Subjetividad y Procesos Cognitivos, Vol. 14, Nº 2, 2010
Pág. 79-88, ISSN impreso: 1666-244X, ISSN electrónico: 1852-7310
UCES

Modelado de la mimetización entre interlocutores para mejorar la naturalidad de sistemas de diálogo hablado

Mimicking the interlocutor's speech: models of speaker entrainment for improving the naturalness of spoken dialogue systems

Agustín Gravano*

Resumen

A medida que las tecnologías de procesamiento del habla continúan mejorando, gradualmente nos acercamos al viejo sueño de crear una máquina que hable. Los actuales sistemas interactivos de diálogo hablado permiten que los usuarios realicen tareas simples, tales como transacciones bancarias y reservas en hoteles, mediante la interacción verbal. Pese a ser relativamente exitosas, estas conversaciones humano-computadora aún tienen un largo camino para recorrer en cuanto a su naturalidad: estos sistemas tienden a ser descriptos por los usuarios como "extraños" o incluso "intimidantes". Entre las razones principales para esta falta de naturalidad, figura el modelado imperfecto de la variación prosódica, o cómo algunas propiedades del habla (tales como la entonación, la intensidad o el ritmo) cambian en las expresiones verbales. Los sistemas actuales todavía son incapaces de manejar estas características en forma correcta, tanto al entender el habla del usuario como para producir respuestas sintetizadas. La variación prosódica es extremadamente compleja en el habla espontánea, y se sabe que la afectan varios niveles de representación lingüística (léxica, sintáctica, semántica y pragmática). En el presente artículo, enfocamos nuestra atención en una dimensión particular de variación prosódica, conocida como "mimetización entre interlocutores", que consiste en la alineación automática de características del habla entre los participantes de un diálogo. Tras un repaso general de la literatura de estos temas, describimos un proyecto de investigación en curso que busca modelar la mimetización prosódica en diálogos.

Palabras clave: procesamiento del habla, diálogo, prosodia, alineamiento.

* Departamento de Computación, FCEyN, Universidad de Buenos Aires. Pabellón I, Ciudad Universitaria. (C1428EGA) Buenos Aires, Argentina. E-mail: gravano@dc.uba.ar

UCES Subjetividad y Procesos Cognitivos, Vol. 14, Nº 2, 2010
Pág. 79-88, ISSN impreso: 1666-244X, ISSN electrónico: 1852-7310

Summary
As speech processing technologies continue to improve, the old dream of creating a machine that talks gradually becomes real. The present interactive speech systems enable users to perform simple tasks such as banking transactions and hotel reservations, through verbal interaction. Despite being relatively successful, these human-computer conversations still have a long way to go regarding their naturalness: these systems tend to be described as "odd" or even "intimidating" by users. Among the main reasons for this lack of naturalness, is the flawed modeling of prosodic variation or the way some properties of speech (such as intonation, intensity and rhythm) change in verbal expressions. Current systems are still unable to handle these features correctly, both to understand the speech of the user as to produce synthesized responses. Prosodic variation is extremely complex in spontaneous speech, and it is well known that it´s affected by several levels of linguistic representation (lexical, syntactic, semantic and pragmatic). The present article focuses on a specific dimension of prosodic variation, known as "mimetization between interlocutors", which consists in the automatic alignment of speech features between the participants of a dialogue. After a general overview of the literature on these subjects, a research project in process that seeks to model the prosodic mimetizatin in dialogues is described.

Key words: speech processing, dialogue, prosody, alignment.

1. Introducción

El objetivo primario del área del procesamiento del habla es la construcción de sistemas informáticos capaces de manipular de manera efectiva el lenguaje hablado. Tal capacidad puede parecer trivial para los seres humanos, puesto que está inherentemente incorporada en nosotros. A diario empleamos el lenguaje oral casi sin percatarnos de la cantidad y la complejidad de los procesos involucrados en algo tan natural como mantener una conversación. Sin embargo, muchos de dichos procesos plantean tremendas dificultades para los sistemas informáticos. En consecuencia, tras unos cincuenta años de investigación en estos temas, todavía estamos relativamente lejos de alcanzar el objetivo mencionado; especialmente, si se toma como referencia a otras disciplinas informáticas, como por ejemplo las comunicaciones o la computación gráfica, las cuales han mostrado avances extraordinarios en las últimas décadas.

