Departamento de Computación, FCEyN, UBA

Procesamiento del Habla

Agustín Gravano

1er Cuatrimestre 2017

Prosodia

Entonación, ritmo y más…

Ejemplo de síntesis

- caperucita-roja.ogg
  - Sintetizado con Bluemix de IBM, voz es-LA, demo: https://text-to-speech-demo.mybluemix.net/
  - Texto del cuento “Caperucita Roja” tomado de: http://ciudadseva.com/texto/caperucita-roja/

- ¿Qué problemas tiene esta voz artificial?

Prosodia

- Mucha información más allá de las palabras.

- ¿Cómo varía el habla?
  - *"paró de llover" [. ? !]*
  - *"no dije a la izquierda, dije a la **derecha**"*
  - *"por un lado…", "entonces…", "…, pero…"*
  - *"no cantes, Victoria"*
  - *"como acá"*
  - *"Ayer se me rompió el auto." vs.*
    "Vengo del mecánico. Ayer se me rompió el auto."

¿Qué es la prosodia?

"Uso de características suprasegmentales para comunicar significados pragmáticos al nivel de la oración." (Ladd, 1996)

- Características suprasegmentales:
  - Abarcan varios fonos.
  - F0, intensidad, duración, calidad de la voz.
- Significados pragmáticos al nivel de la oración:
  - Estructura o función discursiva.
  - Prominencia de una palabra o una frase.
  - Significado afectivo o emocional.

Dimensiones de la Prosodia

- Intensidad.
  - Fuerte/suave. Amplitud, RMS → Decibeles.
- Nivel tonal.
  - Aguda/grave. Frecuencia fundamental → Herz.
- Tasa del habla.
  - Rápida/lenta.
- Calidad de la voz.
  - Susurro, voz tensa, voz rasposa, etc.

Tasa del Habla

- Unidad segmental sobre unidad de tiempo. Ej:
  - Palabras/minuto.
  - Sílabas/segundo.
  - Fonos/segundo.
- ¿Qué necesitamos para medir la tasa del habla?
  - Depende del nivel de detalle que necesitemos.
  - Si tenemos una transcripción alineada (manual o automática), podemos estimar síl/seg y fon/seg.
  - Para síl/seg, detectar núcleos de las sílabas.
    - Ej: syllables.praat, bajado de: https://sites.google.com/site/speechrate/Home/praat-script-syllable-nuclei-v2

Calidad de la voz

- Variables acústicas relacionadas:
  - *Jitter*, *shimmer*: perturbaciones en la periodicidad de la señal (frecuencia y amplitud, respectivamente).
  - Relación ruido-armónico (NHR y HNR): relación entre componentes periódicas (armónicos) y no periódicas (ruido).
- Correlatos perceptuales:
  - Voz clara, limpia vs. rasposa, ronca, crujiente.
  - Voz relajada vs. Tensa.
- Usada para estudiar patologías del habla.

Variación prosódica

- Hasta ahora vimos cómo medir las variaciones en la prosodia.
- ¿Cómo podemos representar esas variaciones?
  - Analogía con espectro → fonos.
- Cómo representar diferencias en:
  - Intensidad
  - Velocidad
  - Prominencia
  - Frases prosódicas
  - Contorno entonacional
  - Calidad del habla

Variación prosódica

Ejercicios

- comoaca.wav y nocantesvictoria.wav
  - Variantes de “como acá” y de “no cantes victoria”.
  - Estudiar pitch track, intensidad, duración de cada sílaba.
  - Intentar caracterizar la prosodia de:
    - comoaca.wav: pregunta vs. afirmación
    - comoaca.wav: significado (*igual que acá* vs. *almuerzo acá*)
    - nocantesvictoria.wav: la pausa al final de “cantes”

Variación prosódica

- Intento temprano: Joshua Steele 1775
  - JoshuaSteel.pdf
- Steel introdujo preguntas importantes:
  - ¿Qué aspectos del habla queremos capturar?
    O sea, ¿cuáles son relevantes?
  - ¿Atributos continuos o categóricos?
  - Si son categóricos, ¿cuáles son las clases?

Variación prosódica

- La variación prosódica abarca:
  - Niveles de prominencia → acentos tonales.
  - Estructura de frases prosódicas.

