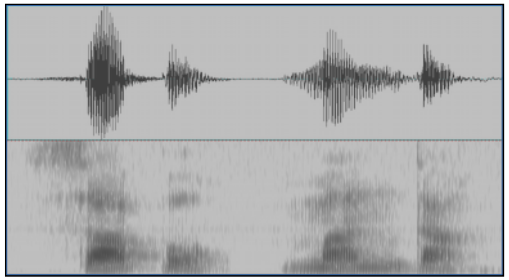
Departamento de Computación, FCEyN, UBA

Procesamiento del Habla

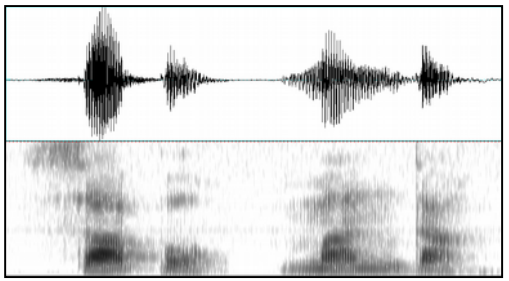
Agustín Gravano

1er Cuatrimestre 2017

Reconocimiento del habla



Preprocesamiento



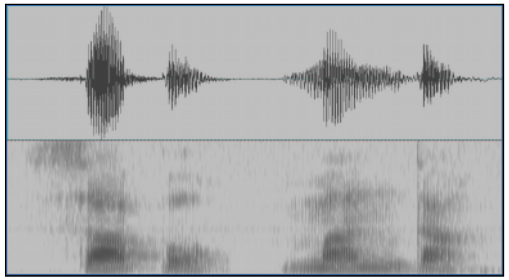
Reconocimiento

“avenida rivadavia veinte doce”

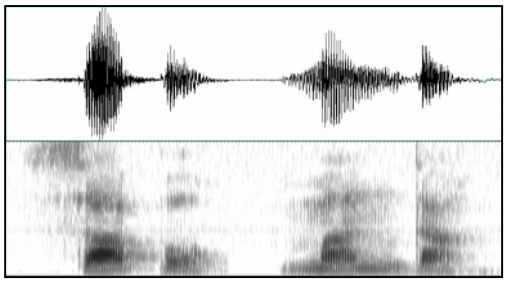
Posprocesamiento

“Av. Rivadavia 2012.”

Reconocimiento del habla



Preprocesamiento



Re[illegible]miento

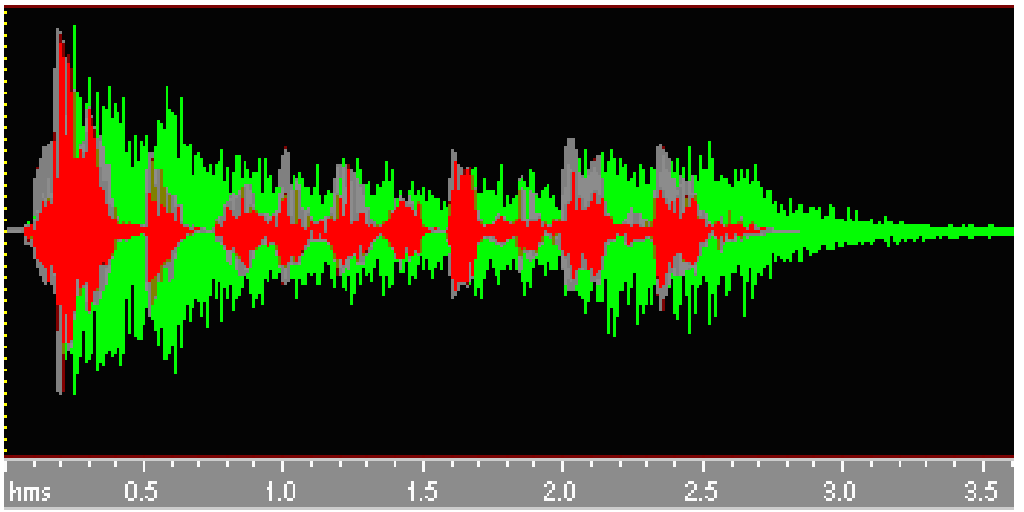
“avenida rivadavia veinte doce”

Posprocesamiento

“Av. Rivadavia 2012.”