1.1. Tecnologías de procesamiento del habla

La disponibilidad de sistemas de procesamiento del habla engendrará un sinnúmero de nuevas tecnologías, hoy en día pertenecientes aún al mundo de la ciencia ficción. Por ejemplo, será posible realizar búsquedas de palabras y frases en bases de datos de contenidos audiovisuales (e.g. *YouTube*, o archivos de audiciones de radios). Las reuniones de trabajo podrán resumirse automáticamente, resaltando los temas abarcados y las decisiones tomadas durante las mismas. También se podrá traducir el habla de una persona de un idioma a otro en forma simultánea, rompiendo persistentes barreras

sociales entre miembros de distintas culturas. Quizá la aplicación más obvia sean las interfaces de usuario: la activación mediante la voz de funcionalidades de computadoras, automóviles y teléfonos celulares, reemplazando interfaces artificiales como teclados y tableros por el medio elemental de comunicación entre humanos. Las tecnologías del habla también abrirán puertas a personas con capacidades especiales: por ejemplo, en la actualidad ya hay sistemas de síntesis del habla empleados por pacientes con impedimentos vocales (el caso más conocido es el del físico británico Stephen Hawking), y sistemas de lectura de textos para personas con visión disminuida. Hoy en día, ya existen sistemas comerciales que implementan algunas de las ideas mencionadas en este párrafo, aunque funcionan todavía en dominios limitados y con una eficacia acotada. En definitiva, son innumerables las oportunidades que aparecerán a partir del desarrollo de las tecnologías de procesamiento del habla; la lista de posibles aplicaciones podría llenar muchas páginas.

Tal vez la aplicación más ambiciosa en esta área sean los *sistemas de diálogo hablado*, también conocidos como *agentes conversacionales*. Idealmente, tales sistemas pueden interactuar con seres humanos mediante el lenguaje hablado; es decir, son capaces de entablar un diálogo con una persona. La construcción de tales sistemas plantea varios problemas de gran dificultad; una única iteración de su ejecución puede resumirse de la siguiente forma: i) cada secuencia de sonidos emitida por el usuario es depurada de ruidos y decodificada a una secuencia de palabras; ii) se extrae el mensaje contenido en dichas palabras; iii) el sistema elige un mensaje para responder al usuario; iv) ese mensaje se transforma a una secuencia de palabras; v) las palabras son finalmente sintetizadas en una secuencia de sonidos para mostrar al usuario. Cuando el usuario vuelve a hablar, comienza una nueva iteración. En la actualidad ya existen sistemas experimentales que permiten al usuario realizar tareas específicas de limitada complejidad, tales como realizar consultas de horarios de autobuses, o diagnosticar y reparar problemas en el sistema de televisión por cable. Sin embargo, la usabilidad de los mismos es todavía muy reducida, sobre todo por deficiencias en el primer paso (filtrado de ruidos y reconocimiento de las palabras), pero también por la total falta de naturalidad en la interacción entre el humano y la computadora. Buena parte de tal falta de naturalidad puede atribuirse a la carencia de modelos adecuados de variación prosódica en el habla espontánea.

1.2. Variación prosódica

El lenguaje oral se diferencia de muchas maneras del lenguaje escrito. Para empezar, no posee puntuación ni estructura explícita: oraciones, párrafos y secciones no están claramente delimitados. Además, el habla espontánea suele estar plagada de disfluencias, incluyendo pausas llenas (*"eh"*, *"esteh"*), autocorrecciones (*"mañana a las cuat- cinco"*) y errores de dicción (*"tres trigres"*), así como de construcciones agramaticales (*"yo fui el único que volví*", "*esa fue la mejor vez que la pasé"*).

