CHARLA DE BORRACHOS:
¿Me escuchás bien? ¡Fuerte y claro!

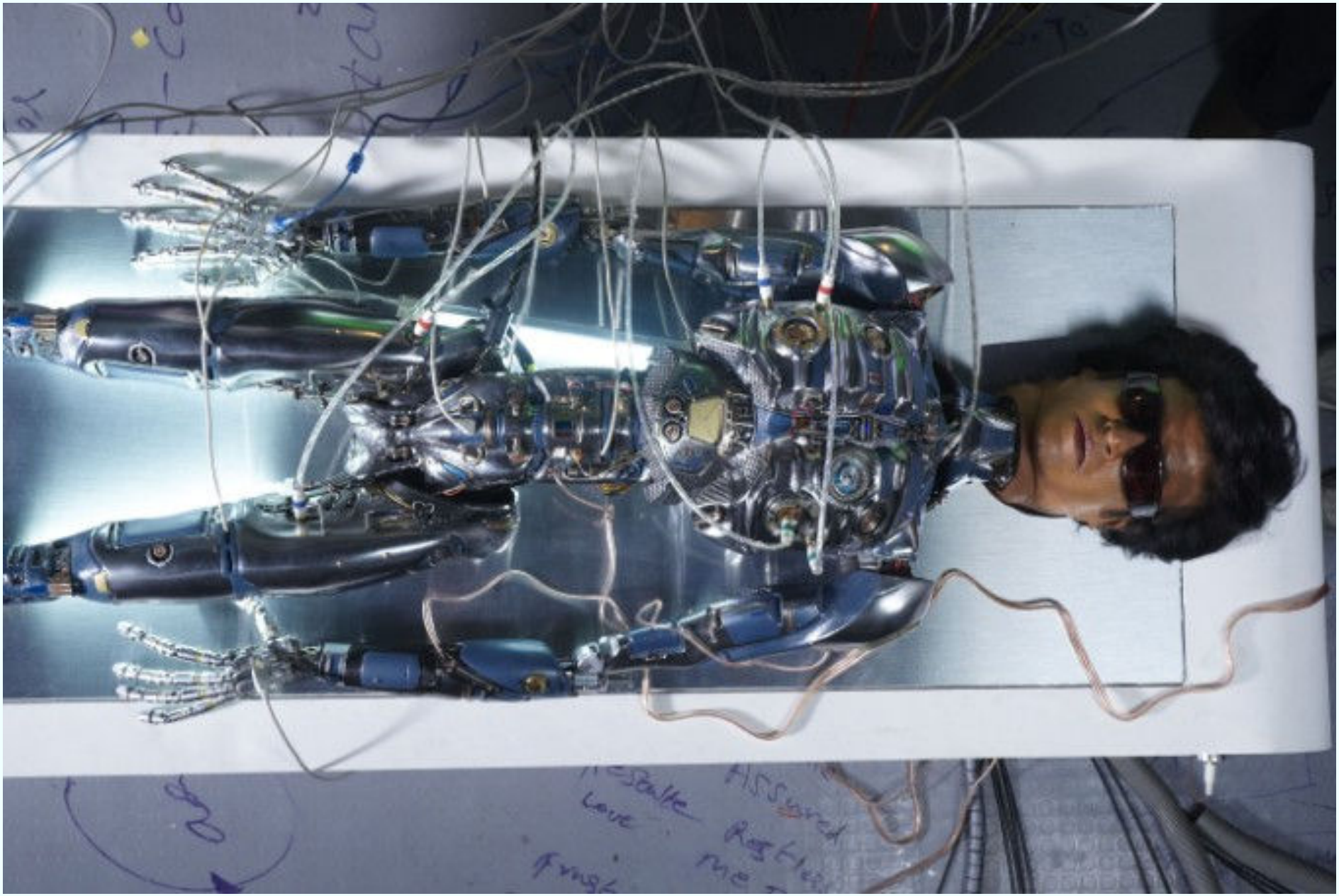
Cómo lograr hacer hablar mejor a una computadora

Christian G. Cossio Mercado

Lab. de Investigaciones Sensoriales, INIGEM, CONICET-UBA, Htal. de Clínicas

Jueves 6 de junio de 2013

¿Qué hacés?



