Modelo sistémico para el procesamiento acústico del habla: el caso de la locución de noticias

Modelo sistêmico para o processamento acústico da fala: o caso da locução de notícias

Systemic model for acoustic del habla el procesamiento: el caso de la locucion news

Ángel Rodríguez Bravo[1]
(angel.rodriguez@uab.es)
Lluís Mas Manchón[2]
(lluis.mas@uab.es)

Resumen
Este artículo propone un modelo para la comunicación hombre-máquina que aporta nuevos puntos de vista para el desarrollo de los sistemas de procesamiento del habla. El modelo presentado se pone a prueba aplicándolo al procesamiento del habla continua, para buscar criterios de procesamiento que permitan la segmentación automática en unidades-noticia de cualquier locución informativa desde una perspectiva intra-locutor.

Palabras-clave: Comunicación. Lenguaje. Unidades-news.

Resumo
Este artigo propõe um modelo para a comunicação homem-máquina que aporta novos pontos de vista para o desenvolvimento dos sistemas de processamento da fala. O modelo apresentado pode ser comprovado ao ser aplicado ao processamento da fala contínua, para buscar critérios de processamento que permitam a segmentação automática em unidades-notícia de qualquer locução informativa a partir de uma perspectiva *intra-locutor*.

Palavras-chave: Comunicação. Fala. Unidades-notícia.

Abstract
This article proposes a model for human-machine communication that brings new perspectives for the development of speech processing systems. The model presented can be proved to be applied to continuous speech processing, processing to search criteria to enable the automatic segmentation units news information of any utterance from a speaker intra perspective.

Keywords: Communication. Speech. Units-news.

[1] Universidad Autónoma de Barcelona (España). Laboratorio de Análisis Instrumental de la Comunicación (LAICOM)
[2] Universidad Autónoma de Barcelona (España). Laboratorio de Análisis Instrumental de la Comunicación (LAICOM)

Introducción

En el año 1981 S. E. Levinson y M. Y. Liberman (1981) afirmaban: "después de más de 40 años de investigación, el reconocimiento del habla natural o conversacional sigue siendo un objetivo utópico". Desde principios de los 80 hasta hoy, el diseño de sistemas automáticos oyentes, parlantes o traductores es una de las corrientes de investigación a la que se han dedicado más esfuerzos económicos y humanos en todo el mundo; sin embargo, tras 70 años de trabajo, los sistemas automáticos de reconocimiento de habla siguen siendo herramientas profundamente limitadas.

La calidad de los resultados obtenidos en esta línea de investigación tiene que hacernos pensar en que los paradigmas que dominan la investigación sobre el procesamiento del habla arrastran errores fundamentales. ¿Pero cuáles son estos errores?

Aparentemente, la dificultad esencial para que las máquinas comprendan el lenguaje oral está en la gran variabilidad de la señal vocal. No obstante, la investigación básica está hoy ya lo bastante madura y desarrollada como para aceptar que la enorme variabilidad que "distorsiona" los estándares sonoros teóricos (fonemáticos, léxicos y sintácticos) de una lengua no es aleatoria sino motivada y sistemática. Y que toda esta variabilidad está orientada por modelos expresivos (BRAVO, 1988-89). Si aceptamos que este planteamiento fenomenológico es correcto, la estructura que gobierna la interpretación automática del habla debería contemplar siempre un nivel de procesamiento capaz de reconocer y de dar sentido a todas estas "distorsiones" expresivas.

Para poder avanzar en la solución de los problemas que todavía paralizan la evolución en el procesamiento del habla hemos de aceptar que se ha producido un cambio profundo y que la necesidad de abordar la comunicación desde la perspectiva de la significación ha alcanzado ya de lleno a la tecnología. El problema tecnológico del reconocimiento de formas, resuelto todavía muy parcialmente y con un enorme gasto de recursos informáticos, tiene ahora la necesidad de utilizar criterios científicos directamente vinculados con la significación. Esto supone la necesidad de romper definitivamente con la idea de que "los aspectos semánticos de la comunicación son irrelevantes desde la perspectiva de la ingeniería" (SHANNON y WEAVER, 1981). ¿Qué es el reconocimiento de formas si no que el establecimiento de criterios de asignación de sentido a determinados parámetros obtenidos a partir del muestreo exhaustivo de una "masa comunicativa"?