- Eventos prosódicos marcados por:
  - Prolongaciones segmentales.
  - Cambios en F0, intensidad y calidad de la voz.
  - Límites de frases → muchas veces seguidos de pausas, pero no siempre.

¿Para qué modelar la prosodia?

- Sistemas de síntesis del habla (TTS):
  - Expresar sin ambigüedad el mensaje deseado.
    - Ej: *María no renunció[,] por el sueldo.*
  - Lograr mayor naturalidad.
    - Digresión: ¿Los robots deberían sonar humanos?
- Sistemas de reconocimiento del habla (ASR):
  - Comprender el mensaje expresado por el hablante.
  - Tareas más específicas:
    - Segmentar en temas.
    - Encontrar los puntos más relevantes.
- Sistemas de diálogo hablado (SDS):
  - Manejar bien los protocolos para ceder y tomar la palabra.

Modelos de variación tonal

- Dos tipos de modelos de variación tonal:
  - Modelo Lineal, o de Secuencia de Tonos
    - Secuencia de eventos discretos de un léxico entonacional.
  - Modelo Superposicional:
    - Jerarquía de componentes fonológicas.

Modelo de Secuencia de Tonos

- Objetivos tonales
  - Acentos tonales.
  - Tonos de final de frase.

- Sistema ToBI (***To****nes and* ***B****reaks* ***I****ndices)*

- *Ejemplos:*
  - made1.{wav,TextGrid}
  - oregano.{wav,TextGrid}

Modelo de Secuencia de Tonos

- Unidad básica de descripción:
  - Frase entonacional.
  - Delimitada por pausas y/o por una prolongación en el final de la frase (acento de final de frase).
- Cada sílaba puede tener acento léxico y/o acento tonal.
  - Todas las palabras tienen acento léxico.
  - Sólo las palabras salientes en la frase tienen acento tonal.
  - Ejemplo: *"siempre"* lleva acento léxico en la primera sílaba; el acento tonal es opcional.
    *"voy **siem**pre" vs.*
    *"**voy** siempre y **cuan**do no **llue**va"*
- Los acentos tonales se marcan mediante **uno o más** de estos:
  - {pico de f0, pico de intensidad, prolongación de la sílaba}.

Sistema ToBI

- Estándar desarrollado entre 1991 y 1994 por expertos del área.

- Objetivos:
  - Armar un sistema de anotación para la prosodia del inglés NA que fuese robusto y confiable.
  - Promover la disponibilidad de cuerpos de datos anotados prosódicamente siguiendo un estándar, que pudieran ser compartidos por la comunidad.

Sistema ToBI

- Existen versiones de ToBI para el japonés, alemán, italiano, español, y otras variantes del inglés.
- Una transcripción de ToBI requiere:
  - Grabación de habla.
  - Contorno de f0 (*pitch track*).
  - 4 capas (*tiers*) de ToBI:
    - capa ortográfica: palabras
    - capa de junturas entre palabras (*break-index tier*)
    - capa tonal: acentos tonales, tonos de final de frase
    - capa miscelánea: disfluencias, toses, risas, etc.

ToBI: Acentos tonales

- ¿Qué palabras son producidas en forma prominente? ¿y cómo?
- Tipos de acentos tonales:
  - H* simple high (declarativa: "*brown*" en brown.wav)
  - L* simple low (pregunta sí/no: "*mariana*" *en* money1.wav)
  - L*+H late rise (incerteza/incredulidad: "*stein*" en stein1.wav)
  - L+H* early rise to stress (contraste: "*mariana*" en noone.wav)
  - H+!H* fall onto stress (supuesta familiaridad: "*theresa*" en theresa2.wav)

ToBI: Junturas entre palabras

- Nivel de juntura entre palabras:
  - 0: sin límite entre dos palabras. Ej: *la Argentina*.
  - 1: límite normal entre dos palabras.
  - 2: fuerte separación pero sin marca tonal.
  - 3: límite de frase intermedia.
  - 4: límite de frase entonacional.

ToBI: Tonos de final de frase

- Frases entonacionales:
  - L-L% falling (declarativa: *"one"* en brown.wav)
  - L-H% low rising (continuation rise: *"man"* en stein1.wav)
  - H-L% plateau ("school", "people" en school.wav)
  - H-H% high rising (pregunta sí/no: "alley" en manitowoc.wav)

- Frases intermedias: L-, H-.