Sin embargo, quizá la característica más distintiva del lenguaje oral sea la *prosodia*: la manifestación acústica de las palabras. Coloquialmente, la prosodia permite distinguir entre *qué* se dice, y *cómo* se lo dice. Posee varias componentes, entre ellas: i) la *entonación*, determinada por las variaciones en el *nivel tonal* de las sucesivas sílabas que conforman una frase; ii) la *intensidad*, informalmente conocida como *volumen*; iii) la *duración segmental*, que determina la velocidad y el ritmo del habla; y iv) la *calidad de la voz*, que describe cualidades de la producción del habla independientes de las tres anteriores, como ser la aspereza, el susurro, la voz tensa y la voz chirriante, entre otras.

Cotidianamente manipulamos la prosodia de nuestra habla para un sinnúmero de funciones: para expresar énfasis, para hacer preguntas, para contestarlas, para estructurar el discurso, para coordinar conversaciones, para expresar emociones, etc. Salvo en contadas excepciones, estas funciones del lenguaje oral escapan al alcance de los sistemas actuales, diseñados para procesar casi exclusivamente *qué* se dice, pero no *cómo*. En consecuencia, oraciones como *"María no renunció por el sueldo que cobraba"*, ambiguas en el plano escrito y desambiguadas mediante la prosodia[1], corren serio riesgo de ser malinterpretadas por los sistemas actuales. Asimismo, la ausencia de un manejo adecuado de la prosodia conduce a que el habla artificial sea descripta como "extraña" o "mecánica" por los usuarios, quienes al escuchar largos pasajes se desconcentran, se cansan y pierden buena parte de la información, incluso cuando la voz resulta perfectamente inteligible.

En consecuencia, resulta de vital importancia para las aplicaciones descriptas en la sección anterior contar con modelos que expliquen efectivamente las distintas dimensiones de variación prosódica presente en el habla. Por un lado, permitirían perfeccionar la interpretación de expresiones producidas oralmente por el usuario; por otro lado, incrementarían la naturalidad del habla sintetizada por la computadora. El proyecto descripto en el presente artículo estudia una dimensión de la variación prosódica que -se postula- ocurre en conversaciones entre dos personas: la *mimetización* de ciertos rasgos prosódicos entre los interlocutores.

1.3. Mimetización

En la literatura de Psicología del Comportamiento se ha observado con frecuencia que, bajo ciertas condiciones, cuando una persona mantiene una conversación, modifica su manera de actuar, aproximándola a la de su interlocutor. En una reseña de este tema, Chartrand & Bargh (1999) describen a este fenómeno como una "imitación no conciente de posturas, maneras, expresiones faciales y otros comportamientos del compañero interaccional" [p. 893][2], y conjeturan que es más fuerte en individuos

[1] Dependiendo de cómo se diga esta frase, puede entenderse que María renunció o que no renunció.

[2] *"Nonconscious mimicry of the postures, mannerisms, facial expressions, and other behaviors of one's interaction partners"*.

con empatía disposicional. En otras palabras, personas con predisposición a buscar la aceptación social modifican su comportamiento en forma más marcada para aproximarlo a sus interlocutores.

Esta modificación del comportamiento ha sido observada también en la manera de hablar. Por ejemplo, los interlocutores adoptan las mismas formas léxicas para referirse a las cosas, negociando tácitamente descripciones compartidas, en especial para cosas que resulten poco familiares (Garrod & Anderson, 1987; Isaacs & Clark, 1987; Brennan, 1996; entre otros). Estudios más recientes sugieren que esto también es cierto para el uso de estructuras sintácticas (Reitter et al., 2006) y para ciertos parámetros prosódicos como el ritmo y la intensidad (Coulston et al., 2002; Ward & Litman, 2007). Este fenómeno subconsciente es conocido como *mimetización, alineamiento, adaptación o convergencia*, y también con el término inglés *entrainment*, y se ha mostrado que juega un rol importante en la coordinación de diálogos, facilitando tanto la producción como la comprensión del habla en los seres humanos (Pickering & Garrod, 2004; Goleman, 2006; Nenkova et al., 2008).

2. Proyecto en curso

El presente proyecto está siendo realizado en colaboración con la Prof. Julia Hirschberg[3], la Prof. Ani Nenkova[4] y el Prof. Štefan Beňuš[5]. Consiste en estudiar el alineamiento entre participantes de diálogos, con respecto a varios parámetros prosódicos, incluyendo el ritmo, la intensidad, los patrones de acentuación y los contornos de entonación.