Lo que estamos defendiendo es que en el momento actual de desarrollo del saber sobre el procesamiento del habla es necesario dirigir la atención hacia los aspectos semánticos de la comunicación y, por tanto, hacia la integración de un modelo comunicativo que deje de separar entre la aproximación matemática y la aproximación semántica del mismo.

1 Propuesta de un nuevo modelo de comunicación hombre-máquina

Siendo consecuentes con esta reflexión, vamos a proponer un modelo apoyado en la teoría de sistemas, que parte de la necesidad de localizar y analizar las diferencias esenciales entre las estructuras de procesamiento del ser humano y las de los sistemas de acción robótica. En tanto que los sistemas de procesamiento automático del habla tienen siempre como objetivo último la comunicación hombre-máquina, partimos de la convicción de que el referente funcional de los sistemas automáticos de reconocimiento ha de ser el ser humano. En consecuencia, pensamos que un modelo capaz de poner de relieve las fases de construcción de la significación en el proceso de comunicación habrá de resultar útil para el avance de los sistemas automáticos que han de ser dirigidos por el valor significante de las formas procesadas.

El modelo que presentamos más abajo muestra la interacción entre un sistema emisor-receptor humano con un sistema automático y pone de relieve las siguientes cuestiones fundamentales: 1) la doble salida del sistema humano que diferencia MENSAJES (secuencias formales con objetivos comunicativos) y ACCIONES EXTERNAS (actuaciones de adaptación al medio); 2) la capacidad del ser humano de obtener significación de cualquier ESTÍMULO PRIMARIO del entorno; 3) la organización del sistema humano en dos módulos de procesamiento diferenciados, uno receptor y otro emisor, que interactúan con "n" memorias de trabajo (tantas como códigos expresivos) en las que se almacenan bibliotecas de formas predefinidas; 4) la capacidad del sistema humano de utilizar recursos tecnológicos para construir artefactos comunicativos eficaces. Frente a esto nos encontramos con que un sistema comunicativo de procesamiento automático: 1) dispone de una única salida (OUTPUTS) que indica el reconocimiento de formas; 2) tiene grandes problemas para discriminar entre el ruido y los mensajes, no es capaz de extraer información útil de los ESTIMULOS PRIMARIOS ya que los trata siempre como ruido; 3) dispone de un módulo de procesamiento receptor pero no de procesamiento emisor; 4) utiliza un módulo de procesamiento receptor que normalmente trabaja con

un código expresivo único (reconocimiento, fonemático, léxico y morfosintáctico) que interactúa solo con una memoria de trabajo.

Pero veamos de qué modo un planteamiento tan general como el que proponemos puede dar respuestas al problema concreto del procesamiento del habla.