¿Qué hacés?

¿Qué hacés?

¿Qué hacés?

¿Qué hacés?

¿Que qué hago?

¿Que qué hago?

- **“Evaluación Automática de la Calidad del Habla Artificial”**
- Ver qué tan bien suena el habla generada por una computadora.
- Tenemos habla sintetizada en muchos lugares! (A1)
  - Lectura de mensajes de texto
  - Sistemas de gestión telefónica (IVR)
  - ...

La comunicación humana

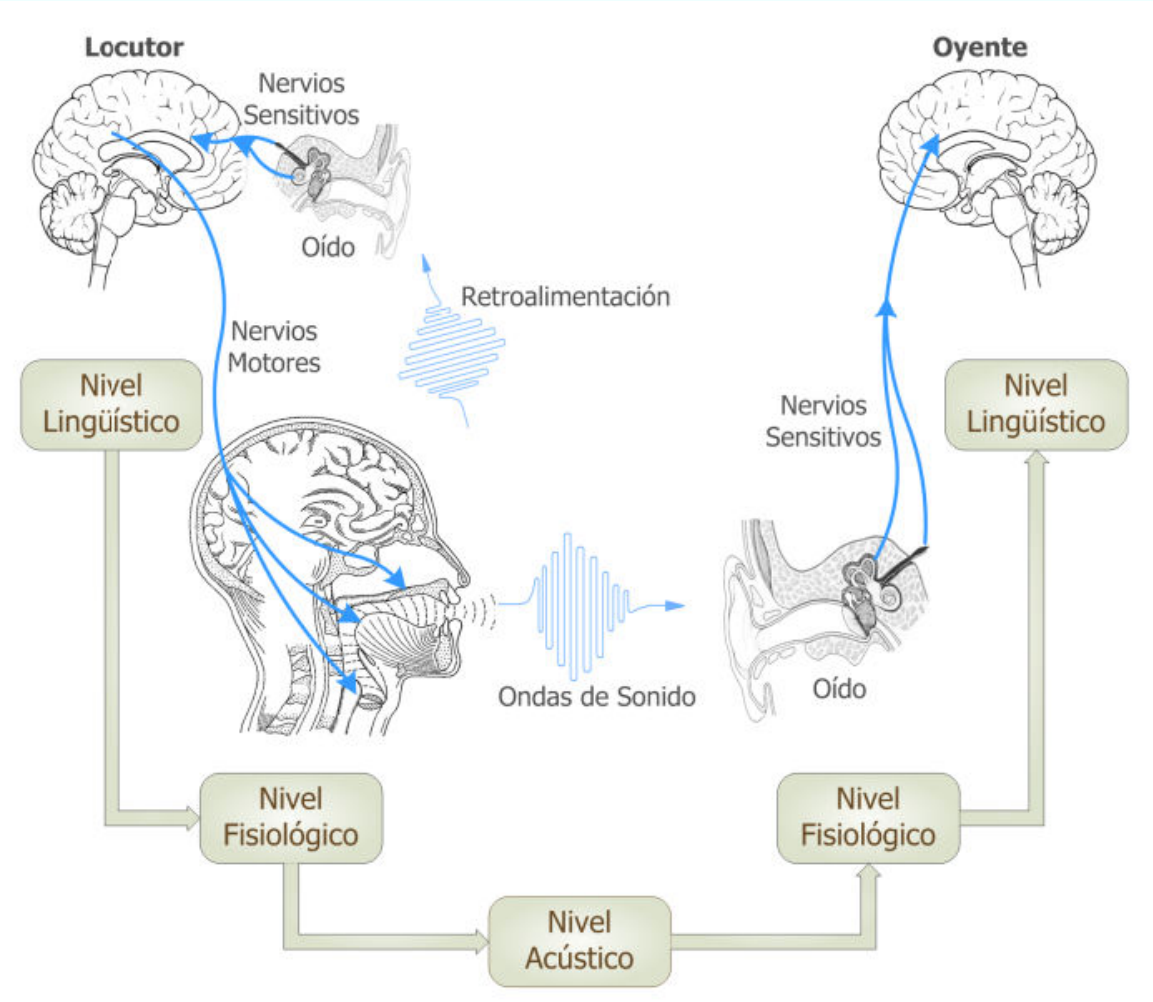


Figura: Cadena del Habla. Tomado de [7]

La comunicación humana

- La comunicación es naturalmente multimodal (V1)
  - escuchamos (e.g., un grito de aviso)
  - vemos (labios, gestos, ademanes, actitudes, . . .)
  - olemos (e.g., algo que se quema)
  - y tocamos (e.g., algo caliente) . . . aunque en menor medida

¡Pero sólo trabajaremos con el habla!

- Nos quedaremos sólo con la audición
- Recordemos que descartamos las otras fuentes de información comunicacional.
- Y, principalmente, perdemos el canal visual!

¿Y cómo entendemos lo que oímos?

- Habría varias alternativas. . .
- ¿Cuál creen que es la que utilizamos?
  - Se parte de lo percibido, y llegamos a lo que se quiso decir ($\uparrow$)
  - Se escucha el todo, y se trata de reconocer las mejores hipótesis posibles de acuerdo a lo que escuchado en forma global ($\downarrow$)
  - Cuando se puede es top-down, y cuando no se hace bottom-up ($\uparrow \downarrow$)

¿Y cómo entendemos lo que oímos?

- ¿Cuál creen que es la que utilizamos? (A2)
  - Se parte de lo percibido, y llegamos a lo que se quiso decir ($\uparrow$)
  - Se escucha el todo, y se trata de reconocer las mejores hipótesis posibles de acuerdo a lo que escuchado en forma global ($\downarrow$)
  - Cuando se puede es top-down, y cuando no se hace bottom-up ($\uparrow \downarrow$)
- Respuesta...
  - La percepción del habla suele top-down ($\downarrow$), pero también es bottom-up ($\uparrow$)
  - Que haya opción a utilizar información top-down econonomiza recursos cognitivos
  - La percepción está directamente conectada con las expectativa

