman, M., Pitrelli, J., Ostendorf, M., Wightman, C., Price, P., Pierrehumbert, J., and Hirschberg, J. (1992). ToBI: a standard for labelling English prosody. In Proceedings of ICSLP92, volume 2, pages 867-870
- Pitrelli, J. F., Beckman, M. E., and Hirschberg, J. (1994). Evaluation of prosodic transcription labeling reliability in the ToBI framework. In ICSLP94, volume 1, pages 123-126
- Pierrehumbert, J., and J. Hirschberg (1990) The meaning of intonation contours in the interpretation of discourse. In P. R. Cohen, J.Morgan, and M. E. Pollack, eds., Plans and Intentions in Communication and Discourse, 271-311. MIT Press.
- Beckman and Elam. Guidelines for ToBI Labelling. Web.

Recap: Joining Units (+F0 + Duration)

- Dalam seleksi unit, seperti pada diphone, unit-unit perlu digabungkan secara pitch-sinkron.
- Pada sintesis diphone, perlu memodifikasi F0 dan durasi.
  - Pada seleksi unit, secara prinsip juga perlu memodifikasi F0 dan durasi unit yang dipilih.
  - Namun, dalam praktiknya, jika database seleksi unit cukup besar (seperti pada sistem komersial), modifikasi prosodik tidak diperlukan karena unit yang dipilih mungkin sudah mendekati prosodi yang diinginkan.

Unit Selection Summary

Keuntungan:

- Kualitas jauh lebih baik daripada diphone.
- Pemilihan prosodi alami terdengar lebih baik.

Kekurangan:

- Kualitas bisa sangat buruk di beberapa bagian.
  - Masalah HCI: campuran antara suara yang sangat baik dan sangat buruk cukup mengganggu.
- Sintesis memerlukan komputasi yang mahal.
  - Tidak dapat menyintesis semua yang diinginkan:
  - Seleksi unit (tidak seperti sintesis diphone) tidak dapat mengubah penekanan.
  - Seleksi unit memberikan hasil yang bagus (tetapi mungkin tidak sepenuhnya benar).