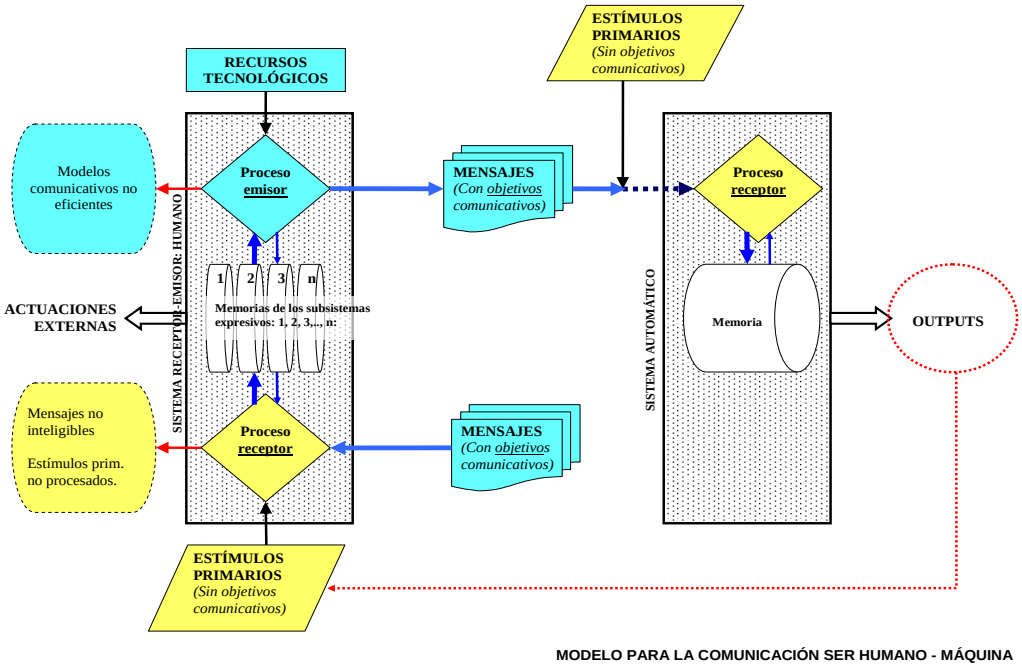
Entrando al modelo por la parte inferior central – MENSAJES *(con objetivos comunicativos)-*, y entendiendo que estos MENSAJES son secuencias concretas de discurso oral, por ejemplo: locuciones informativas radiofónicas, podemos obtener deducciones interesantes. Las "distorsiones" de la locución de un noticiario de radio, se organizan en distintos niveles de expresión que comunican al oyente información sobre los objetivos comunicativos: la relevancia diferencial de cada uno de los contenidos, la actitud emocional del locutor, el inicio y el final de la unidad informativa, los rasgos fisiológicos diferenciales de emisor; o bien sobre la forma, el tamaño, el tipo de movimiento, etc. de aquello que describe oralmente el locutor. Son transgresiones que en realidad constituyen rasgos sonoros interpretables, y que no pertenecen a las formas léxicas o a las formas gramaticales, ni son tampoco definitorios de la estructura acústica estándar de los fonemas, aunque se encabalgan sobre ellas modificándolas. Dicho de otro modo, e interpretando la cuestión que nos ocupa en función del modelo que proponemos, las locuciones de noticias son procesados como una estructura compleja que articula un flujo de formas primarias (tensión emocional, ataque de inicio y caída de final, evolución de la velocidad de locución, formas entonativas, etc.) con un segundo flujo de formas expresivas culturales (fonemas, lexemas, estructura sintáctica, modelos tonales etc.).

Estamos procesando, en suma, un doble flujo significante: cultural y primario, que articula simultáneamente datos estructurados en formas según "n" códigos expresivos, que han de ser confrontados con las formas predefinidas de las respectivas bibliotecas soportadas por "n" memorias de trabajo.

Comparemos ahora la parte humana del proceso con la parte del modelo que representa el procesamiento automático. A la vista del modelo, se hace evidente que una máquina debería estar provista de diferentes subrutinas de procesamiento y distintas memorias simultáneas de trabajo que le permitan seleccionar cuales son las formas de referencia adecuadas (en función de la actitud emocional del locutor, su velocidad de locución, los ataques y caídas de la intensidad, sus macro estructuras tonales etc.) para reconocer los rasgos expresivos que se articulan y confluyen en el mensaje oral. Así, disponer de un proceso receptor de subrutinas múltiples, permitiría a los sistemas

de reconocimiento identificar y extraer el significado de formas acústicas que han sido fuertemente modificadas en función de sus objetivos comunicativos y sus contextos expresivos, para confrontar semánticamente los valores de esta expresividad pre-verbal con el sentido resultante de la estructura léxica y morfosintáctica, matizándola o incluso modificándolo profundamente.

Figura 1: modelo para la comunicación ser humano - máquina



Fonte: autores

1 Hacia una solución para la segmentación automática en unidades-noticia de los noticiarios audiovisuales

Nuestra hipótesis de partida es considerar todas las dimensiones prosódicas de las locuciones informativas como un discurso integral, coherente respecto de su objetivo comunicativo básico. En este sentido, hemos aplicado el tipo de análisis del discurso que se deriva de la Teoría Intencional del Discurso (1986), de la Estructura Retórica (2005), y Atención Dinámica (1989). La interpretación de este análisis es congruente con la Teoría General de la Información y Teoría de la Noticia (1998).