Sistemas de Diálogo

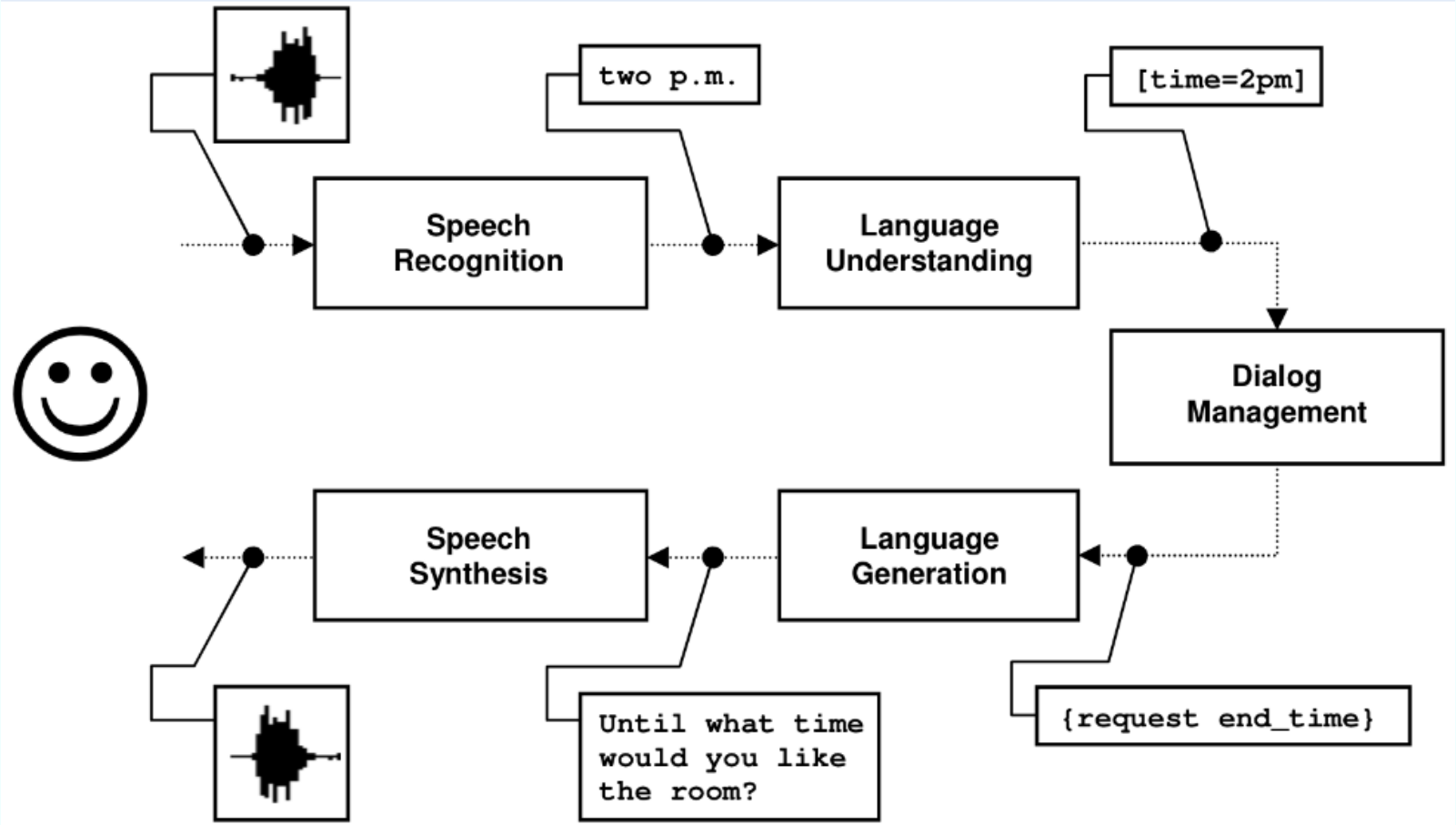


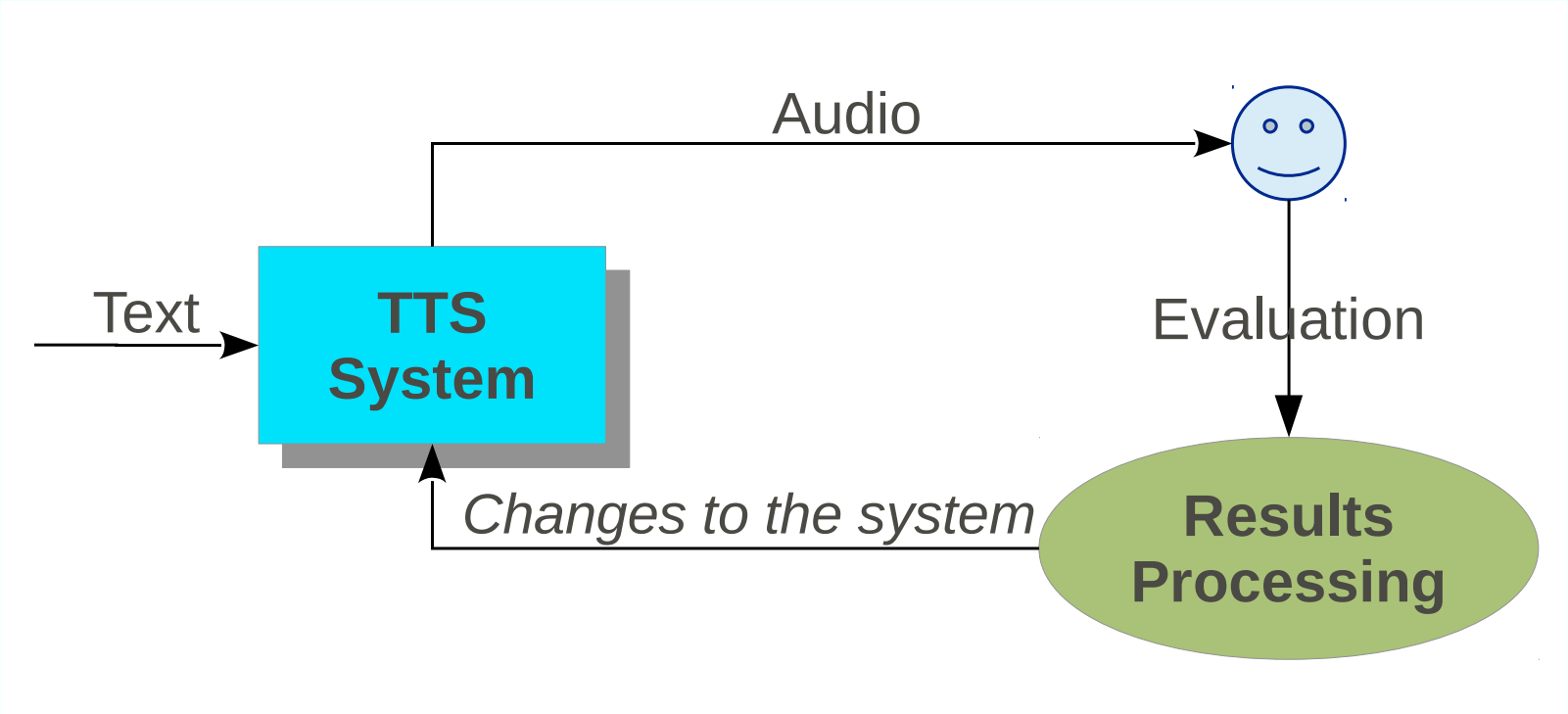
Figura: Arquitectura clásica de un sistema de diálogo. Tomado de [3]

.

Evaluaciones Actuales

- Hay muchas técnicas y procedimientos. . .
  - ¿Cómo suena en general? (MOS)
  - ¿Qué se entiende de lo que se dice? (SUS)
  - ¿Cómo suena, qué se entiende y es aceptable lo que se dice? (ITU)

Evaluaciones Actuales



Evaluaciones Actuales

Inteligibilidad

Impresión global

¿Cómo juzga la calidad sonora de lo que acaba de oír?

- ☐ Excelente
- ☐ Buena
- ☐ Pasable
- ☐ Mediocre
- ☐ Mala

Esfuerzo de escucha

¿Cómo describiría Vd el esfuerzo de escucha necesario para entender el mensaje?

- ☐ Descanso absoluto; ningún esfuerzo
- ☐ Atención necesaria; esfuerzo requerido no apreciable
- ☐ Esfuerzo moderado
- ☐ Esfuerzo considerable
- ☐ Significado incomprensible a pesar de todos los esfuerzos posibles

Dificultades de comprensión

¿Ha encontrado algunas palabras difíciles de entender?

- ☐ Nunca
- ☐ Raramente
- ☐ De vez en cuando
- ☐ A menudo
- ☐ Todo el tiempo

Nitidez

¿Eran nítidos los sonidos?

- ☐ Sí, muy claros
- ☐ Sí, suficientemente claros
- ☐ Medianamente claros
- ☐ No, no muy claros
- ☐ No, en absoluto

Aceptación

¿Piensa que esta voz podría utilizarse para un servicio telefónico de este tipo?