Sintesis bentuk gelombang: ikhtisar

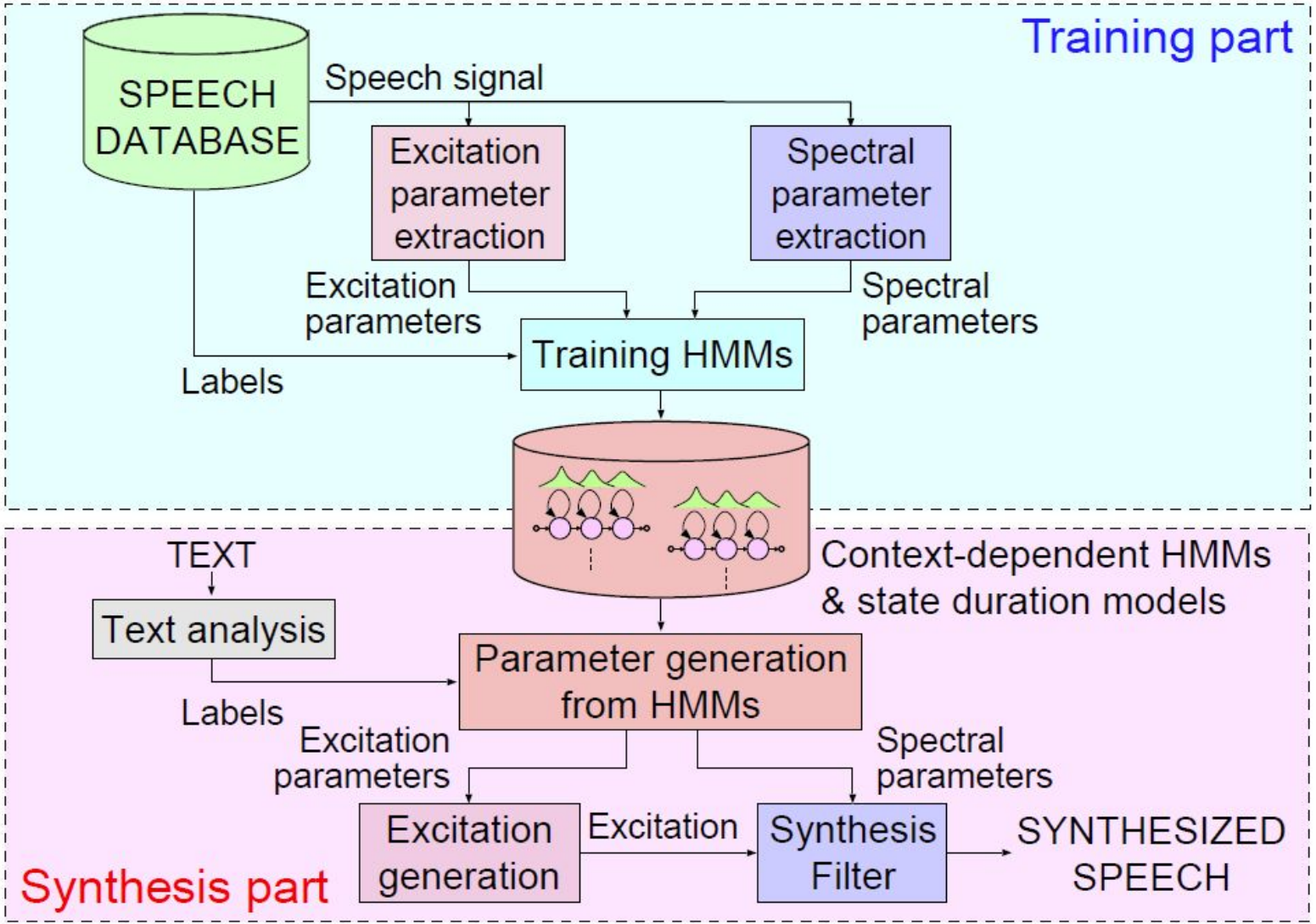
- Sintesis berbasis forman
- Sintesis konkatenatif
- Sintesis parametrik
- Membangun sistem teks-ke-ucapan

Pertanyaan kunci dalam sintesis parametrik

- Parameter apa yang kita prediksi?
  - Biasanya MFCC untuk spektrum, log F0, voicing/eksitasi
- Bagaimana kita menggabungkannya (vocoding)?
  - Parameterisasi yang tepat dan menggabungkannya dengan baik mengurangi efek buzzy robotik
- Bagaimana kita membuat prediksi?
  - Pilihan HMM, pendekatan pembelajaran mesin
  - Kurang penting daripada masalah vocoding/kombinasi
- Untuk representasi input/output yang dipilih untuk TTS, bagaimana bisa mendapatkan label berkualitas tinggi untuk representasi?

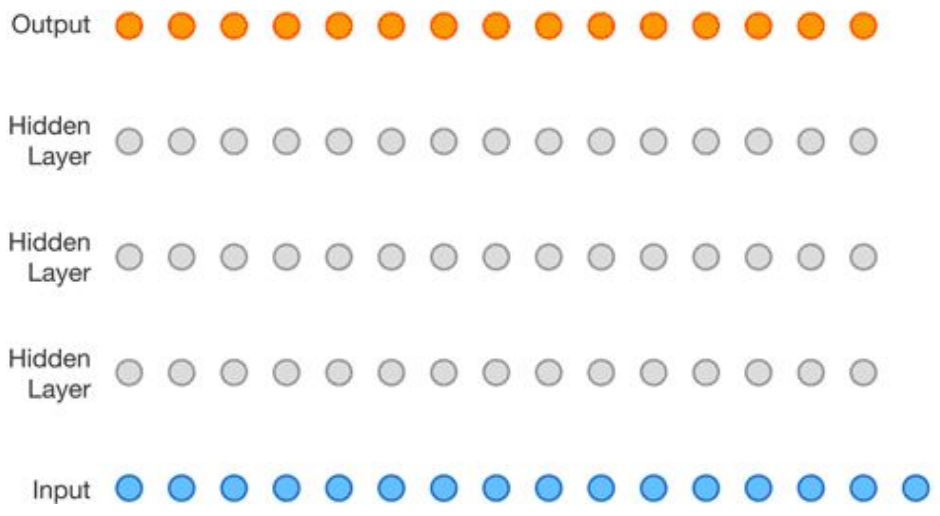
Apa yang dihasilkan sintesis berbasis HMM?