2.1. Hipótesis a evaluar

En concreto, las siguientes son las principales preguntas que este proyecto intenta responder:

1. ¿Existen otros tipos de alineamiento prosódico, además del postulado en trabajos previos para el ritmo y la intensidad? Las variables de estudio incluyen: el nivel tonal, el nivel de intensidad, el uso de contornos de entonación, la velocidad del habla, el empleo de pausas llenas, y la aparición de silencios y superposiciones entre los hablantes.

2. ¿Cómo y cuándo tiene lugar el alineamiento prosódico? ¿Varía según propiedades de la conversación, como por ejemplo el tópico de la misma?

[3] Department of Computer Science, Columbia University, Nueva York, EE.UU.

[4] Department of Computer and Information Science, University of Pennsylvania, Filadelfia, EE.UU.

[5] Department of English and American Studies, Constantine the Philosopher University, Nitra, Eslovaquia.

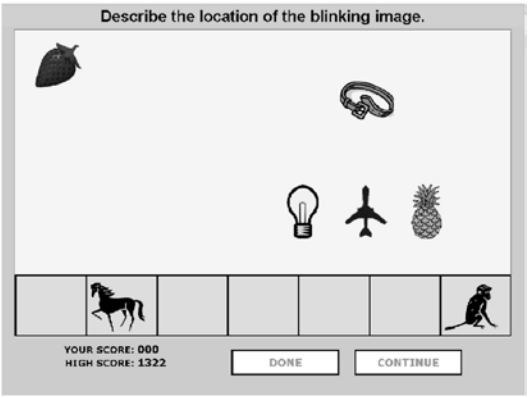
3. ¿Qué efecto tiene en las características de la conversación la existencia de alineamiento prosódico? Por ejemplo, ¿mejora la coordinación entre los hablantes, o la efectividad de la comunicación?

Nuestro objetivo primario consiste en desarrollar un modelo computacional de alineamiento prosódico, potencialmente útil para mejorar la naturalidad de los actuales sistemas de diálogo hablado. Dado que se espera que dichos sistemas puedan interactuar con los usuarios de manera coordinada para ejecutar tareas eficientemente, la respuesta a estas preguntas resulta de inmenso interés.

2.2. Materiales y métodos

Los materiales estudiados en este proyecto provienen de dos cuerpos de datos de diálogo espontáneo en inglés norteamericano: *el Columbia Games Corpus* y *el Switchboard Corpus*. El primero de ellos consiste en doce conversaciones diádicas (i.e., con dos participantes) entre trece

Figura 1. Juego del Columbia Games Corpus



personas distintas. En cada sesión, se sentó a dos participantes (quienes no se conocían previamente) en una cabina profesional de grabación, cara a cara a ambos lados de una mesa, y con una cortina opaca colgando entre ellos para evitar la comunicación visual. Los participantes contaron con sendas computadoras portátiles conectadas entre sí, en las cuales jugaron una serie de juegos simples que requerían de comunicación verbal. Por ejemplo, en uno de tales juegos, ambas computadoras muestran un tablero con varios objetos (Figura 1), todos en la misma posición excepto por uno, el *objetivo*, que aparece en un lugar distinto en cada computadora. Uno de los jugadores, para quien el objetivo aparece titilando, debe entonces describir la ubicación exacta del mismo usando los otros elementos como referencia, de modo que el otro jugador pueda mover su propia

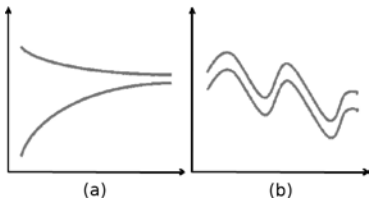
instancia del objetivo a la posición correcta; al terminar cada juego, se otorga un puntaje según la precisión de la tarea realizada. Las grabaciones se hicieron en 44 kHz, 16 bits, con un canal separado para cada hablante; luego fueron guardadas en 16 kHz para el presente estudio. Cada sesión duró aproximadamente 45 minutos, totalizando 9 horas de diálogos, 70.259 palabras (2.037 únicas) para todo el cuerpo de datos.