Preprocesamiento

- Problemas de la entrada de ASR:
  - Ruidos de fondo.
  - Múltiples hablantes:
    - *Cocktail party problem*.
    - Conversación con varios participantes.
  - Reverberación:



reverberacion.wav

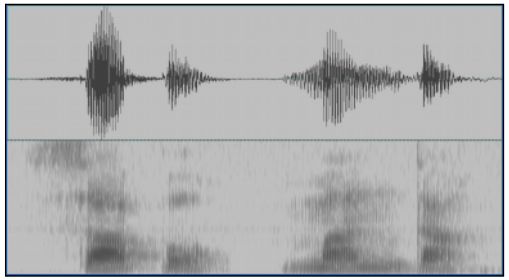
Imagen y audio tomados de http://www.mcsquared.com/y-reverb.htm.

Preprocesamiento

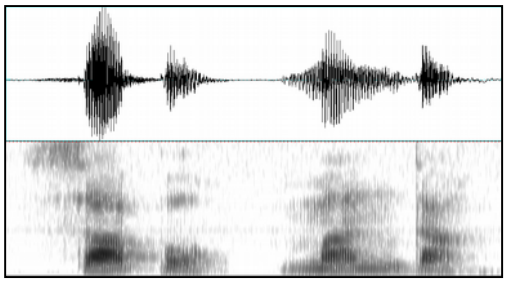
- Soluciones:
  - Técnicas de procesamiento de señales.
  - Filtros acústicos.
  - Filtrado vs. modelado (ej: para reverberación).
  - Separación de las fuentes de audio.
    - Con 1, 2 o múltiples micrófonos.
    - Ejemplo (2 mics): separacion-{1,2}.wav

      Tomado de http://cnl.salk.edu/~tewon/Blind/blind_audio.html.

Reconocimiento del habla



Prep[illegible]amiento



Re[illegible]miento

“avenida rivadavia veinte doce”

Posprocesamiento

“Av. Rivadavia 2012.”

Posprocesamiento

- Transcripción sin errores de parte de un diálogo:

  sí este eh soy socio del de un gimnasio de acá a la vuelta sobre juan be justo este vo- en realidad trato de ir tres días a la semana aunque no siempre puedo qué tipo de ejercicios hacés en el gimnasio y hago de todo un poco cuánto te cobran mil trescientos pesos por mes epa

Posprocesamiento

A: *Sí, soy socio de un gimnasio de acá a la vuelta, sobre Juan B. Justo. Trato de ir tres días a la semana, aunque no siempre puedo.*

B: *¿Qué tipo de ejercicios hacés en el gimnasio?*

A: *Y… hago de todo un poco.*

B: *¿Cuánto te cobran?*

A: *$1300 por mes.*

B: *¡Epa!*

Posprocesamiento

sí este eh soy socio del de un gimnasio de acá a la vuelta | sobre juan be justo | este vo- en realidad trato de ir tres días a la semana aunque no siempre puedo || ¿qué tipo de ejercicios hacés en el gimnasio? || y hago de todo un poco || ¿cuánto te cobran? || mil trescientos pesos por mes || ¡epa!

- Segmentación de hablantes y de oraciones, disfluencias, mayúsculas, puntuación, números, etc.

Segmentación de hablantes

- Noticias de radio, reuniones, diálogos telefónicos.
- Primer paso: Diarización de hablantes.
  - Proceso de particionar el audio en segmentos homogéneos según la identidad del hablante.
  - Habla → [segmento 1] [segmento 2] [segmento 3] …
  - El output **no** tiene información de la identidad de los hablantes.

Segmentación de hablantes

sí este eh soy socio del de un gimnasio de acá a la vuelta

sobre juan be justo este vo- en realidad trato de ir tres días a la semana aunque no siempre puedo

qué tipo de ejercicios hacés en el gimnasio

y hago de todo un poco

cuánto te cobran

mil trescientos pesos por mes

epa

Segmentación de hablantes

sí este eh soy socio del de un gimnasio de acá a la vuelta

sobre juan be justo este vo- en realidad trato de ir tres días a la semana aunque no siempre puedo

qué tipo de ejercicios hacés en el gimnasio

y hago de todo un poco

cuánto te cobran

mil trescientos pesos por mes

epa

- Cada segmento tiene solamente un hablante.
- Dos segmentos adyacentes tienen hablantes distintos.

