Métricas de mimetización acústico-prosódica en hablantes y su relación con rasgos sociales de diálogos

Juan Manuel Pérez

Springer-Verlag, Computer Science Editorial,
Tiergartenstr. 17, 69121 Heidelberg, Germany
{alfred.hofmann,ursula.barth,ingrid.haas,frank.holzwarth,
anna.kramer,leonie.kunz,christine.reiss,nicole.sator,
erika.siebert-cole,peter.strasser,lncs}@springer.com
http://www.springer.com/lncs

Resumen Keywords: Procesamiento del Habla, Series de Tiempo, Entrainment, Regresión Linea

1. Introduction

Los sistemas de diálogo humano-computadora son cada vez más frecuentes, y sus aplicaciones comprenden una amplia gama de rubros: desde aplicaciones móviles, motores de búsqueda, juegos o tecnologías de asistencia para ancianos y discapacitados. Si bien es cierto que estos sistemas logran captar la dimensión lingüística de la comunicación humana, tienen un déficit importante a la hora de procesar y transmitir el aspecto superestructural de la comunicación oral, que radica en el intercambio de afecto, emociones, actitudes y otras intenciones de los participantes. Este problema puede verse en cualquier sistema que interactúe sintetizando lenguaje humano: por ejemplo, las aplicaciones telefónicas que atienden automáticamente a sus clientes [PH05,RBL$^+$06]. Stanley Kubrick y Arthur C. Clarke predijeron esto a la perfección, cuando en "2001: Una Odisea en el Espacio"(1968) dotaron a *HAL* de una voz monótona y robótica, casi lobotomizada. Otro problema grave que sufren estos sistemas humano-computadora es que asumen que sus interacciones de "a turnos", cuando las conversaciones entre humanos suelen distar bastante de ese modelo.

Dentro de las cualidades del lenguaje oral, una de las más distintivas es la *prosodia*, qué es la dimensión que capta *cómo* se dicen las cosas, en contraposición a *qué* se está manifestando. Posee varias componentes acústico-prosódicas: por ejemplo, el tono o pitch, la intensidad o volumen, la calidad de la voz, la velocidad del habla y otras. Un manejo adecuado de estas componentes es lo que, hoy día, distingue una voz humana de una artificial. Esta carencia de habilidad sobre la prosodia conlleva cierta dificultad en la interacción con agentes conversaciones, que suelen ser calificados como "mecánicos" o "extraños" en su forma de comunicarse. [RBL$^+$06,WRWN05]

En pos de mejorar el entendimiento entre agentes conversacionales y sus usuarios, resulta de vital importancia poder entender y modelar las variaciones prosódicas de la comunicación oral. Esto se traduciría tanto en una mejor apreciación de lo que quiere comunicar el usuario, como en una mayor naturalidad de la voz sintetizada por el agente.

1.1. Mimetización

En la literatura de Psicología del Comportamiento se ha observado con frecuencia que, bajo ciertas condiciones, cuando una persona mantiene una conversación, ésta modifica su manera de actuar aproximándola a la de su interlocutor. En una reseña de este tema se describe a este fenómeno como una "imitación no consciente de posturas, maneras, expresiones faciales y otros comportamientos del compañero interaccional" [CB99, p. 893] y conjeturan que es más fuerte en individuos 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 [Bre96]. Estudios más recientes sugieren que esto también es cierto para el uso de estructuras sintácticas [RKM06]. Este fenómeno subconsciente es conocido como mimetización, alineamiento, adaptación o convergencia y también con el término inglés *entrainment*. 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[NGH08,GBLH15]. En nuestro caso, nos interesa principalmente el *entrainment* de la prosodia.

1.2. Midiendo la mimetización

Muchos estudios han examinado la mimetización prosódica, listados en [DLSVC14]. Un número importante de ellos se han basado en la premisa de la mimetización como un fenómeno lineal, en el cual la convergencia "va sucediendo" a lo largo de la conversación [BSD95]. Estos estudios dividen las conversaciones en varias partes, y verifican que la diferencia absoluta entre los valores medios (de las variables acústico-prosódicas) y sus desviaciones se aproxime en las últimas partes de la interacción. Sin embargo, este enfoque de la mimetización niega su faceta dinámica: los interlocutores pueden estar inactivos y luego hablar, pueden pasar por varias etapas como escuchar, pensar, discutir un punto, etc. En [LH11] se reportó que éste es un fenómeno no sólamente lineal, sino también dinámico, donde los interlocutores van coincidiendo en el análisis por turnos.

Un problema común que surge a la hora de calcular estas métricas es el hecho de que las conversaciones no están alineadas en el tiempo, ni se dan en turnos de duración constante. Nos preguntamos entonces qué partes del diálogo de un hablante deberían compararse con qué otras partes de su par. Un enfoque de

comparar interlocuciones uno a uno es demasiado simple y no captura situaciones de diálogo reales, mucho más dinámicas y con solapamiento casi constante.

Para atacar estos inconvenientes, utilizamos el método *TAMA* (Time Aligned Moving Average) [KDW+08], que consiste en separar en ventanas de tiempo el diálogo, y promediar los valores de las variables prosódicas dentro de cada una. Este método es muy similar a aplicar un filtro de Promedio Móvil (Moving Average), lo que da el nombre a la técnica. Al separar el diálogo en ventanas de tiempo, podemos construir dos series de tiempo en base a cada interlocutor. Estas abstracciones son mucho más tratables que tener una secuencia de elocuciones de parte de cada hablante, y nos permiten efectuar análisis bien conocidos, uno de los cuáles nos permite construir una medida del *entrainment*.

1.3. Objetivo del estudio

En el presente estudio, aplicamos la técnica de *TAMA* para definir dos métricas de *entrainment*. Utilizamos un corpus de diálogo entre dos participantes angloparlantes, quienes interactúan mediante un juego a través de computadoras. El corpus ha sido anotado manualmente con variables que describen la percepción social de la conversación; por ejemplo: ¿el sujeto parece comprometido con el juego? ¿al sujeto no le agrada su compañero?

Luego, veremos si existe, para cada una de las variables acústico-prosódicas, alguna relación significante entre las métricas definidas y las percepciones sociales sobre las conversaciones. Uno esperaría que valores altos de nuestras métricas del *entrainment* se relacionen con valores altos de variables sociales positivas, tales como mostrarse colaborativo o compenetrado en la tarea.

2. Materiales y Método

2.1. Columbia Game Corpus

Empleamos el Columbia Game Corpus [Gra09] consiste en doce conversaciones diádicas (i.e., con dos participantes) entre trece personas angloparlantes distintas. Todos los participantes reportaron hablar inglés americano estándar, y no tener problemas de audición. La edad de los participantes se encuentra en el rango de los 20 a 50 años.

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. Todas las conversaciones cuentan con transcripciones textuales alineadas temporalmente a la señal de audio, realizadas por personal especialmente entrenado.