Sistema ToBI

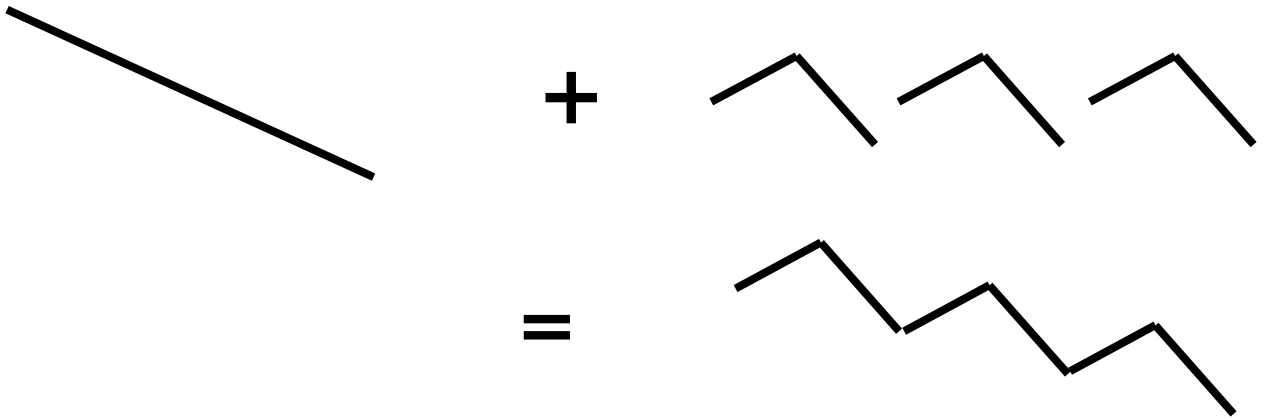
- Material de entrenamiento online:
  - http://anita.simmons.edu/~tobi/index.html
- Evaluación
  - Buena concordancia entre anotadores:
    (Silverman et al. '92; Pitrelli et al '94)
    - 88% acuerdo en presencia/ausencia de tono
    - 81% acuerdo en categoría tonal
    - 91% acuerdo en indices de juntura

Anotación automática de ToBI

- Ejemplo:
  - Escudero-Mancebo, D., González-F., C., Vivaracho-P., C., & Cardeñoso-P., V., "A fuzzy classifier to deal with similarity between labels on automatic prosodic labeling". Computer Speech & Language, 28(1):326-341, 2014.
- Tema abierto de investigación.
- Sistemas algo imprecisos todavía, pero **consistentes**.
  - Aportan información prosódica muy útil para otras tareas de procesamiento automático del habla.

Modelo superposicional de Fujisaki

- Usado principalmente para síntesis del habla.
- Modela el patrón de F0 con una superposición lineal de dos componentes: de frase y de acentos.
  - La frase tiene una forma básica (ej: descendente).
  - Cada acento tiene su propia forma parametrizable.