Los tres pasos básicos del análisis que proponemos son los siguientes:

1. Dar prioridad a la estructura prosódica suprasegmental (en este nivel, el idioma es irrelevante).
2. Categorizar la información según su grado de noticiabilidad en: *nueva, mediada, vieja*; y según su contenido en *tema* y *rema* del evento.
3. Buscar en el análisis las variaciones de tono, tiempo e intensidad, que guardan relación con las categorías anteriores

Una vez experimentado este procedimiento sobre una muestra de noticias llegamos a los siguientes planteamientos:

1. La noticia es la manifestación física de un proceso de comunicación eficiente: el mensaje-noticia esta acústicamente marcado en su inicio y su final.
2. El mensaje noticia tiene dos niveles: el *estructural*: (fonemas, vocabulario y sintaxis) y el *superestructural* (entonación, tempo e intensidad). Nuestro objeto de estudio son las marcas superestructurales en las fronteras entre noticias.

Estas marcas dependen de:

a) La relación entre estructura y superestructura: el sentido de lo que se quiere comunicar depende completamente de la correcta utilización de la superestructura.
b) La reiteración de unos pocos modelos concretos en el formato superestructural de la noticia:
c) La variabilidad dependiente del locutor: en una locución de noticias correcta, esa variabilidad quedaría restringida al estilo personal del emisor.

Lógicamente, la búsqueda de marcas de inicio y final útiles para la segmentación de noticias se centrará en b).

En consecuencia, el modelo que estamos utilizando nos orienta a someter todo patrón y unidad de análisis de la señal acústica a la estructura de las formas expresivas del procesamiento humano, y no a la inversa. El procesamiento humano de la locución siempre se realiza de mayor a menor nivel de complejidad, y nunca parte de un conjunto de datos; sino que directamente se procesan formas expresivas que son grandes unidades complejas, significativas y autónomas.

Nuestro modelo nos indica que debemos definir algoritmos a partir de una mirada selectiva de la señal sonora que derive, a la vez, en un tratamiento selectivo de los datos y discrimine con claridad los distintos subsistemas expresivos que confluyen en la construcción del habla.

Finalmente, el modelo nos indica que las formas expresivas del nivel suprasegmental, que sirven a los humanos para segmentar los mensajes-noticia, se organizan en 3 subsistemas expresivos

funcionales: el de la *entonación*, el del *ritmo* y el de la *intensidad*, que interactúan con el subsistema del *contenido*.

Conclusiones: aproximación a un algoritmo para la segmentación de la locución informativa en noticias

Los resultados de aplicar nuestro modelo al problema de la segmentación automática de la locución informativa en unidades-noticia arrojan, a modo de hipótesis, el siguiente listado de criterios básicos que han de ser utilizados simultáneamente para el modelado de las formas en la señal acústica:

Nivel 0: *Obtención selectiva de datos* con el siguiente criterio:

- **Pitch:** dato de vibración vocal de cada sílaba en el centro de su parte sonora.

- **Sílaba:** duración de la parte sonora de cada golpe de voz.

-**Palabra-clave**: unidades mínimas de contenido nuevo asociadas a un campo semántico predeterminado (tópicos informativos) (RODRÍGUEZ BRAVO, 2006).

- **Pausa:** ausencia de locución de duración igual o mayor a 0.1 segundos (VILAR, 1999). La pausa es una unidad mínima de delimitación: todo segmento de locución está delimitado por una pausa, pero una pausa no siempre delimita un segmento.

Nivel 1: *Formas simples*. Se trata de los modelos de organización de los datos del nivel 0.

- **Reset:** el locutor realiza una *pausa* para tomar aire. A la pausa le sucede un pico tonal más alto que el pico inmediatamente anterior.