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

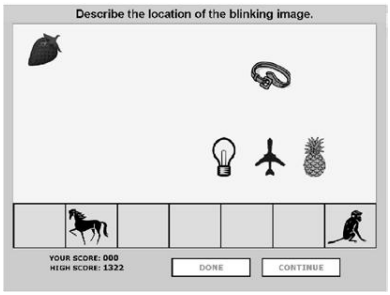

Figura 1. Juego de objetos del Columbia Games

conectadas entre sí, en las cuales jugaron una serie de juegos simples que requerían de comunicación verbal. El primero de ellos es un juego de cartas que no consideramos en el presente estudio por tratarse esencialmente de monólogos o diálogos con poca interacción. Luego de esto, pasaron al juego que analizamos, denominado 'juego de objetos'.

En el juego de objetos, la pantalla de cada jugador mostró un tablero con varios objetos, entre 5 y 7, como se ve en la Figura 1. Para uno de los jugadores (el Descriptor) el objeto *Objetivo* aparecía en una posición aleatoria entre otros objetos. Para el otro jugador, a quien llamaremos el Seguidor, el objetivo aparecía en la parte baja de la pantalla. Entonces, al Descriptor se le encargaba describir la posición del Objetivo de manera que el Seguidor pudiera mover su representación del objeto a la misma posición en su pantalla. Luego de una negociación entre ambos jugadores para decidir la mejor posición del objeto, se les asignó a los jugadores una puntuación entre 1 y 100 puntos de acuerdo a qué tan acertado fue el posicionamiento del objetivo por parte del Seguidor.

Cada sesión consistió de 14 tareas como ésta, cambiando los objetos y sus ubicaciones. En las primeras cuatro tareas, uno de los sujetos tomó el papel del Descriptor; en los siguientes cuatro invirtieron roles, y en las finales seis fueron alternando los roles de Descriptor y Seguidor.

2.1.1. Anotaciones sobre comportamiento social Varios aspectos del comportamiento de los jugadores durante los juegos de objetos fueron anotados mediante la herramienta de crowdsourcing *Amazon Mechanical Turk*[1]. Cada anotador escuchó el audio correspondiente a una tarea del juego y tuvo que responder a varias preguntas sobre cada uno de los sujetos, entre las que se encuentran:

- ¿el sujeto contribuye para el éxito del equipo? (*contributes-to-completion*)
- ¿el sujeto parece comprometido con el juego? (*engaged-with-game*)
- ¿el sujeto se expresa correctamente? (*making-self-clear*)
- ¿el sujeto piensa lo que va a decir? (*planning-what-to-say*)

[1] https://www.mturk.com

- ¿el sujeto alienta a su compañero? (*gives-encouragement*)
- ¿el sujeto le hace difícil hablar a su compañero? (*difficult-for-partner-to-speak*)
- ¿el sujeto está aburrido con el juego? (*bored-with-game*)
- ¿al sujeto no le agrada su compañero? (*dislikes-partner*)

Cada uno de estos audios fue puntuado por cinco anotadores, que respondieron por sí o por no para cada una de las preguntas. El puntaje que recibe cada una de las preguntas (a las cuales llamaremos a partir de ahora *variables sociales*) consiste en la cantidad de respuestas afirmativas que recibió, teniendo un rango de 0 a 5. Por ejemplo, una tarea dada podría tener puntaje 3 para la variable social 'el sujeto A se expresa correctamente' o puntaje 5 para la variable 'el sujeto B dirige la conversación'.

2.2. Extracción de variables acústico/prosódicas

La herramienta *Praat* [2] fue utilizada para extraer automáticamente las variables acústico-prosódicas del corpus. Las variables que medimos fueron el tono, la intensidad, la proporción de vocalizaciones, jitter, shimmer, cantidad de sílabas, cantidad de fonemas, y la proporción de ruido sobre armónicos. Estos atributos fueron medidos en cada uno de los segmentos de habla del corpus.

En la siguiente tabla resumimos estas features. Recordemos nuevamente que estas features son medidas en un intervalo de tiempo.

Variable	Descripción
F0 Mean	Valor medio de la frecuencia fundamental
F0 Max	Valor máximo de la frecuencia fundamental
Int Mean	Valor medio de la intensidad
Int Max	Valor máximo de la intensidad
NHR	Noise-to-harmonics ratio
Shimmer	Shimmer medido
Jitter	Jitter medido
Sílabas/seg	Cantidad de sílabas por segundo
Fonemas/seg	Cantidad de fonemas por segundo

3. Descripción del método TAMA

En [KDW+08] se introdujo un método novedoso para el análisis del *entrainment* acústico-prosódico. Esta técnica consiste, a grandes rasgos, en armar dos series de tiempo para cada uno de los interlocutores y luego utilizar herramientas de análisis sobre las series construídas. Una serie de tiempo, en términos

[2] http://www.fon.hum.uva.nl/praat/

coloquiales, es una colección cronológica de observaciones, como pueden ser los valores de las acciones de una empresa a lo largo del tiempo, o la cantidad de lluvia medida en milímetros para cada mes de cierto año.

Uno de los problemas que resuelve esta técnica es el del alineamiento: si intentásemos comparar cada segmento del habla (*elocución* o *utterance*) con otros, ¿cómo los alinearíamos? Una posibilidad sería alinear cada segmento de un hablante con el próximo de su interlocutor. Esto, sin embargo, es muy simplista y poco representativo de la realidad ya que los diálogos entre humanos no suelen darse en ese formato. Al introducir el concepto de series de tiempo, podemos olvidarnos de los segmentos del habla y simplemente utilizar estas construcciones.

Para construir la serie de tiempo de cada hablante debemos, en primer lugar, dividir el diálogo en ventanas solapadas de igual tamaño. A la diferencia entre ventana y ventana llamaremos *frame step*, y al tamaño de ventana *frame length*. Consideraremos sólo los segmentos de habla que se encuentren dentro de cada ventana; aquellos que atraviesen los límites de las ventanas serán cortados para que se mantengan dentro de éste. En la Figura 2 se ilustra el proceso: las líneas punteadas marcan los límites de la ventana, los intervalos coloreados en negro los segmentos de habla, y en gris los segmentos cortados.

Como producto de esto, nuestro corpus queda dividido en una sucesión ventanas solapadas. En el trabajo original [KDW+08], se usa un step de 10 segundos, y un tamaño de ventana de 20 segundos, dando como resultado un solapamiento del 50 %.