- ☐ Sí
- ☐ No

Evaluaciones Actuales

ITU Calidad

Impresión global

¿Cómo juzga la calidad sonora de lo que acaba de oír?

- ☐ Excelente
- ☐ Buena
- ☐ Pasable
- ☐ Mediocre
- ☐ Mala

Claridad

¿Ha detectado algunas anomalías en la pronunciación?

- ☐ No
- ☐ Sí, pero no molestas
- ☐ Sí, un poco molestas
- ☐ Sí, molestas
- ☐ Sí, muy molestas

Velocidad al hablar

La velocidad media de lo anunciado era:

- ☐ Demasiado rápido
- ☐ Un poco demasiado rápido
- ☐ Satisfactorio
- ☐ Un poco demasiado lento
- ☐ Demasiado lento

Agrado

¿Cómo describiría Ud. la voz?

- ☐ Muy agradable
- ☐ Agradable
- ☐ Pasable
- ☐ Desagradable
- ☐ Muy desagradable

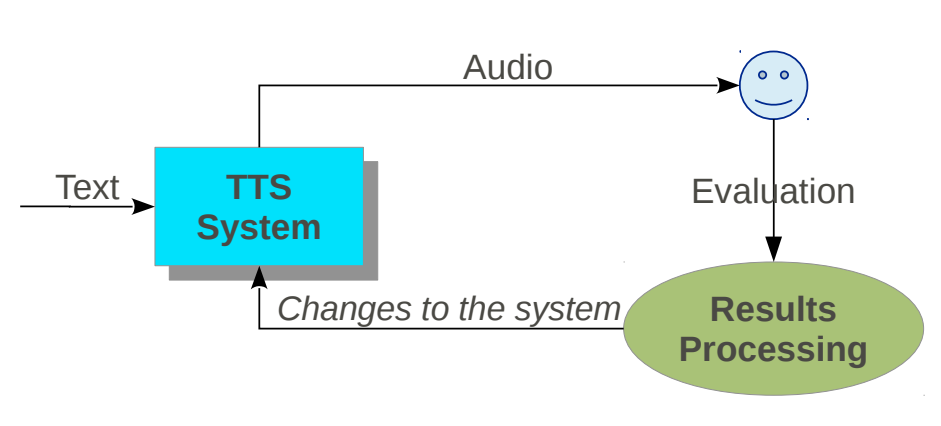
Aceptación

¿Piensa que esta voz podría utilizarse para un servicio telefónico de este tipo?

- ☐ Sí
- ☐ No

Observaciones:

Evaluaciones Actuales



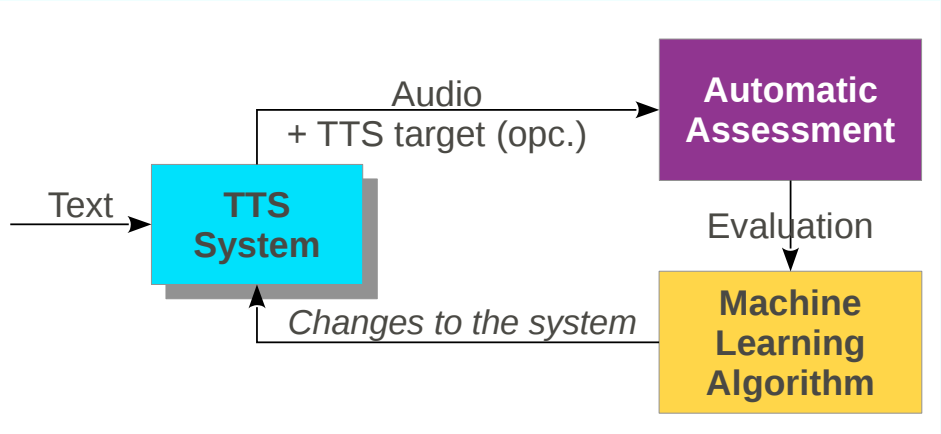
- Los métodos actuales tienen muchos costos asociados
  - Tiempo
  - Dinero ($$$)
  - De registro y procesamiento de los resultados

- **Se deben buscar otras opciones si se quiere utilizar métodos de aprendizaje automático para el ajuste de los sistemas TTS**

¿Qué cosas debemos evaluar?