- **Contorno entonativo**: el pitch debe ser considerado como curva o contorno. Este contorno es continuo entre *resets* y actúa como unidad entonativa mínima.

- **Formas temporales**: están determinadas por la duración de las sílabas que constituyen cada contorno.

- **Pico tonal**: vértice superior (pitch máximo) del contorno coincidente con una palabra-clave.

- **Coda final:** las fronteras de los contornos entonativos están marcadas como una fuerte caída progresiva del tono y un aumento de la duración de las formas temporales.

Nivel 2: *Formas complejas*. Unidades formales que agrupan dos o más *formas simples.*

- **Ritmo de habla**: Forma que depende de la evolución de las *formas temporales*. El ritmo es más rápido cuantas más *sílabas* y *picos tonales* sean detectados por unidad de tiempo. Siguiendo el criterio perceptivo y tomado como ejemplo de referencia la sílaba:

velocidad ≤ 70 sílabas/minuto → ritmo lento;

velocidad = 70-80 sílabas/minuto → ritmo medio;

velocidad ≥ 80 sílabas/minuto → ritmo rápido.

- **Downtrend**: Forma que depende de la altura de los picos tonales de los *contornos entonativos*. Se caracteriza por la disminución sucesiva de la altura de los *picos tonales* como efecto de la pérdida progresiva de fuerza articulatoria entre *resets* (TERKEN, 2003).

- **Grupo entonativo**: Forma constituida por *contornos entonativos*, cuyo final está marcado siempre por la aparición consecutiva de una *coda fina*l y un *reset*.

- **Plateau:** Forma constituida por ausencia de variación tonal durante un espacio temporal de un duración mínima. Tiende a situarse en el centro temporal de la noticia.

- **Rema:** Forma constituida la aparición sucesiva de *palabras clave*.

Nivel 3: *Formas expresivas*: son patrones de significación del ser humano. Es un nivel de procesamiento que implica simultáneamente varios subsistemas expresivos

- **Prominencia**: Esta forma expresiva está determinada por la correlación entre los picos tonales, la disminución del ritmo de habla (MAEKAWA, 2000) y el aumento de la intensidad. El Modelo Penta desarrollado por Yi Xu (2004) coincide con nuestra propuesta en tanto que se articula, también, a partir de las funciones significativas. Xu define tres categorías para la intensidad: "fuerte", "débil" y "normal". En nuestro caso, el patrón de significación es la novedad y el interés informativo del *rema* donde se sitúa la prominencia.

- **Grupo fónico**: Es una forma expresiva que puede incluir uno o más *grupos entonativos* con sus correspondientes *downtrends*. Está delimitada por un pico tonal al principio del grupo, y una *coda final* seguida de un *reset* al terminar el grupo. Su significación es el efecto de agrupamiento del sentido en torno a un tópico informativo al cohesionar las *novedades* comunicadas.

- **Macromelodía**: Esta forma expresiva determina la sintaxis sonora global de la locución y, en consecuencia, su valor significante es priorizar unos tópicos informativos respecto a otros. Es fundamental para la inteligibilidad del mensaje.

La macromelodía está determinada por tres formas expresivas básicas:

- **Patrones de discurso**: Son estructuras que articulan varios *grupos entonativos* con un *ritmo* y un rango de variabilidad específicos, cuyo valor significante orienta al receptor sobre los objetivos comunicativos (informativo, persuasivo, emocional, etc.)

Estos patrones están determinados por *contornos entonativos* concretos y la presencia de puntos de inflexión (*picos tonales* más altos) en los grupos fónicos.

- **Párrafo**: Forma expresiva compuesta por un conjunto de remas separados por pausas sustancialmente mayores que las de los grupos fónicos. Engloba dos o más grupos fónicos.