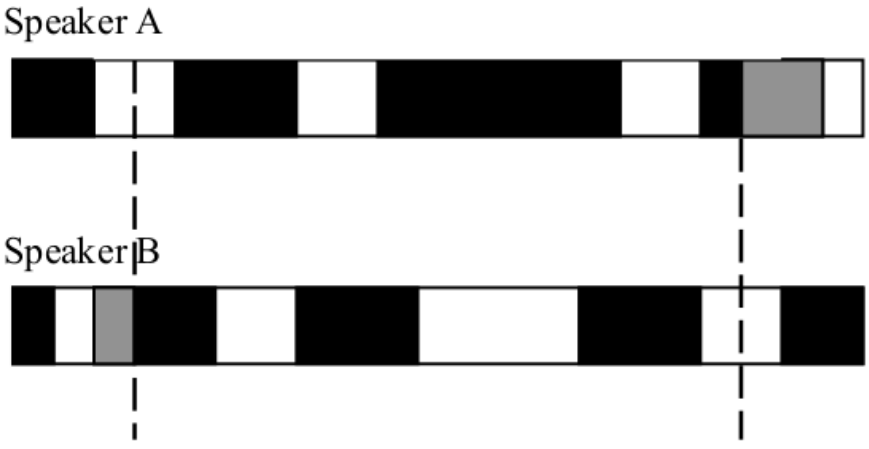

Figura 2. Gráfico de la separación del diálogo en ventanas (Fuente [KDMC08])

Una vez que la conversación se ha partido en ventanas mediante el proceso descripto, se calculan los valores de la serie de tiempo para cada hablante y cada variable acustico-prosódicas (por ejemplo el pitch) mediante el siguiente cálculo:

$$\mu = \sum_{i=1}^{N} f_i d'_i \tag{1}$$

donde i itera sobre las elocuciones dentro del *frame*, d'_i es la duración relativa del segmento (respecto del tiempo total hablado en toda la ventana) y f_i es el valor de la *variable acústico-prosódica* que estamos midiendo. d'_i se calcula con la fórmula

$$d'_i = \frac{d_i}{\sum_{i=1}^{N} d_i} \tag{2}$$

donde d_i es la longitud en segundos de los segmentos del habla en el frame.

Como se ve en la ecuación 1, el valor que calculamos es una media ponderada del valor de la variable por la duración de las elocuciones. Así, por ejemplo, al calcular una serie de tiempo sobre la intensidad, la contribución de interjecciones (*ah!* por ejemplo), que suelen tener altos valores, estará atenuada por sus breves duraciones.

Dada una variable acústico-prosódico y una conversación, una vez obtenidas dos series de tiempo mediante el cálculo ventana a ventana de la ecuación 1, necesitamos efectuar algún tipo de análisis sobre éstas para obtener una medida del *entrainment*.

3.1. Análisis bivariado

En [KDMC08] se continúa el trabajo en series de tiempo, y se efectúan análisis tanto para cada serie por separado como para las dos en conjunto, lo cual se llama "análisis bivariado" en la terminología de series de tiempo. En este análisis pretendemos analizar ambas series como parte de un sistema y ver cómo se influyen y retroalimentan mutuamente.

Una posible medida del *entrainment* se puede obtener midiendo cuánto influye una serie sobre otra, considerándolas a ambas como parte de un sistema donde ambas interactúan. Este *entrainment*, entonces, sería direccional: queremos medir cuánto influye el interlocutor A sobre el interlocutor B y viceversa. Puede darse el caso en que ambos tengan fuerte interacción, en tal caso hablamos de *feedback*.

Para medir cuánto se mimetizan las dos series, se usa la función de correlación cruzada (f.c.c) [Cha13], que mide cuánto se parecen la serie X e Y aplicando un desplazamiento k, lo cual nos arroja como resultado un valor entre -1 y 1 (similar al coeficiente de correlación de la estadística clásica). Podemos aproximar la c.c.f. mediante la fórmula de la correlación cruzada muestral.

$$r_{AB}(k) = \begin{cases} \dfrac{\sum_{t=k+1}^{n}(A_t-\mu_A)(B_{t-k}-\mu_B)}{\sqrt{\sum_{t=1}^{n}(A_t-\mu_A)^2 \sum_{t=1}^{n}(B_t-\mu_B)^2}} & \text{si } k \geq 0 \\ \dfrac{\sum_{t=-k+1}^{n}(B_t-\mu_B)(A_{t+k}-\mu_A)}{\sqrt{\sum_{t=1}^{n}(A_t-\mu_A)^2 \sum_{t=1}^{n}(B_t-\mu_B)^2}} & \text{si } k < 0 \end{cases} \quad (3)$$

Podemos ver que, si $k \geq 0$, lo que hacemos es, a grandes rasgos, calcular la correlación de Pearson entre A y B, pero tomando los $n-k$ últimos valores de A y los $n-k$ primeros de B. Si $k < 0$, lo hacemos entre A y B, pero desplazando en sentidos inversos. Viéndolo de otra forma, si $k \geq 0$, estamos midiendo cuánto influye B sobre A contemplando un desplazamiento de k puntos; si $k \leq 0$ medimos la influencia de A sobre B a misma distancia. La utilización de estos desplazamientos está explicada en [GBLH15], donde se menciona que la influencia de los hablantes no es necesariamente inmediata sino que puede tener algunos segundos de demora para tomar lugar.

Para cada conversación, se estima entonces el correlograma cruzado, considerando desplazamientos tanto positivos como negativos. Hecho esto, en el estudio [KDMC08] sólo analizan la significancia de los resultados de la correlación cruzada, enumerando aquellos lags en los cuales esto ocurrió. En la sección 5 comentaremos cómo utilizamos la técnica descripta para la medición del entrainment direccional.

3.2. Modificaciones a TAMA

En primer lugar, [KDMC08] discute la disyuntiva de elegir un tamaño de ventana y step para el método: ventanas demasiado chicas pueden causar que no hayan segmentos de habla en ellas, mientras que un tamaño de ventana demasiado grande suavizaría en exceso la serie de tiempo. A colación de esto, dicho trabajo menciona dos posibles soluciones para el problema de los puntos

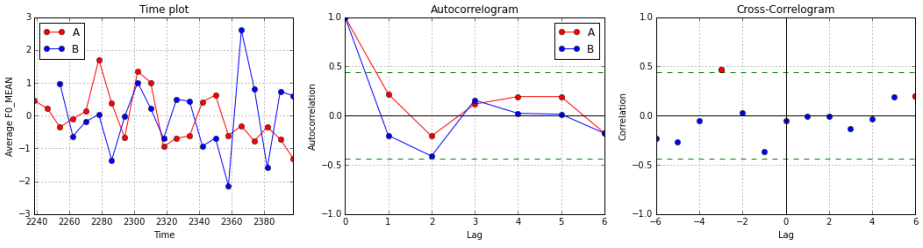

Figura 3. Time-plot producido por TAMA, junto a su autocorrelación y correlación cruzada

faltantes: interpolar (también mencionado en [DLSVC14]) o repetir el punto anterior de la serie.

Estos enfoques, sin embargo, pueden dar lugar a valores de *entrainment* artificialmente altos por la construcción misma de la serie, ya que nos generaría puntos correlacionados fuertemente entre sí en cada una de las series de los hablantes. Por otro lado, descartar aquellas conversaciones que tengan puntos faltantes puede ser demasiado restrictivo y eliminar de nuestro corpus una gran cantidad de datos valiosos. Teniendo estas cosas en mente, decidimos aceptar series de tiempo con datos faltantes, que pueden ser producto de ventanas sin segmentos de habla o con algunos demasiado pequeños que imposibiliten la medición de las variables acústico-prosódicas: por ejemplo interjecciones o backchanneling (*uh-huh* o *hmmm* en inglés).