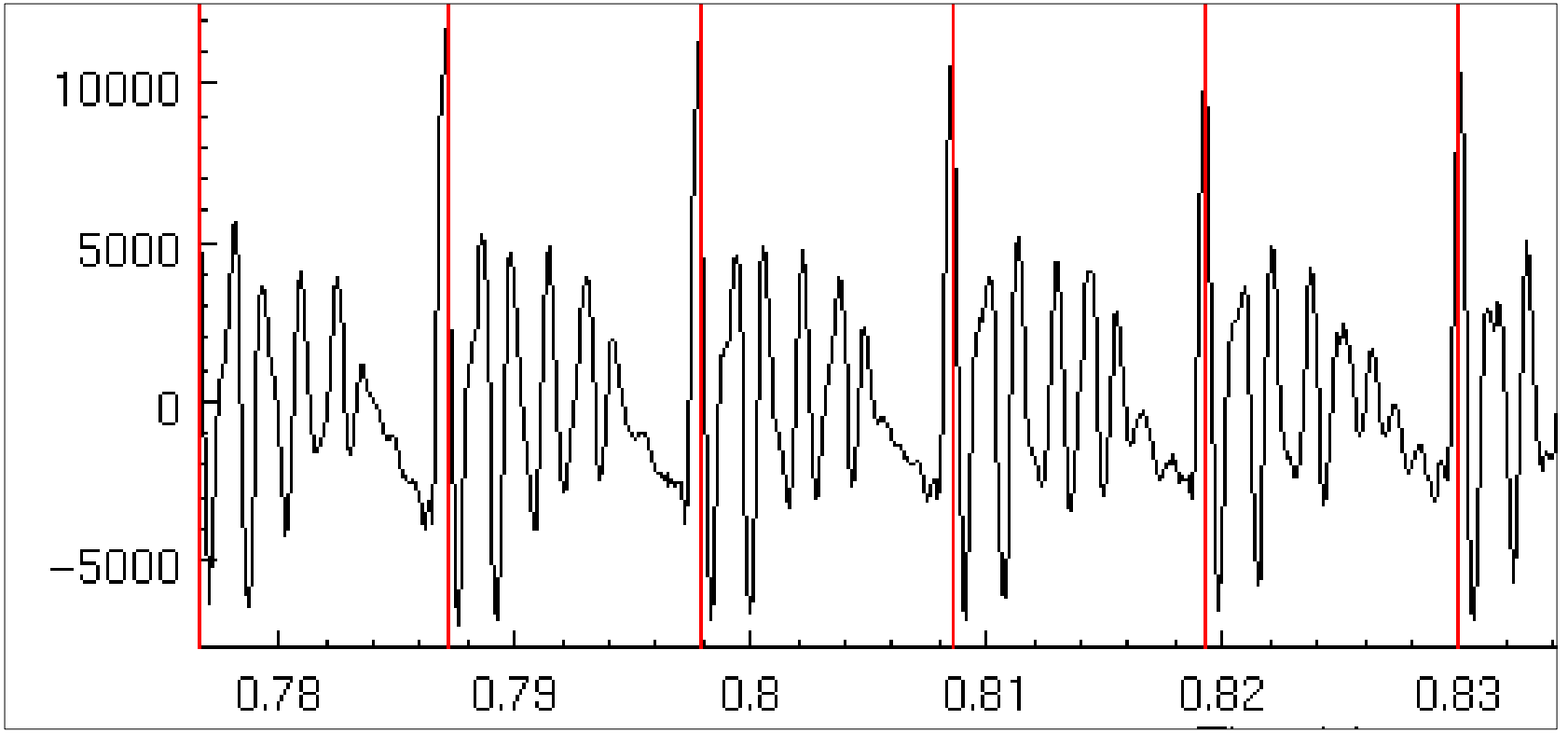
Modelo superposicional de Fujisaki

- Parámetros a determinar para cada frase:
  - Frecuencia base (piso de F0 del hablante).
  - Cantidad de comandos de frase y de acentos.
  - Duración y amplitud de cada comando de frase y de acento.
- En un corpus, entrenamos reglas para derivar estos parámetros a partir de un texto.
- Luego, dado un texto a sintetizar, se derivan esos parámetros para estimar el contorno de F0 a usar.
- Desventajas:
  - No modela los diferentes tipos de acentos, ni las variaciones en finales de frase.
  - Muy específico para síntesis; no adecuado para estudiar la variación prosódica en general.

Modificación de la prosodia

- En síntesis concatenativa, todas las unidades tienen las mismas variables prosódicas (f0, int, dur).
- La prosodia deseada se consigue con proc. de señales.
- La intensidad se puede modificar fácilmente.
- ¿Cómo modificar tono y duración?
  - Aumentar la duración de una señal disminuye el tono.
- TD-PSOLA:
  - *Time-Domain Pitch-Synchronous Overlap-and-Add*
  - Identificar ciclos básicos de la señal.
  - Para cambiar la duración: duplicar/borrar ciclos.
  - Para cambiar el tono: juntar o separar ciclos.