Segmentación de hablantes

- Segundo paso: identificación de hablantes.
  - Dado el output de la diarización, agrupar los segmentos correspondientes al mismo hablante.

[segmento 1]
[segmento 2]
[segmento 3]
[segmento 4]
...

→

[segmento 1 - hablante 1]
[segmento 2 - hablante 2]
[segmento 3 - hablante 1]
[segmento 4 - hablante 3]
...

Segmentación de hablantes

sí este eh soy socio del de un gimnasio de acá a la vuelta

sobre juan be justo este vo- en realidad trato de ir tres días a la semana aunque no siempre puedo

qué tipo de ejercicios hacés en el gimnasio

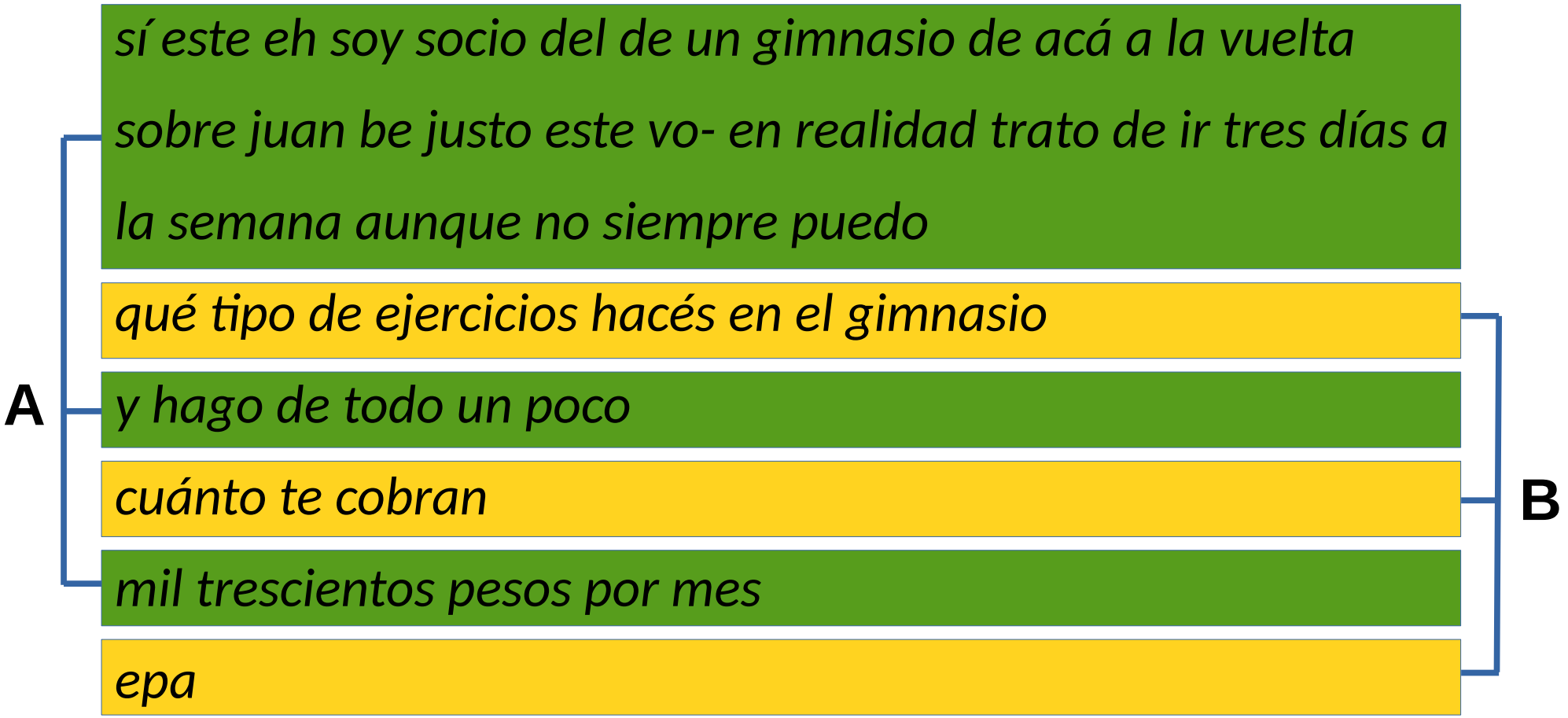
y hago de todo un poco

cuánto te cobran

mil trescientos pesos por mes

epa

Segmentación de hablantes



Segmentación de hablantes

- Diarización e identificación.
- Técnicas de clustering; HMMs con GMMs.

Segmentación del discurso

- Organización del discurso:
  - Estructura jerárquica de segmentos.
    - oraciones, párrafos, temas, parentéticos (*Quito, la capital de Ecuador, está situada en…)*, etc.
  - Puede modelarse con una estructura de pila (*stack*).
    - Nuevo segmento: *push frame*; fin segmento: *pop*.
    - *Frame*: contiene las entidades activas en el segmento.
  - Marcadores de discurso: *pero, sin embargo, además, cambiando de tema, a propósito, antes que me olvide…*
  - B. Grosz & C. Sidner *"Attention, intention, and the structure of discourse"* , Computational Linguistics, 1986.

Segmentación del discurso

- ¿Indicadores prosódicos de la segmentación del discurso?
  - Idea: Variamos nuestra prosodia para comunicar cambios en la estructura discursiva.
  - ¿Esos cambios son sistemáticos?
  - Para buscar evidencia:
    - Observaciones en cuerpos de datos grabados.