El *Switchboard Corpus* consiste en 2.430 diálogos telefónicos espontáneos de características muy diferentes al corpus anterior. Tienen menor duración (6 minutos en promedio) y una calidad de audio sustancialmente inferior (grabaciones en 8 kHz con ruido de línea y de fondo). Además, no existió restricción alguna al tópico de conversación y los participantes podían conocerse de antemano, hechos que abren las puertas a complejas interacciones sociales que escapan al alcance de este estudio. Sin embargo, gracias al elevado número de diálogos disponibles en este corpus, aspiramos a detectar patrones estadísticamente significantes que trasciendan los problemas que puedan surgir de dicha heterogeneidad.

Ambos cuerpos de datos cuentan con transcripciones textuales alineadas temporalmente a la señal de audio, realizadas por personal especialmente entrenado. Todas las variables acústico-prosódicas analizadas en este estudio fueron extraídas en forma automática de la señal de audio y de las transcripciones alineadas, empleando la herramienta de libre disponibilidad *Praat* (Boersma & Weenink, 2001).

El presente proyecto involucra la utilización de técnicas estadísticas y de aprendizaje automático (en inglés, *machine learning*) para encontrar modelos descriptivos de alineamiento, tanto a nivel global (considerando toda una conversación como una unidad, en la cual el alineamiento ocurre o no ocurre), como a nivel local (considerando la evolución dinámica del alineamiento, a medida que transcurre la conversación). Consideramos además dos tipos posibles de alineamiento, ilustrados esquemáticamente en la Figura 2: *convergencia* y *sincronía*, usando la terminología propuesta por Edlund et al. (2009). En la convergencia, los dos hablantes comienzan con valores disímiles en una

Figura 2. Convergencia (izq.) y sincronía



variable determinada (p. ej., intensidad); a medida que la conversación avanza, dichos valores van aproximándose entre sí. En la sincronía, los dos hablantes poseen valores

posiblemente distintos para una variable, pero esta fluctúa de manera similar para ambos a lo largo de la conversación.

2.3. Resultados preliminares

Los resultados obtenidos hasta el momento en este proyecto resultan confusos y contradictorios. Para cada una de las distintas variables de estudio, cada diálogo se encuadra en uno de los siguientes tres escenarios:

1. existencia de alineamiento entre los hablantes;
2. existencia de ***des***alineamiento entre los hablantes (es decir, el fenómeno opuesto al postulado, con un distanciamiento gradual de las características prosódicas de los hablantes, en lugar de un acercamiento);
3. ninguno de los escenarios anteriores.

Es importante resaltar que estos resultados claramente difieren de una distribución aleatoria. La evidencia de la existencia de alineamiento y desalineamiento es en todos los casos pronunciada y estadísticamente significativa, lo cual descarta que los resultados observados puedan ser explicados como una consecuencia del azar. En cambio, parece probable que existan factores -aun no incluidos en nuestro estudio- que determinen cuál de los tres escenarios tendrá lugar en cada conversación.

En este contexto, la clave puede yacer en la observación realizada por Chartrand & Bargh (1999): que la mimetización tiene una presencia más fuerte en individuos con empatía disposicional. A partir de esta conjetura, surge la pregunta de si los factores faltantes en nuestro análisis están relacionados a la personalidad de los participantes y a la empatía interpersonal entre ellos. En otras palabras, lo que deseamos saber es: ¿cómo afectan a la aparición de alineamiento prosódico los rasgos personales e interpersonales de los hablantes?

2.4. Trabajo futuro

Para buscar la respuesta a este interrogante, es necesario incorporar nueva información a los diálogos del cuerpo de datos; al momento de la presentación de este trabajo, nos encontramos evaluando diferentes sistemas de rotulado manual. En primer lugar, estamos considerando diferentes métodos de determinación del tipo de personalidad (por ejemplo, dependiente, antisocial, narcisista, etc.). En segundo lugar, estamos diseñando una serie de preguntas que permitan medir rasgos relacionados al nivel de empatía interpersonal en una conversación (por ejemplo, ¿considera que al hablante A le importaba ser aceptado por el hablante B?).