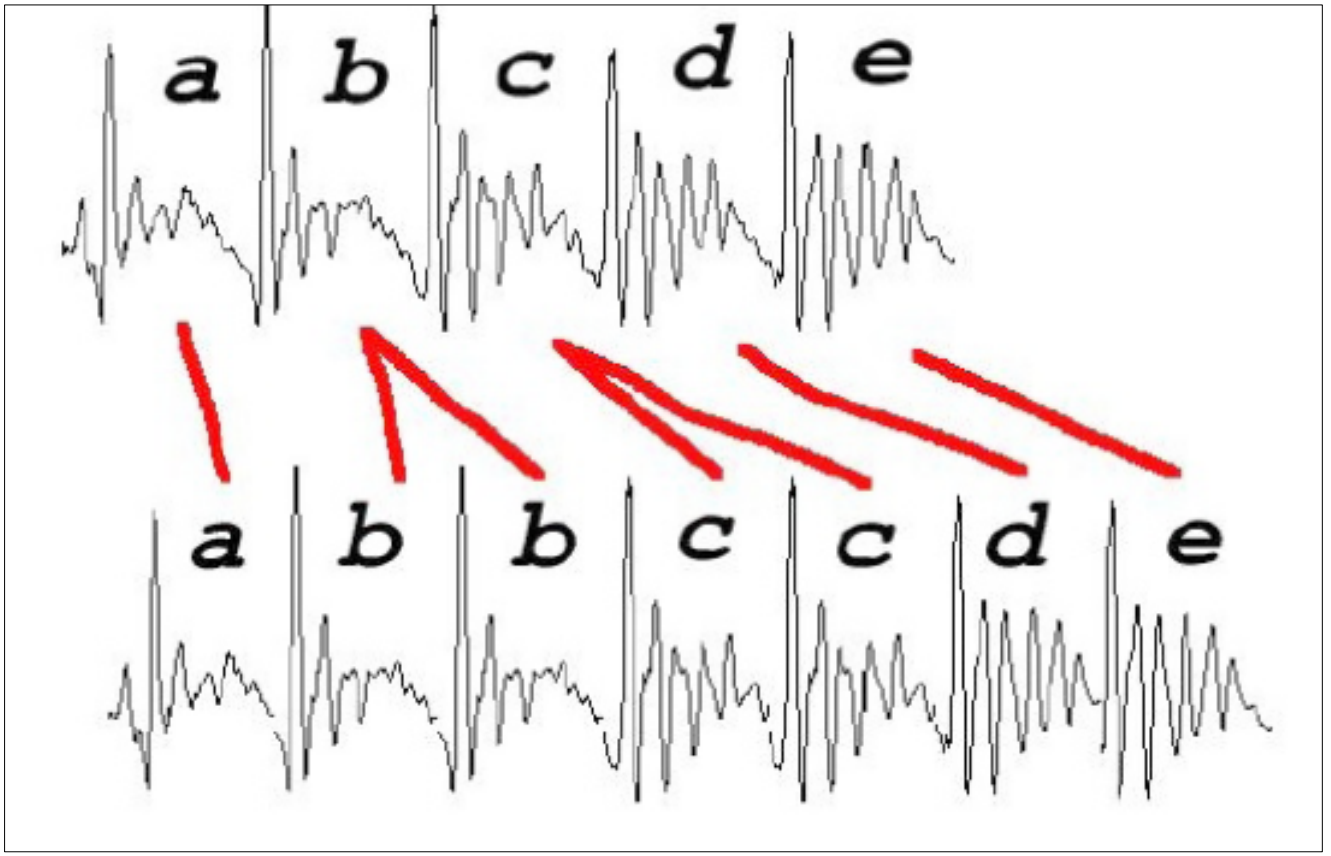
TD-PSOLA: Cómo identificar ciclos básicos



- *Pitch marking*
  - Electroglotógrafo (EGG) durante la grabación.
  - Automáticamente: algoritmos aproximados (ej: autocorrelación).