    - Experimentos de laboratorio: percepción y producción.

Segmentación del discurso

- Indicadores prosódicos de la segmentación del discurso:
  - Pausas entre segmentos.
  - Ritmo/velocidad del habla: Lento <SB> Rápido
  - Rango tonal (*pitch range*): Comprimido <SB> Expandido
  - Amplitud: Baja <SB> Alta
  - Contorno entonacional: H* L-L% <SB> H* L-H%

<SB> = Separador de dos segmentos.

Segmentación del discurso

- Problema práctico: Detectar automáticamente límites de oraciones y tópicos/historias en transcripciones ASR.
- Motivación:
  - Oraciones: para análisis sintáctico y semántico.
  - Tópicos/historias: para resúmenes, extracción de información.
- Algoritmos de ML entrenados sobre cuerpos de datos segmentados manualmente (ej: *Broadcast News*)
  - Atributos léxicos:
    - Dependientes del dominio (radio ≠ diálogo telefónico).
    - Sensibles a la performance de ASR.
  - Atributos acústicos y prosódicos:
    - Menos dependientes del dominio.
    - Sensibles a la identidad del hablante: estilo, cultura, etc.

Disfluencias

- El habla espontánea es *gramaticalmente incorrecta:*
  - Vacilaciones: *ch- change strategy* disfl1.wav
  - Pausas llenas: *um Baltimore* disfl2.wav
  - Reparaciones: *Ba- uh Chicago* disfl3.wav
- **Muy** comunes en el habla espontánea.
- Gran problema para ASR.
- ¡Pero no para los humanos!
  - Procesamos/ignoramos las disfluencias casi sin notarlo.
  - Ironía: Por ello mismo, es difícil rotularlas en grabaciones.

Disfluencias

- **Pausas llenas**
  - *eh*, *em*, *mh*, *este*, ...
  - ASR: En general se las trata como palabras comunes.
  - Función importante en conversaciones: ganar o conservar la palabra, mientras se prepara la frase.