- Kita tidak menggunakan HMM lagi, tetapi sistem ML mana pun memiliki pertanyaan yang sar



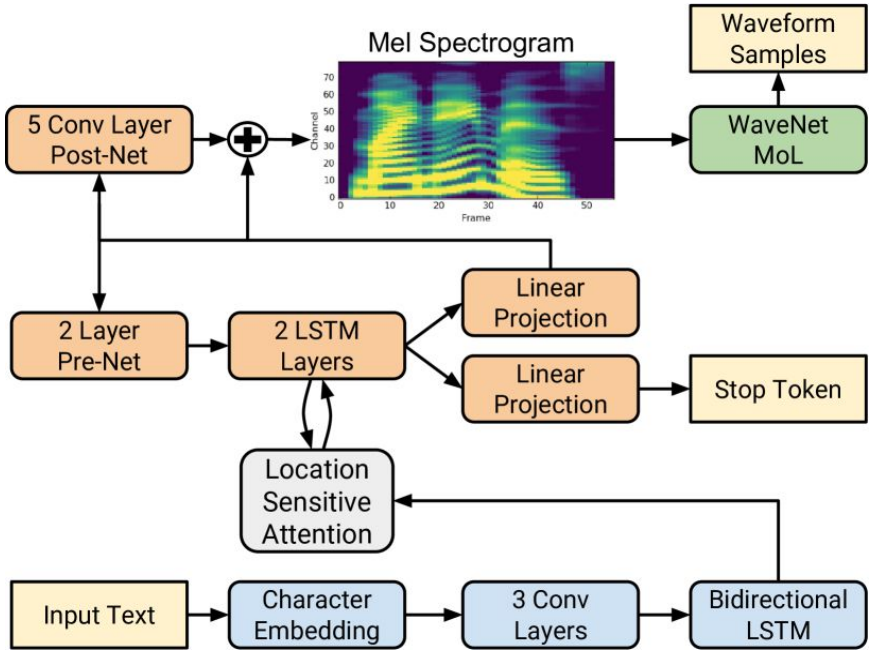
WaveNet

- Menghapus fitur akustik – pembuatan bentuk gelombang langsung
  - Tidak perlu spektrum
- Berdasarkan NN konvolusional
  - 16k langkah/detik → perlu dependensi yang sangat panjang
  - Dilated convolution – melewatkan langkah-langkah
  - Bidang reseptif eksponensial berkenaan dengan # lapisan
  - Dikondisikan pada fitur linguistik
  - Memprediksi gelombang terkuantisasi menggunakan softmax
- Tidak terikat pada bingkai ±stasioner
  - Dapat menghasilkan gelombang yang sangat non-linier
- Sangat alami, penawaran terbaik Google saat ini



Tacotron

- Pendekatan berbeda: menghilangkan fitur linguistik
- Dilatih langsung dari pasangan bentuk gelombang & transkripsi
- Menghasilkan spektogram (pada tingkat bingkai)
  - T1 – linier: konversi Griffin-Lim (memperkirakan fase gelombang yang hilang)
  - T2 – skala mel: perlu sesuatu yang lebih baik, seperti WaveNet, kualitas yang lebih baik
- Berdasarkan model seq2seq dengan attention
  - Diadaptasi – hanya LSTM yang tidak berfungsi dengan baik
  - T2 – encoder: konvolusional + LSTM
  - T2 – decoder: Linear pre-net (mengurangi spektrum sebelumnya)
    - LSTM + attention
    - Stop classification
    - Post-net – konvolusional: menghasilkan spektrum
  - T1: serupa, lebih kompleks (lapisan khusus)



Sintesis bentuk gelombang: ikhtisar

- Sintesis berbasis forman
- Sintesis konkatenatif
- Sintesis parametrik
- Membangun sistem teks-ke-ucapan