- Inteligibilidad (¿se entiende lo dicho?)
- Naturalidad (e.g., velocidad del habla)
- Aceptabilidad (¿es usable?)
- Agradabilidad (¿suena lindo?) (!)

¿Cómo las podemos evaluar?



- Desde el punto de vista púramente acústico (e.g., disrupciones en la señal, clicks, plops)
- Desde la percepción subjetiva del habla (e.g., se escucha suave/aspero)
- Desde la consistencia audio-texto (e.g., .Estoy muy triste" leído en voz alegre)
- Si las características de la voz se corresponden con las del texto (A3)

CHARLA DE BORRACHOS:
¿Me escuchás bien? ¡Fuerte y claro!

Cómo lograr hacer hablar mejor a una computadora

Christian G. Cossio Mercado

Lab. de Investigaciones Sensoriales, INIGEM, CONICET-UBA, Htal. de Clínicas

Jueves 6 de junio de 2013

Preguntas, Dilemas, Disyuntivas, Dudas existenciales...

Cosas varias

- En la próxima ECI 2013 estaremos dando un curso
  - *N3 - Reconocimiento automático de habla e identificación de locutores*
- Recomendación: Hacer la materia de Agustín Gravano de Procesamiento de Habla, y la de José Castaño sobre Proc. de Lenguaje Natural.

Cosas varias

- El tema de investigación es muy amplio
- Siempre es bienvenido que haya gente que se quiera sumar
- Me ubican en. . .
  - Por mail: cgcossio [arroba] gmail
  - En twitter: @CGCossioM
  - En el LIS: http://lis.secyt.gov.ar

¡MUCHAS GRACIAS POR VENIR!

Videos recomendados

- Síntesis de Habla
  - Charla de Simon King (Edinburgh University): http://youtu.be/xzL-pxcpo-E
  - Charla de Kim Silverman (Apple): http://youtu.be/7mjh0PSUv0M
- Efecto McGurk
  - Explicación: http://youtu.be/G-IN8vWm3m0
  - Ejemplo: http://youtu.be/PWGeUztTkRA
- Cómo escuchamos: http://youtu.be/stiPMLtjYAw

Algunas referencias útiles I

[1] Antons, J-N, Schleider, R., Arndt, S., Möller, S., Porbadnigk, A.K., and Curio, G., "Analyzing Speech Quality Perception Using Electroencephalography". En *IEEE Journal of Selected Topics in SP*, vol. 6, no. 6, 2012.

[2] Benoît, C., Grice, M., Hazan, V., "The SUS test. A method for the assessment of TTS synthesis intelligibility". En *Speech Communication*, vol. 18, 381–392, 1996.

[3] Bohus, D., Rudnicky, A.I., "The RavenClaw dialog management framework: Architecture and systems". En *Computer Speech & Language*, 23 (3), pp. 332–361, 2009.

[4] Campbell, N., "Evaluation of Speech Synthesis". En Dybkjær, L., Hemsen, H., Minker W. (Eds.), *Evaluation of Text and Speech Systems*, pp. 29–64, Springer, 2007.

Algunas referencias útiles II

[5] Davis, M.H., Johnsrude, I.S., "Hearing Speech Sounds. Top-Down Influences on the Interface between Audition and Speech Perception". En *Hearing Research*, vol. 229, pp. 135–147, 2007.

[6] Delogu, C., Conte, S., Sementina, C., "Cognitive Factors in the Evaluation of Synthetic Speech". En *Speech Communication*, vol. 24, pp. 153–168, 1998.

[7] Evin, D.A., *Incorporación de Información Suprasegmental en el Proceso de Reconocimiento Automático del Habla*, Tesis de Doctorado. Fac. de Ciencias Exactas y Naturales, Universidad de Buenos Aires, 2011.