Idealmente, este tipo de información debería haber sido recolectada con cuestionarios completados por los mismos participantes inmediatamente antes y después de cada sesión. Dado que esto no ocurrió así (los datos fueron recolectados para otros proyectos), ahora solo resta apelar a anotadores que rotulen los datos como observadores externos, basándose en sus percepciones. Asimismo, las tareas de rotulado serán

objeto de análisis en sí mismas: la concordancia entre anotadores será evaluada para cada tarea mediante el índice Kappa. Solamente incorporaremos a nuestro estudio principal aquellos datos que muestren una concordancia razonablemente alta.

Una vez que contemos con estos nuevos rotulados, podremos enriquecer nuestro estudio de alineamiento prosódico. Como ya se explicó, analizaremos si las variables relacionadas a los rasgos personales e interpersonales tienen un rol activo en la presencia o ausencia de alineamiento de variables prosódicas de los hablantes.

3. Conclusiones

El objetivo del presente proyecto es modelar el (hipotético) alineamiento de algunas variables prosódicas entre los participantes de un diálogo. Tales variables incluyen el nivel tonal, el nivel de intensidad, el uso de contornos de entonación, y la velocidad del habla, entre otras. Las claras contradicciones en nuestros resultados preliminares sugieren fuertemente la existencia de factores ignorados hasta ahora, tales como los rasgos personales e interpersonales de los participantes de la conversación. La etapa actual del proyecto consiste en explorar estas nuevas dimensiones del problema.

La eventual disponibilidad de un modelo de alineamiento prosódico servirá para mejorar la calidad de diversas aplicaciones de procesamiento del habla, tanto para perfeccionar la interpretación de expresiones producidas oralmente por personas, como para incrementar la naturalidad del habla artificial. En particular, los sistemas de diálogo hablado se verán beneficiados por este conocimiento, puesto que para dichos sistemas resultan críticas la naturalidad y la coordinación de las interacciones.

Reconocimientos

Este proyecto es financiado parcialmente por el subsidio NSF IIS-Robust Intelligence 0803148, de la *National Science Foundation* de EE.UU.

Bibliografía

Boersma, P. y Weenink, D. (2001). Praat: Doing phonetics by computer. http://www.praat.org.

Brennan, S.E. (1996). Lexical entrainment in spontaneous dialog. *Int'l Symposium on Spoken Dialogue*, pp. 41-44.

Chartrand, T.L.; Bargh, J.A. (1999). The chameleon effect: the perception-behavior link and social interaction. *Journal of Personality and Social Psychology*, 76(6), pp. 893-910.

Coulston, R.; Oviatt, S. y Darves, C. (2002). Amplitude convergence in children's conversational speech with animated personas. *ICSLP,* pp. 2689-2692.

Edlund, J.; Heldner, M. y Hirschberg, J. (2009). Pause and gap length in face-to-face interaction. *Interspeech*, pp. 2779-82.

Garrod, S. y Anderson, A. (1987). Saying what you mean in dialogue: A study in conceptual and semantic co-ordination. *Cognition*, 27(2), pp. 181-218.

Goleman, D. (2006). *Social Intelligence: The New Science of Human Relationships*. Bantam, New York.

Isaacs, E. y Clark, H. (1987). References in conversation between experts and novices. *Journal of Experimental Psychology*, 116(1), pp. 26-37.

Nenkova, A.; Gravano, A. y Hirschberg, J. (2008). High frequency word entrainment in spoken dialogue. *ACL/HLT*, pp. 169-172.

Pickering, M.J. y Garrod, S. (2004). Toward a mechanistic psychology of dialogue. *Behavioral and Brain Sciences*, 27, pp. 169-226.

Reitter, D.; Keller, F. y Moore, J.D. (2006). Computational modelling of structural priming in dialogue. *HLT/NAACL*, pp.121-124.

Ward, A. y Litman, D. (2007). Automatically measuring lexical and acoustic/prosodic convergence in tutorial dialog corpora. *SLaTE Workshop on Speech and Language Technology in Education*.

Fecha de recepción: 15/12/09
Fecha de aceptación: 10/05/10