En segundo lugar, se modificó el tamaño de la ventana para ajustarlo a nuestro corpus. En vez del step de 10" y tamaño de 20", optamos por 8 y 16 segundos respectivamente luego de efectuar un análisis con el fin de encontrar un balance entre el tamaño de la ventana y la cantidad de indefiniciones.

4. Time plots

Usando la técnica descripta con las variaciones que consideramos en la anterior sección, generamos dos series de tiempo para cada tarea. Como antes mencionamos, la ventana elegida es de 16 segundos con un step de 8 segundos lo cual da un overlap del 50 %.

Dada una ventana, puede ocurrir que alguno de los interlocutores no haya hablado, o su interacción haya sido demasiado breve como para medir sus variables acústico-prosódicas. Como ya mencionamos en la Sección 3.2, y a diferencia de [KDMC08], construimos las series sin ese punto, y sin interpolarlo tampoco.

De estas tareas, sólo nos quedamos con aquellas que tengan al menos 5 puntos definidos para cada serie, de manera que tenga sentido poder calcular la correlación cruzada más adelante. Con esto, no sólo nos interesa la duración de la charla, sino cierta calidad de las series generadas. En la Tabla 1 pueden verse las tareas que tuvimos en consideración, junto a su duración.

Como primer paso siempre recomendado en el análisis de series de tiempo [Cha13], graficamos los time plots conjunto de cada par de series, a la vez que sus autocorrelogramas. En la Figura 4 podemos observar un ejemplo de esto.

A priori, las series tienen aspecto de series autoregresivas de orden uno. Es decir, series que son de la forma $X_t = \alpha X_{t-1} + e_t + c$, con e_t ruido blanco, α y c constantes. Este hecho es esperable por la construcción misma del método TAMA, ya que la ventana de cada punto tiene un solapamiento con la ventana anterior. Más aún, uno esperaría que $\alpha \sim 0{,}5$ ya que nuestras ventanas tienen ese índice de overlap. Los autocorrelogramas de las series, por otro lado, tienen en su mayoría un valor significativo en $k = 1$, el valor del α de la autoregresión.

El hecho de que los autocorrelogramas desciendan rápidamente a cero es un indicio de que las series de tiempo construidas son estacionarias. Esto nos habilita a efectuar el análisis bivariado de las series.

Task	S-01	S-02	S-03	S-04	S-05	S-06	S-07	S-08	S-09	S-10	S-11	S-12
01	–	–	149.8	–	–	–	–	–	54.5	106.0	–	56.1
02	–	–	–	–	–	–	–	–	41.7	63.8	–	–
03	–	51.7	–	80.7	77.9	69.2	68.4	49.6	–	122.2	81.0	–
04	–	187.2	93.3	76.1	79.9	99.2	84.3	–	58.0	129.6	67.9	95.2
05	–	–	–	86.3	–	126.7	145.8	90.7	45.7	134.2	–	–
06	–	–	–	–	–	148.2	50.6	60.2	46.1	66.7	46.7	40.2
07	–	66.0	–	117.7	–	72.4	–	87.7	85.9	110.6	65.7	–
08	–	458.8	98.6	203.8	–	188.7	59.9	48.1	–	157.4	–	81.1
09	–	–	–	75.5	134.2	83.0	108.7	–	62.1	404.0	41.0	92.5
10	50.1	231.3	162.8	242.5	–	122.4	71.1	74.7	–	356.0	69.8	92.7
11	–	74.4	–	98.6	70.1	–	58.9	–	72.9	104.0	59.4	101.9
12	61.3	90.1	129.1	182.9	–	130.3	75.8	57.6	–	101.6	–	64.8
13	55.1	124.0	108.1	144.1	114.7	–	–	83.8	94.0	174.0	84.8	91.5
14	–	75.3	–	–	107.3	–	52.5	144.3	75.5	108.4	91.6	98.4

Cuadro 1. Tabla de tareas seleccionadas y sus duraciones

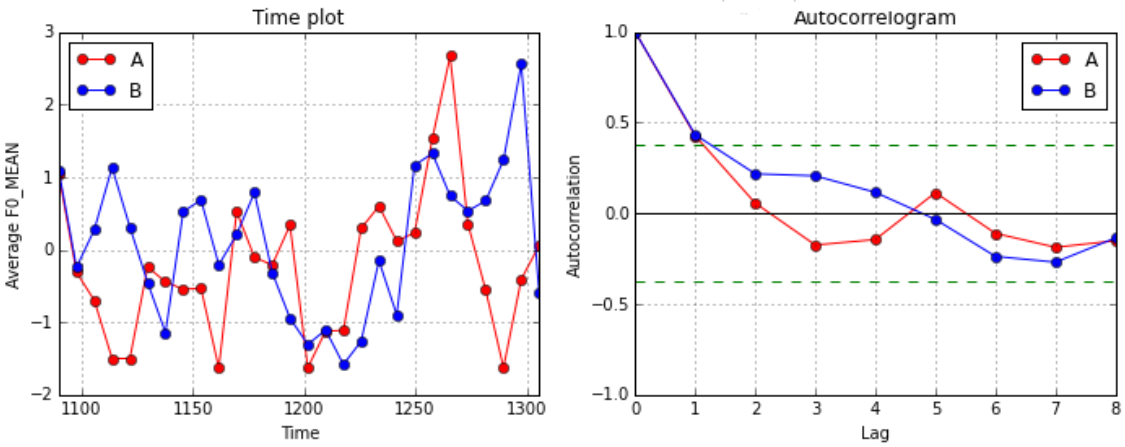

Figura 4. Time-plot generado por el método TAMA, junto a su autocorrelograma

5. Medición del Entrainment

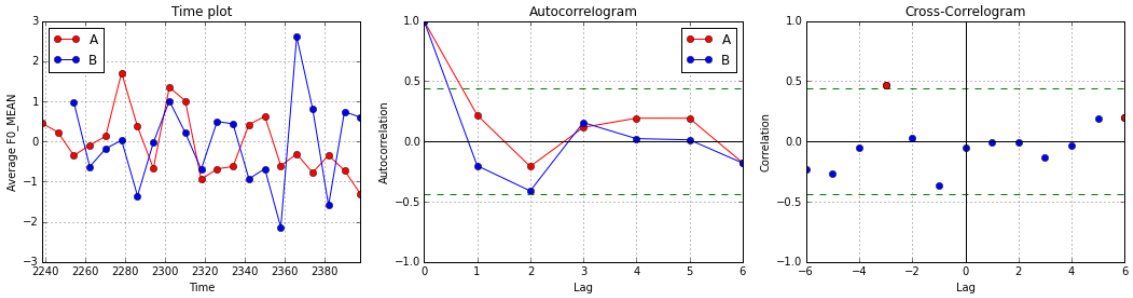

Figura 5. Time-plot generado por el método TAMA, junto a su autocorrelograma y correlograma cruzado

Considerando todo lo mencionado en la Sección 3.1, procedimos a definir una medida de *entrainment* basándonos en el cálculo de la correlación cruzada muestral. Recordemos que, bajo la definición dada en la ecuación 3 de $r_{AB}(k)$, al tomar $k \geq 0$ medíamos cuánto influía B sobre los futuros valores de A, y viceversa cuando $k \leq 0$. Se tomó la decisión de que este cálculo sólo se realice cuando el desplazamiento resulta en al menos 4 puntos que se solapan; si esto no ocurre, dejamos indefinido el valor en el correlograma cruzado.