- **Ritmo melódico**: Esta última forma expresiva está determinada por la distribución de formas temporales y picos tonales de los grupos fónicos en los niveles de párrafo y de noticia. Los picos tonales más altos de cada grupo fónico se definen como "top" (T), los medios como "mid" (M) y los bajos como "bottom" (B) (BAQUÉ y ESTRUCH, 2003). A partir de aquí, el fenómeno downtrend también se produce en el tercer nivel del discurso. El algoritmo MOMEL (ESPESSER y HIRST, 1993) etiqueta los picos como T_0, T_1, T_2, los valles como B_0, B_1, B_2…, y los niveles medios como M_0, M_1, M_2… en función de su pertenencia a los diferentes niveles: unidad entonativa (T_0, B_0, M_0), grupo fónico (T_1, B_1, M_1), el párrafo (T_2, B_2, M_2) y noticia (T_3, B_3, M_3).

Si nuestro *"Modelo para la comunicación ser humano - máquina"* representa y se ajusta con eficacia a la fenomenología del proceso que estamos investigando, los algoritmos para la segmentación en noticias de las secuencias de locución informativa diseñados a partir de estos criterios deberían ser eficientes. Si en nuestra siguiente etapa de investigación se cumple esta premisa, las hipótesis que estamos planteando quedarían contrastadas y el modelo sistémico que proponemos se consolidaría como una herramienta conceptual fértil para orientar la investigación del procesamiento del habla desde la perspectiva comunicológica.

Artigo submetido em 18/09/2013 e aceito em 25/09/2013.

Referencias

LEVINSON, Stephen; LIBERMAN, Mark. Reconocimiento del habla por medio de ordenadores. **Scientific American**, n. 57, p. 38-51, jun. 1981.

BRAVO, Ángel Rodríguez. Máquinas que hablan y escuchan. **Telos,** Madrid, n. 16, p. 117-124, dic./feb. 1988-89.

SHANNON, C.E; WEAVER,W. Teoría matemática de la comunicación. Madrid: Ediciones Forja, 1981.

Grosz, B.J.; Sidner, C.L. Attention, intentions, and the structure of discourse. **Computational Linguistics,** v. 12, n. 3, p. 175-204, jul./sept. 1986.

TABOADA, M.; MANN, W. C. Applications of rethorical structure theory. **Discourse Studies,** v. 8, n. 4, p. 567-588, 2005. Disponible en: <http://www.sfu.ca/~mtaboada/docs/Taboada_Mann_ RST_ Part2.pdf>. Acceso en: 19 oct. 2007.

JONES, M. R.; Boltz, M. Dynamic attending and responses to time. **Psychological Review**, v. 2, n. 3, p. 459-491, 1989.

Martínez Albertos, J.L. **Curso general de redacción periodística**. Madrid: Paraninfo, 1998.

RODRÍGUEZ BRAVO, et al. **Informe de recerca sobre els resultats obtinguts en la primera fase del projecte CLAIS (Classificador Automàtic d'Informació Sonora)**. Clave L. Ref.: Informe de investigación entregado al "Consell de l'Audiovisual de Catalunya" (CAC) en mayo de 2006. Se ha localizado 1000 en grupos para cada tema.

VILAR, N. M. **El uso de la voz en la publicidad audiovisual dirigida a los niños y su eficacia persuasiva** (Tesis Doctoral). Universidad Autónoma de Barcelona, 1999.

TERKEN. In: HORNE, M. (Ed.). **Prosody**: theory and experiment. Studies presented to Gösta Bruce. Dordrecht: Kluwer Academic Publishers, 2003.

MAEKAWA et al. In: HORNE, M. (Ed.). **Prosody**: theory and experiment. Studies presented to Gösta Bruce. Dordrecht: Kluwer Academic Publishers, 2000.

Xu, Y. **The penta modelo d speech melody**: transmitting multiple communicative functions in parallel. [S.l.]: Sound to Sence, at MIT, jun. 2004. p. 11-13.

BAQUÉ, L.; ESTRUCH, M. Modelo de Aix-en-Provence. In: PRIETO, P. (Ed.). **Teorías de la entonación**. Alfabeto Intsint, Universidad de Aix en Provence. Barcelona: Ariel (Ariel Lingüística), 2003. p. 123-154.

ESPESSER, R.; HIRST, D. Automatic modelling of fundamental frequency using a quadratic spline function. IPA, **Travaux de l´Institut de Phonétique d´Aix,** 15, 75-8. 1993.