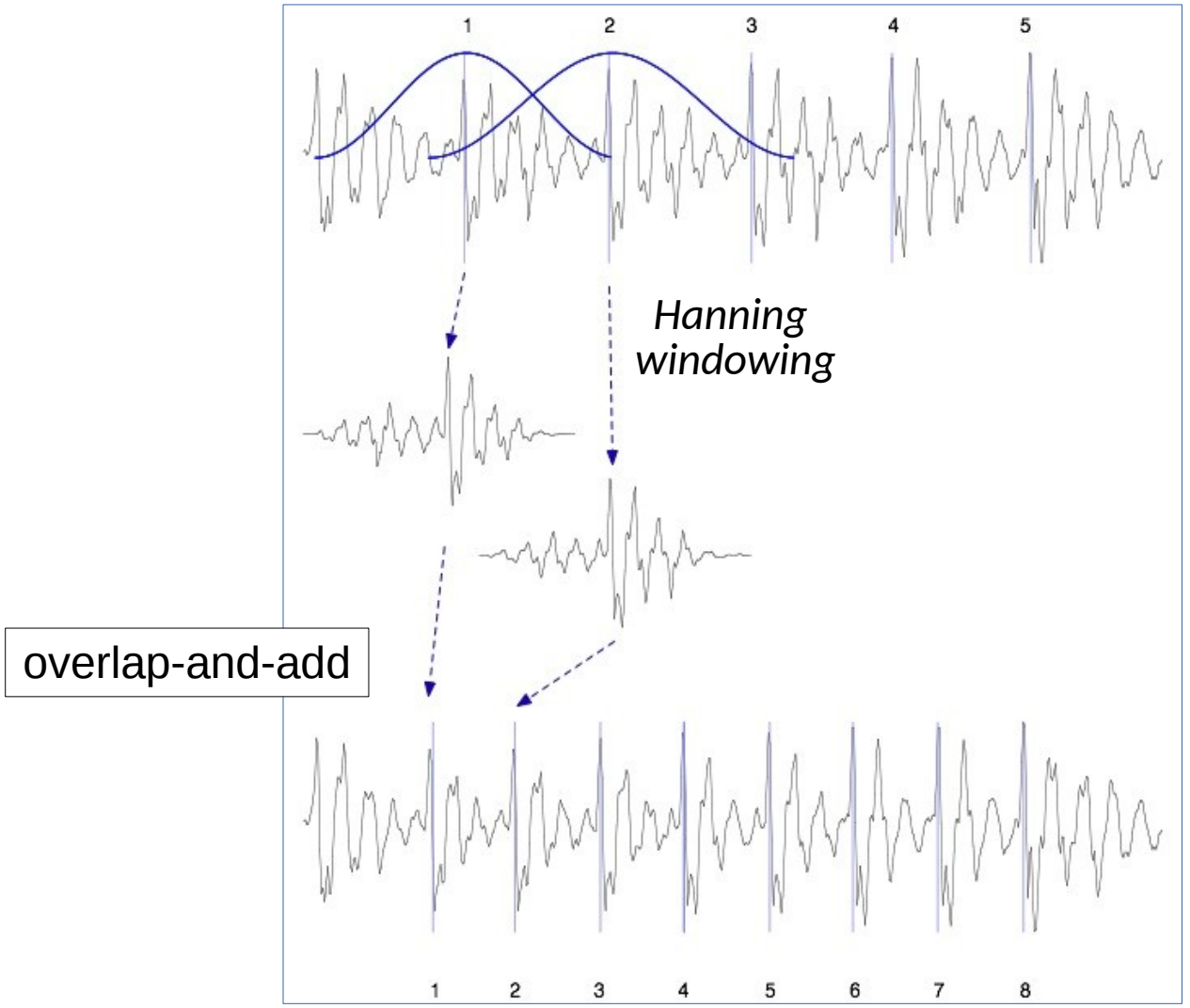
TD-PSOLA: Cómo modificar la duración

- Duplicar ciclos para alargar la señal.
- Eliminar ciclos para acortar la señal.



TD-PSOLA: Cómo modificar el tono

- Juntar/separar ciclos para aumentar/disminuir la frecuencia.
- Agregar ciclos cuando sea necesario para mantener la duración.



Ejercicio: Manipulación de prosodia

- Seleccionar un objeto de tipo 'Sound'
- *Manipulate → To manipulation…* (Usar rango tonal 100-400Hz.)
- *View & Edit.*
- *Pitch → Stylize pitch (2st)*
- Para modificar el tono, en la capa *"Pitch manip"*:
  - Arrastrar los puntos. Agregar puntos con: *click*, *CTRL+T.*

- Para modificar la duración, en la *"Duration manip"*:
  - Arrastrar los puntos. Agregar puntos con: *click*, *CTRL+D.*

- **Ejercicio 1:** Abrir cena.wav. Modificar su prosodia para que suene a una pregunta (mejor aún si suena indignada).
- **Ejercicio 2:** Abrir lamparita.wav. Modificar su prosodia para que diga "al**can**za" en lugar de "alcan**zá**".

Resumen

- Prosodia: Características suprasegmentales del habla para comunicar significados pragmáticos.
- Eventos prosódicos:
  - Prominencia (acentos tonales) y estructura de frases.
  - Marcados con cambios en:
    - duración, f0, intensidad, calidad de voz.
- Modelo de secuencia de tonos: ToBI
- Modelo superposicional de Fujisaki
- Modificación de la prosodia. TD-PSOLA.