Disfluencias

- **Reparaciones**: varios tipos.
  - Repeticiones
    - *bueno si hubieran * si hubieran hecho…*
  - Sustituciones por palabras de la misma clase
    - *era la * esa forma de …*
  - Sustituciones por constituyentes sintácticos similares
    - *me parece que hay * es más estricto en …*
  - Reinicios
    - *como te decía la * querés algo de tomar?*
  - Con o sin fragmentación
    - *la informática * computación es una disciplina…*
    - *la informát- * computación es una disciplina…*

Disfluencias

- Estructura de las reparaciones
  - *reparando * edición reparación* (* = punto de interrupción)

    edición = {Ø, eh, este, digo, …}
- Ejemplos:
  - *bueno si hubieran si hubieran hecho…*
  - *y la em lo que te decía era…*
  - *como te decía la eh querés algo de tomar?*
  - *la informát- digo la computación es una disciplina…*

Disfluencias

- ¿Cómo sabe el oyente (o una computadora) con qué quedarse y qué descartar?
  - Fragmentos de palabras: *informát-*, *cuand-*.
  - Pausas llenas: *eh*, *este*.
  - Palabras explícitas: *digo*, *es decir*.
  - ¿Existen fenómenos acústicos/prosódicos que ayuden a identificar disfluencias automáticamente?

Disfluencias

- Problemas prácticos de ASR: Detección y corrección de disfluencias.
- Enfoques:
  - Parsing sintáctico (si hay constituyentes donde debería haber solo uno, borrar el primero).
  - Pattern matching (~~DET uh~~ DET)
  - ML con atributos léxicos (modelo del lenguaje, POS), acústicos (tono, duración de fonos y pausas), calidad de la voz (jitter, spectral tilt).
- En general, detección más fácil que corrección.

Otras tareas de posprocesamiento

- Fechas

veinte de junio de mil nueve sesenta → *20/06/1960* ó *20 de junio de 1960*

en los sesentas → *en los ’60s*

- Abreviaturas

doctor roberts → *Dr. Roberts*

la de ge i → *la DGI*

Otras tareas de posprocesamiento

- Números

dos pesos con diez → $ 2,10

cero ochocientos diez dos tres dos ocho cinco tres siete → 0-810-232-8537

- Puntuación y capitalización

los beatles eran cuatro john lennon paul mccartney george harrison y ringo starr → Los Beatles eran cuatro: John Lennon, Paul McCartney, George Harrison y Ringo Starr.

Otras tareas de posprocesamiento

- Fechas
- Abreviaturas
- Números

} Transductores de estados finitos (FST)

- Puntuación
- Capitalización

} Fuertemente ligadas a la segmentación en oraciones. Clasificadores ML con atributos léxicos y prosódicos.

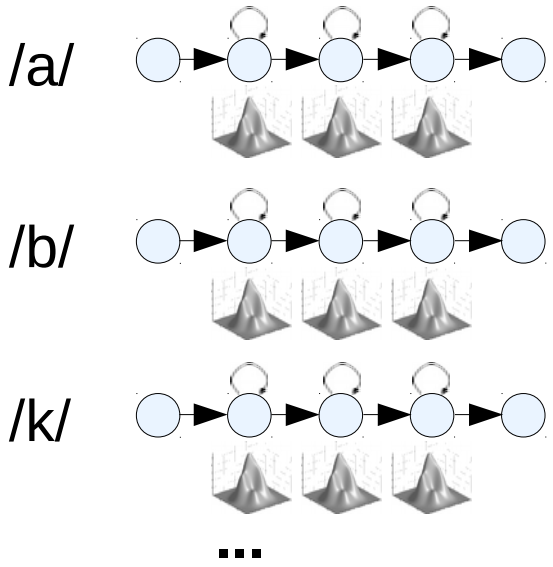
3 Herramientas Gratuitas para ASR

- HTK (Hidden Markov Model Toolkit)
  - http://htk.eng.cam.ac.uk
  - Licencia restrictiva para fines no académicos.
- Kaldi
  - http://kaldi.sourceforge.net
  - Código abierto, licencia Apache.
- CMUSphinx y PocketSphinx
  - https://cmusphinx.github.io/
  - Código abierto, licencia BSD.

PocketSphinx

- https://cmusphinx.github.io/
- Instalación: https://cmusphinx.github.io/wiki/tutorialpocketsphinx/
- Modelos entrenados para el español:
  - https://sourceforge.net/projects/cmusphinx/files/Acoustic%20and%20Language%20Models/Spanish/

Modelo acústico



Diccionario Fonético

a a
aaron a a r o n
ab a b
abajo a b a j o
abandona a b a n d o n a
abandonada a b a n d o n a d a
abandonadas a b a n d o n a d a s
...
cabaña k a b a gn a
cabañas k a b a gn a s
cabe k a b e
cabecilla k a b e s i l l a
cabello k a b e ll o
...