Con esto en mente, definimos una primer métrica $\mathcal{E}^{(1)}_{AB}$ como el valor de $r_{AB}(k)$ con mayor valor absoluto, dado $k \leq 0$. Análogamente lo definimos para $\mathcal{E}^{(1)}_{BA}$. En 5 podemos observar en el lag -3 y 6 los valores de *entrainment* elegidos del correlograma.

En segundo lugar, definimos una segunda métrica $\mathcal{E}^{(2)}_{BA}$, como el valor absoluto de la primera, es decir:

$$\mathcal{E}^{(2)}_{BA} = |\mathcal{E}^{(1)}_{BA}| \tag{4}$$

Por último, cabe mencionar que a diferencia de [KDMC08] dónde sólo se hacía un análisis de significancia, nosotros vamos a utilizar esta medida independientemente de si es o no estadísticamente diferente de cero.

6. Panel de datos

Luego de construir las series de tiempo para cada una de las conversaciones que seleccionamos anteriormente, pasamos a construir una gran tabla que se utilizó en los análisis de regresión detallados en la siguiente sección. Para condensar todos nuestros datos, armamos una tabla por cada variable acústico-prosódica que contiene información definida para cada interlocutor y tarea de nuestro corpus.

session	speaker	task	entrainment	bored	engaged	encourages	clear
1	0	10	0.581475	0	5	5	5
1	0	12	-0.569677	1	5	5	5
1	0	13	0.533701	2	4	5	4
1	1	10	-0.917101	0	5	2	3
1	1	12	0.467112	0	5	4	2
1	1	13	-0.602364	0	5	4	3
2	0	3	0.520696	0	4	5	5
2	0	4	-0.241060	0	5	4	4
2	0	7	0.743719	0	5	4	5
2	0	8	0.147362	0	5	4	2

Cuadro 2. Extracto de la tabla generada para *F0 Mean*

Cada fila de esta tabla representa los datos de un hablante dentro de una tarea. Este hecho lo usamos fuertemente a la hora de definir los grupos en nuestro modelo de Efectos Fijos. En la Tabla 3 se describen las columnas generadas.

La tabla generada tuvo una dimensión de 210 x 14, siendo 210 la cantidad de tareas (contadas dos veces por cada hablante) y 14 las columnas mencionadas en la Tabla 3. Una forma de ver esta tabla es que, para cada sesión y hablante, tenemos una serie de tiempo sobre las tareas siendo los datos el grado del *entrainment* y las variables sociales. En la jerga econométrica, llamamos a este tipo de datos *de panel*[GP99]: un conjunto de mediciones temporales sobre un mismo sujeto a lo largo del tiempo. En este caso el sujeto es un hablante en una sesión, el tiempo son las tareas, y las mediciones son los valores medidos de *entrainment* y las diferentes variables sociales.

En la Tabla 2 tenemos una sección de la tabla. Los sujetos que tenemos en este ejemplo son 3: $speaker = 0$ y $session = 1$, $speaker = 1$ y $session = 1$, y

Campo	*Descripción*
session	número de sesión del corpus (1-12)
speaker	0 si corresponde al interlocutor A; B en otro caso
task	número de tarea en la sesión (1-14)
count	La cantidad de puntos definidos que tiene la serie
entrainment	Si $speaker = 0$, es $\mathcal{E}_{AB}$; $\mathcal{E}_{BA}$ en otro caso
best_lag	el lag del cross-correlogram donde se logra el *entrainment*
engaged_in_game	¿el sujeto parece comprometido con el juego?
difficult_for_partner_to_speak	¿al interlocutor se le dificulta hablar?
contributes_to_successful_completion	¿el sujeto contribuye para el éxito del equipo?
gives_encouragement	¿el sujeto alienta a su compañero?
making_self_clear	¿el sujeto se expresa con claridad?
planning_what_to_say	¿el sujeto piensa lo que va a decir?
bored_with_game	¿el sujeto se muestra aburrido?
dislikes_partner	¿al sujeto no le agrada su compañero?

Cuadro 3. Columnas de la tabla generada para ser utilizada en los análisis de regresión lineal

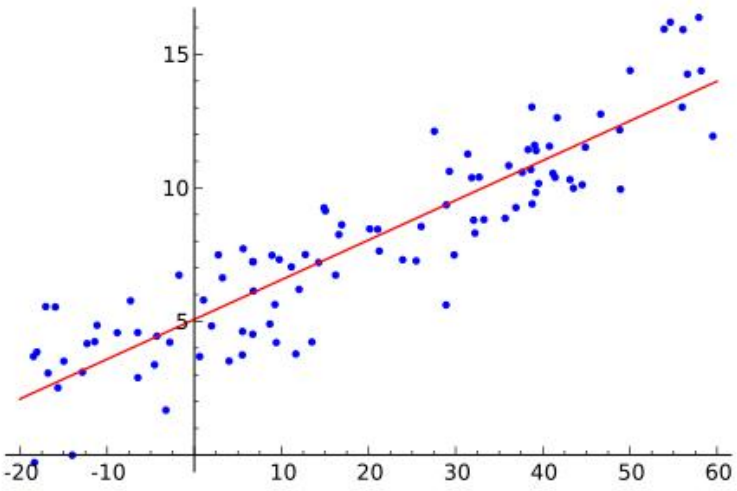

Figura 6. Ejemplo de Regresión Lineal

speaker = 0 y *session* = 2. También tenemos cinco series de tiempo para cada sujeto: *entrainment*, *bored*, *engaged*, *encourages* y *clear*. Vale la pena remarcar que estas series de tiempo, al igual que las que consideramos en la construcción de TAMA, pueden tener datos faltantes ya que, como fue descripto en la Sección 4, no tomamos todas las tareas de todas las sesiones sino aquellas que tienen cierta calidad de diálogo.

7. Análisis de regresión

Llegado a este punto, dada una variable acústico-prosódica, nos interesó evaluar la relación entre el entrainment o mimetización sobre dicha variable y las distintas variables sociales. Con esto en mente, planteamos un modelo de regresión lineal tomando como nuestra variable *explicativa* la mimetización, y la variable *dependiente* será la variable social elegida. Este análisis de regresión nos permitió observar cuál es la variación conjunta de ellas.

Nuestra hipotesis consistió en que la mimetización (por ejemplo, en la intensidad o pitch) se relacionaría de manera directa con ciertas variables sociales de connotación positiva (por ejemplo, la compenetración en el juego) y que se relacionaría de manera inversa con aquellas de carácter negativo (el aburrimiento o el desagrado por su compañero), siguiendo la línea de trabajos previos [GBLH15]. En la siguiente sección se describe el primer análisis, en el cual utilizamos el modelo "pooled" o agrupado, donde utilizamos todos los datos juntos indistintamente de la sesión y hablante del que provengan.