[8] Falk, T.H. and Möller, S., "Towards Signal-Based Instrumental Quality Diagnosis for Text-to-Speech Systems", *Signal Processing Letters, IEEE*, vol.15, pp. 781–784, 2008.

Algunas referencias útiles III

[9] Grancharov, V., Kleijn, W.B., "Speech Quality Assessment". En Benesty, J., Mohan Sondhi, M., Huang, Y. (Eds.) *Springer Handbook of Speech Processing*, pp. 83–98. Springer, 2008.

[10] Gurlekian, J.A., Cossio-Mercado, C.G., Torres, H.M and Vaccari, M.E., "Subjective Evaluation of a High Quality Text-to-Speech System for Argentine Spanish", En *Proceeding of IberSPEECH 2012*, pp. 241–250, Madrid, noviembre 2012.

[11] ITU-T Rec. P.85, *A Method for Subjective Performance Assessment of the Quality of Speech Voice Output Devices*. ITU, 1994.

[12] Jekosch, U., *Voice and Speech Quality Perception, Assessment and Evaluation*. Springer, 2005.

[13] Hinterleitner, F., Möller, S., Norrenbrock, C. and Heute, U., "Perceptual Quality Dimensions of Text-to-Speech Systems". En *Proc. Interspeech'11*, pp. 2177–2180, 2011.

Algunas referencias útiles IV

[14] Lattner, S., Maess, B., Wang, Y., Schauer, M., Alter, K., Friederici, A.D., "Dissociation of Human and Computer Voices in the Brain: Evidence for a Preattentive Gestalt-like Perception". En *Human Brain Mapping*, vol. 20, pp. 13–21, 2003.

[15] Mayo, C., Clark, R.A.J., King, S., "Listeners' weighting of acoustic cues to synthetic speech naturalness". En Speech Communication, vol. 53, pp. 311-326, 2011.

[16] Möller, S., Hinterleitner, F., Falk, T.H. and Polzehl, T., "Comparison of Approaches for Instrumentally Predicting the Quality of Text-to-Speech Systems". En *Proc. Interspeech'10*, 2010.

[17] Norrenbrock, C.R., Hinterleitner, F., Heute, U., and Möller, "Instrumental Assessment of Prosodic Quality for Text-to-Speech Signals". En *IEEE Signal Processing Letters*, vol. 19, no. 5, 2012.

Algunas referencias útiles V

[18] Patel, A.D., *Music, Language and the Brain*. Oxford University Press, New York, 2008.

[19] Pisoni, D.B., "Perception of Synthetic Speech". En van Santen, J.P.H., Sproat, R.W., Olive, J.P. and Hirschberg, J. (Eds.), *Progress in Speech Synthesis*, pp. 541–560, Springer, 1997.

[20] Pisoni, D.B. and Remez, R.E. (Eds.), *The Handbook of Speech Perception*. Blackwell Publishing, 2005.

[21] Sutton, R.S., and Barto, A.G., *Reinforcement Learning: An Introduction*. MIT Press, Cambridge, MA, 1998.

[22] Taylor, P., *Speech Synthesis*. Cambridge Press, England, 2009.

[23] Torres, H.M., Gurlekian, J.A., Cossio-Mercado, C.G., "Aromo: Argentine Spanish TTS system", En *Proceeding of IberSPEECH 2012*, pp. 416–421, Madrid, nov. 2012.

Algunas referencias útiles VI

[24] Vazquez-Alvarez, Y. and Huckvale, M., "The Reliability of the ITU-P.85 Standard for the Evaluation of Text-to-Speech Systems". En *Proc. ICSLP-2002*, pp. 329–332, 2002.

[25] Winters, S.J., and Pisoni, D.B., "Speech synthesis: Perception and Comprehension". En *Encyclopedia of Language and Linguistics, Second Edition*, Vol.12, 31–49. Elsevier, 2005.