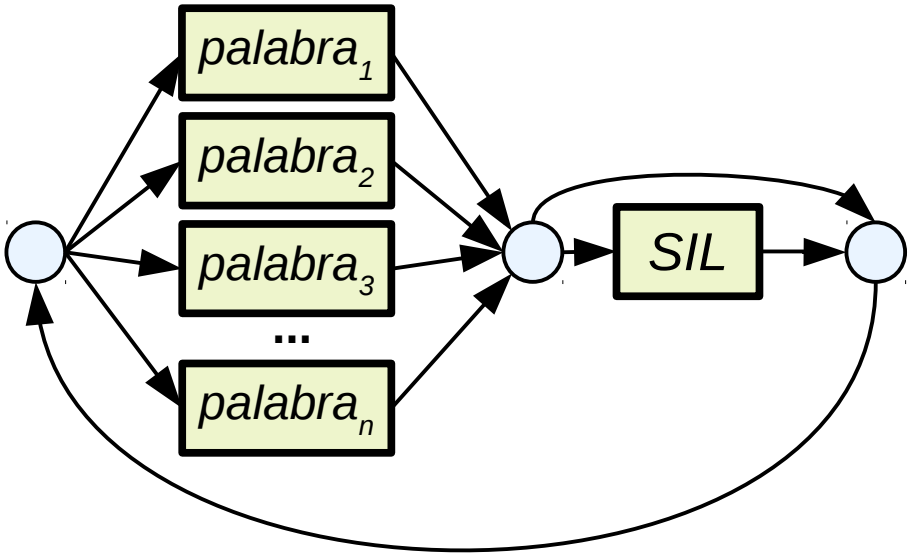
Modelo del Lenguaje

Probabilidades de:

- Unigramas
- Bigramas
- Trigramas

Ejemplos con PocketSphinx

- Ver archivo README.txt
- Correr los comandos “preparativos” (0).
- Correr el comando “(1) Por micrófono, usando el modelo de lenguaje”.
- Probar el sistema, hablando cuando dice “Ready…”.
  - “uno dos tres cuatro cinco”
  - “hola mundo”
  - “esto es una prueba”
  - …



+ LM

Ejemplos con PocketSphinx

- Correr el comando “(2) Por micrófono, usando una gramática específica”.

```
#JSGF V1.0;
grammar isolated_phrases;
<saludo> = (hola | buen día | buenos días);
<destinatario> = (mundo | argentina | universidad);
public <grammar> = <saludo> <destinatario>;
```

- Probar el sistema, hablando cuando dice “Ready...”.
  - “hola mundo”
  - “buen día argentina”
  - “buen día mundo”
  - …

Ejemplos con PocketSphinx

- Procesamiento en modo batch:
  - Comando (3): procesa los archivos wav listados en **`grabaciones.txt`**, usando un LM.
    - Resultados → **`$OUTDIR/salida-con-lm.txt`**
  - Comando (4): idem, pero guardando las lattices con las hipótesis más probables.
    - Lattices → **`$OUTDIR/lattices/`**
  - Comando (5): idem (3), pero usando una gramática restringida.
    - Resultados → **`$OUTDIR/salida-con-gramatica.txt`**

Ejemplos con PocketSphinx

Ejercicio:

Pensar una gramática restringida que sirva para implementar parte de un sistema de diálogo hablado que resuelva una tarea acotada.

Ejemplos:

- Respuestas simples (ej: “sí/no”, fechas, horas).
- Ingresar la clave bancaria de 4 dígitos.
- Comandos de navegación (“arriba”, “abajo”, “izquierda”, “adelante”, etc.).
- Comandos para leer los mails (“siguiente”, “archivar”, “borrar”, etc.).
- Elegir una provincia, luego una ciudad/localidad.

Se pide:

1) Escribir la gramática.
2) Si faltan palabras en el diccionario fonético, agregarlas.
3) Probar el reconocedor con esta gramática, y compararlo con un reconocedor que no use una gramática restringida.