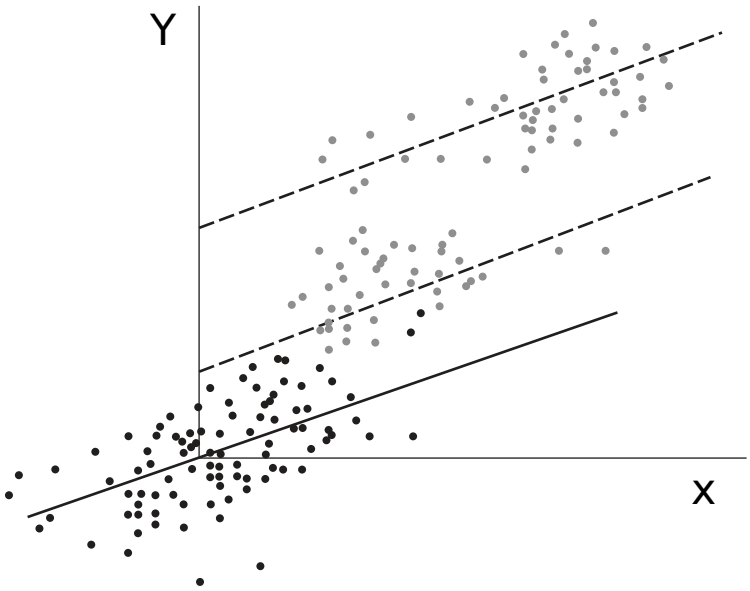

Figura 7. Ejemplo de datos de diferentes sujetos

8. Análisis mediante Regresión Lineal con Efectos Fijos

En esta sección, mostraremos el segundo análisis realizado, que consistió en aplicar un modelo de regresión lineal de cada variable social sobre el *unsigned entrainment*, pero esta vez utilizando un modelo que contemple la heterogeneidad por sesión y hablante .

9. Modelo de Efectos Fijos

El modelo agrupado o *pooled* que vimos en el anterior capítulo ignora la posibilidad de heterogeneidad observada y no observada; esto es, variables no medidas que afectan al sistema planteado. En el caso concreto de nuestro corpus, dicha heterogeneidad puede deberse a factores como la identidad o el género de los hablantes.

Los modelos de efectos fijos nos ayudan a controlar la heterogeneidad no observada cuando ésta es constante en el tiempo, dado un sujeto del sistema. Asumimos que estos factores son inherentes a la conversación entre el hablante y su interlocutor. Para este modelo, definimos los sujetos (en el lenguaje del modelo estadístico) como cada uno de los hablantes y sus respectivas sesiones. No nos importa si el mismo sujeto se repite en otra sesión: cada hablante de una sesión es un sujeto distinto para el modelo de efectos fijos.

10. Modelo de efectos fijos *within group*

Existen (al menos) dos variantes del modelo de efectos fijos: el modelo de variables ficticias y el modelo "dentro del grupo" (*within group*). Ambas técnicas son matemáticamente equivalentes, pero utilizaremos la segunda que nos provee el estimador de la pendiente, que es lo único que nos interesa.

La técnica utilizada consiste en, dentro de cada grupo, restarle tanto a la variable "dependiente" como a la "independiente" las medias dentro de cada grupo (de aquí proviene el nombre del método). Esto resulta en "sacarle" los efectos fijos no-temporales: en nuestro caso, estos están contemplados dentro de la ordenada al origen. Luego de efectuar esta transformación, se aplica regresión lineal "pooled", como en el análisis anterior. Producto del pre-procesamiento de los datos, la ordenada al origen es negligible. En [GP99, chap 16] se describe extensamente este procedimiento.

11. Resultados

Este modelo, utilizando como variable independiente al valor absoluto del *entrainment*, dio valores sustancialmente apreciables. Casi todas las variables acústico-prosódicas poseen al menos un valor significativo de $\widehat{\beta_2}$, destacándose *Int Mean*, *NHR* y *F0 Mean* con 3, 3 y 4 valores significativos respectivamente. Una versión simplificada tabla la podemos ver en la tabla 4 que grafica mediante tabla de doble entrada aquellos pares de variables acústico-prosódicas y variables sociales con coeficientes significativos y su signo.

Con respecto a las variables sociales, podemos observar que:

- *contributes-to-completion* se relaciona positivamente con el *unsigned entrainment* cuando la variable acústico-prosódica medida es *F0 Mean* o bien *NHR*. Esto significa que, cuando sube el valor absoluto del *entrainment*, esta variable positiva también lo hace con buena probabilidad. Esto es un efecto esperable: cuando hay mimetización, hay colaboración para el éxito en el juego.
- *making-self-clear*, otra variable que refleja una visión positiva del juego, también se relaciona positivamente con el *unsigned entrainment* para las variables *F0 Mean*, *NHR*, *Int Max* como a su vez para *Fonemas/seg*
- *engaged-with-game*, de la misma manera que las dos anteriores, relaciona positivamente pero sólo con *F0 Mean*
- *difficult-for-partner-to-speak*, se relaciona de la manera esperada con el *unsigned entrainment* cuando la variable acústico prosódica es *Int Max*; esto es, con $\widehat{\beta_2} < 0$. Esto tiene sentido, ya que a mayor mimetización de los interlocutores, la dificultad de estos para hablar debería disminuir. Por otro lado, $\widehat{\beta_2} > 0$ cuando la variable acústico-prosódica es *Int Mean*, lo cual no era un resultado esperado, pero bien puede ser parte del error estadístico.
- La variable *bored-with-game* se comporta de idéntica manera, sólo que con *F0 Mean*.

- *planning-what-to-say* y *gives-encouragement*, otras variables positivas, no presentan valores significativos.
- *dislikes-partner* no presenta valores significativos

En resumen, encontramos fuerte evidencia empírica en favor de la hipótesis de que el valor absoluto del *entrainment* se relaciona de manera positiva con atributos sociales de características positivas, mientras que lo hace de manera inversa con los que tienen connotaciones negativas.

Un hecho a destacar es que esta medida del entrainment es consistente con otras métricas definidas en otros trabajos, como las construídas en [GBLH15] sobre anotaciones discretas de los patrones entonacionales usando la convención ToBI[PBH94].

12. Conclusiones y trabajo futuro

	Int Max	*Int Mean*	*F0 Mean*	*F0 Max*	NHR	Fon/seg	Sil/seg	SHIMMER	JITTER
contributes		+	+ + +	+	++				
clear	+ + +		+		+ + +		+		
engaged		+	+ + +						+
planning									
encourages								+	
difficult	−−	++				−			
bored			− − −		+				
dislikes									

Cuadro 4. Tabla que representa los resultados significantes del análisis. En una de las entradas, tenemos los nombres abreviados de las variables sociales, y en la otra las variables a/p. El símbolo + representa valor significante y positivo de la pendiente de la regresión de efectos fijos, mientras que − representa significante y negativo. + representa $p < 0{,}10$, ++ $p < 0{,}5$, y + + + $p < 0{,}01$. Análogamente para − , −− , y − − −

Int Max	$\widehat{\beta_2}$	Std. Error	t value	Significance
contributes_to_successful_completion	0.0720	0.4258	1.689631E-01	0.8660
making_self_clear	1.6914	0.3820	4.427376E+00	0.0000
engaged_in_game	0.3456	0.2528	1.367266E+00	0.1732
planning_what_to_say	0.5655	0.5208	1.085851E+00	0.2790
gives_encouragement	0.4739	0.3744	1.265523E+00	0.2073
difficult_for_partner_to_speak	-0.6925	0.2863	-2.418510E+00	0.0166
bored_with_game	0.2110	0.2543	8.298495E-01	0.4077
dislikes_partner	-0.4254	0.3438	-1.237312E+00	0.2175
Int Mean	$\widehat{\beta_2}$	Std. Error	t value	Significance
contributes_to_successful_completion	0.6552	0.3610	1.814712E+00	0.0712
making_self_clear	0.9470	0.6080	1.557502E+00	0.1211
engaged_in_game	0.7091	0.3847	1.843187E+00	0.0669
planning_what_to_say	0.3636	0.5756	6.316937E-01	0.5284
gives_encouragement	0.4051	0.3482	1.163506E+00	0.2461
difficult_for_partner_to_speak	0.5287	0.2515	2.101960E+00	0.0369
bored_with_game	-0.0036	0.4106	-8.663987E-03	0.9931
dislikes_partner	0.5307	0.3889	1.364514E+00	0.1741
F0 Mean	$\widehat{\beta_2}$	Std. Error	t value	Significance
contributes_to_successful_completion	0.9752	0.3058	3.188448E+00	0.0017
making_self_clear	0.6998	0.3907	1.791239E+00	0.0749
engaged_in_game	0.8538	0.2773	3.078945E+00	0.0024
planning_what_to_say	0.6430	0.5363	1.198966E+00	0.2321
gives_encouragement	0.0006	0.3885	1.577445E-03	0.9987
difficult_for_partner_to_speak	-0.5323	0.3835	-1.388190E+00	0.1667
bored_with_game	-0.7663	0.2582	-2.968508E+00	0.0034
dislikes_partner	0.0688	0.3808	1.806265E-01	0.8569
F0 Max	$\widehat{\beta_2}$	Std. Error	t value	Significance
contributes_to_successful_completion	0.7628	0.4381	1.741129E+00	0.0833
making_self_clear	0.6718	0.4129	1.626984E+00	0.1054
engaged_in_game	0.5308	0.3776	1.405582E+00	0.1615
planning_what_to_say	0.0489	0.4210	1.161167E-01	0.9077
gives_encouragement	0.4724	0.5464	8.647145E-01	0.3883
difficult_for_partner_to_speak	-0.3208	0.2821	-1.136927E+00	0.2570
bored_with_game	-0.2584	0.3764	-6.865032E-01	0.4933
dislikes_partner	0.1249	0.3884	3.216226E-01	0.7481

Cuadro 5. Tablas con los resultados de la regresión de efectos fijos sobre el va,or absoluto de *entrainment* para *Int Max*, *Int Mean*, *F0 Mean* y *F0 Max*. En la segunda columna se cita el valor de $\widehat{\beta_2}$, la desviación estándar calculada, el t-valor obtenido y la significancia. Las columnas resaltadas corresponden a aquellas significantes, con diferentes matices de gris según $p < 0{,}10$, $p < 0{,}5$, o $p < 0{,}01$

NHR	$\widehat{\beta_2}$	Std. Error	t value	Significance
contributes_to_successful_completion	0.7271	0.3439	2.114275E+00	0.0358
making_self_clear	1.3576	0.3613	3.758007E+00	0.0002
engaged_in_game	0.1270	0.3431	3.702043E-01	0.7117
planning_what_to_say	-0.1625	0.4264	-3.811856E-01	0.7035
gives_encouragement	0.7665	0.4860	1.577201E+00	0.1165
difficult_for_partner_to_speak	-0.1683	0.3400	-4.951813E-01	0.6211
bored_with_game	0.5527	0.3084	1.792251E+00	0.0747
dislikes_partner	0.3457	0.3279	1.054410E+00	0.2931
Fonemas/seg	$\widehat{\beta_2}$	Std. Error	t value	Significance
contributes_to_successful_completion	0.5557	0.3577	1.553747E+00	0.1220
making_self_clear	0.7598	0.5085	1.494093E+00	0.1369
engaged_in_game	0.2440	0.2586	9.438356E-01	0.3465
planning_what_to_say	0.3614	0.5174	6.984626E-01	0.4858
gives_encouragement	0.0604	0.3829	1.576928E-01	0.8749
difficult_for_partner_to_speak	-0.6264	0.3374	-1.856257E+00	0.0650
bored_with_game	-0.0158	0.3204	-4.921947E-02	0.9608
dislikes_partner	0.0975	0.3137	3.108070E-01	0.7563
Sílabas/seg	$\widehat{\beta_2}$	Std. Error	t value	Significance
contributes_to_successful_completion	0.2451	0.3663	6.692398E-01	0.5042
making_self_clear	0.7934	0.4094	1.937743E+00	0.0542
engaged_in_game	0.4956	0.3642	1.360687E+00	0.1753
planning_what_to_say	0.4429	0.5189	8.535430E-01	0.3945
gives_encouragement	0.2363	0.4192	5.637211E-01	0.5736
difficult_for_partner_to_speak	0.1856	0.3481	5.332959E-01	0.5945
bored_with_game	-0.2909	0.3606	-8.067536E-01	0.4208
dislikes_partner	0.1768	0.3452	5.120454E-01	0.6092
Jitter	$\widehat{\beta_2}$	Std. Error	t value	Significance
contributes_to_successful_completion	0.5770	0.3759	1.534821E+00	0.1265
making_self_clear	0.5057	0.4881	1.036143E+00	0.3015
engaged_in_game	0.4972	0.2515	1.977130E+00	0.0495
planning_what_to_say	-0.0417	0.4628	-9.000210E-02	0.9284
gives_encouragement	-0.0160	0.3502	-4.554031E-02	0.9637
difficult_for_partner_to_speak	-0.2788	0.3126	-8.917922E-01	0.3737
bored_with_game	0.1233	0.3155	3.906725E-01	0.6965
dislikes_partner	-0.1171	0.2788	-4.198582E-01	0.6751
Shimmer	$\widehat{\beta_2}$	Std. Error	t value	Significance
contributes_to_successful_completion	0.3745	0.2754	1.359709E+00	0.1756
making_self_clear	-0.0097	0.3821	-2.544762E-02	0.9797
engaged_in_game	0.2434	0.2881	8.449092E-01	0.3993
planning_what_to_say	-0.6040	0.4735	-1.275476E+00	0.2037
gives_encouragement	0.3638	0.2057	1.768094E+00	0.0787
difficult_for_partner_to_speak	0.2707	0.2720	9.952034E-01	0.3209
bored_with_game	-0.3635	0.2772	-1.311203E+00	0.1914
dislikes_partner	-0.1895	0.2667	-7.105564E-01	0.4783

Cuadro 6. Tablas con los resultados de la regresión de efectos fijos para *NHR*, *Sílabas/seg*, *Fonemas/seg*, *Shimmer* y *Jitter*. En la segunda columna se cita el valor de $\widehat{\beta_2}$, la desviación estándar calculada, el t-valor obtenido y la significancia. Las columnas resaltadas corresponden a aquellas significantes, con diferentes matices de gris según $p < 0{,}10$, $p < 0{,}5$, o $p < 0{,}01$

En el presente trabajo, analizamos cómo dos métricas dinámicas del fenómeno conocido como *entrainment* o mimetización en el plano acústico-prosódico se relacionan con las percepciones, por parte de terceros, de aspectos sociales de la interacción entre los participantes. Ambas métricas pueden computarse en forma automática a partir de las grabaciones de las conversaciones, con un hablante por canal, y con transcripciones alineadas temporalmente al audio. Todo este análisis se da en el contexto de un juego orientado a tareas, que comprende interacciones de una naturaleza muy similar a las de una interfaz humano-computadora.

Estas métricas fueron construidas a través del análisis de series de tiempo, y apuntan a cuantificar cuánto se imitan o mimetizan los hablantes en términos de sus variables acústico-prosódicas. En primer lugar, contemplamos una métrica que penaliza el dis-entrainment con valores negativos. Se aplicó análisis de regresión sobre esta métrica, y los resultados que dio no fueron significativos. En segundo lugar, construimos una métrica que valora de igual manera el *entrainment* y el dis-*entrainment*, de acuerdo a trabajos previos que sugerían que el segundo fenómeno puede considerarse en algunas circunstancias como un mecanismo de adaptación cooperativa. Al efectuar el análisis de regresión sobre esta métrica, los resultados fueron significativos y consistentes con la hipótesis planteada de que el *entrainment* se relaciona positivamente con características sociales favorables de la conversación, mientras que lo hace de manera inversa con aquellas negativas.

Respecto a trabajos anteriores que construyen medidas del *entrainment* acústico-prosódico, la métrica usada en esta tesis se comporta de manera consistente preservando las relaciones expuestas en otros trabajos entre el *entrainment* y aquellas variables sociales de carácter positivo y negativo. Esta métrica, además, se puede efectuar sin intervención manual, a diferencia de aquellas que utilizan ToBI o anotaciones de otro tipo. A su vez, la cuantificación presentada evita el problema del alineamiento de turnos mediante la abstracción de éstos usando series de tiempo.

Una contribución importante de este trabajo es la validación de la métrica introducida en [KDMC08], dando indicios de que ésta efectivamente captura rasgos relevantes de la interacción, que a su vez guardan relación con la percepción social de la conversación. Igual de importante es el uso del valor absoluto de la correlación cruzada, como medida unificadora del *entrainment* y *disentrainment* y que remarca la importancia del segundo fenómeno dentro de la comunicación verbal, a la luz de últimos trabajos acerca de la divergencia en el diálogo.

A pesar de que los resultados son prometedores, siguen siendo preliminares y su robustez requiere de más validaciones. Como trabajo futuro, proponemos reproducir estos análisis sobre otros corpus de habla, como por ejemplo Switchboard [3]. Adicionalmente, se debería verificar el impacto del proceso de pre-whitening, ya que un análisis preliminar no mostró grandes diferencias entre usar o no este filtro. Otra dirección posible es utilizar herramientas de análisis multivariado de series de tiempo sobre las diferentes variables acústico-prosódicas y sobre la base de esto construir nuevas métricas del *entrainment* prosódico.

[3] https://catalog.ldc.upenn.edu/LDC97S62

Referencias

[Bre96] Susan E Brennan. Lexical entrainment in spontaneous dialog. *Proceedings of ISSD*, 96:41–44, 1996.

[BSD95] Judee K Burgoon, Lesa A Stern, and Leesa Dillman. Interpersonal adaptation: Dyadic interaction patterns. 1995.

[CB99] Tanya L Chartrand and John A Bargh. The chameleon effect: the perception–behavior link and social interaction. *Journal of personality and social psychology*, 76(6):893, 1999.

[Cha13] Chris Chatfield. *The analysis of time series: an introduction*. CRC press, 2013.

[DLSVC14] Céline De Looze, Stefan Scherer, Brian Vaughan, and Nick Campbell. Investigating automatic measurements of prosodic accommodation and its dynamics in social interaction. *Speech Communication*, 58:11–34, 2014.

[GBLH15] Agustın Gravano, Štefan Benuš, Rivka Levitan, and Julia Hirschberg. Backward mimicry and forward influence in prosodic contour choice in standard american english. In *Sixteenth Annual Conference of the International Speech Communication Association*, 2015.

[GP99] Damodar N Gujarati and Dawn C Porter. Essentials of econometrics. 1999.

[Gra09] Agustin Gravano. *Turn-taking and affirmative cue words in task-oriented dialogue*. Columbia University, 2009.

[KDMC08] Spyros Kousidis, David Dorran, Ciaran McDonnell, and Eugene Coyle. Times series analysis of acoustic feature convergence in human dialogues. In *Proceedings of Interspeech*, 2008.

[KDW+08] Spyros Kousidis, David Dorran, Yi Wang, Brian Vaughan, Charlie Cullen, Dermot Campbell, Ciaran McDonnell, and Eugene Coyle. Towards measuring continuous acoustic feature convergence in unconstrained spoken dialogues. 2008.

[LH11] Rivka Levitan and Julia Bell Hirschberg. Measuring acoustic-prosodic entrainment with respect to multiple levels and dimensions. 2011.

[NGH08] Ani Nenkova, Agustín Gravano, and Julia Hirschberg. High frequency word entrainment in spoken dialogue. In *Proceedings of the 46th annual meeting of the association for computational linguistics on human language technologies: Short papers*, pages 169–172. Association for Computational Linguistics, 2008.

[PBH94] John F Pitrelli, Mary E Beckman, and Julia Hirschberg. Evaluation of prosodic transcription labeling reliability in the tobi framework. In *ICSLP*, 1994.

[PH05] Roberto Pieraccini and Juan Huerta. Where do we go from here? research and commercial spoken dialog systems. In *6th SIGdial Workshop on Discourse and Dialogue*, 2005.

[RBL+06] Antoine Raux, Dan Bohus, Brian Langner, Alan W Black, and Maxine Eskenazi. Doing research on a deployed spoken dialogue system: one year of let's go! experience. In *INTERSPEECH*, 2006.

[RKM06] David Reitter, Frank Keller, and Johanna D Moore. Computational modelling of structural priming in dialogue. In *Proceedings of the Human Language Technology Conference of the NAACL, Companion Volume: Short Papers*, pages 121–124. Association for Computational Linguistics, 2006.

[WRWN05] Nigel G Ward, Anais G Rivera, Karen Ward, and David G Novick. Root causes of lost time and user stress in a simple dialog system. In *Ninth European Conference on Speech Communication and Technology*, 2005.