Universidad de Buenos Aires
Facultad de Ciencias Exactas y Naturales
Departamento de Computación

Mimetización entre interlocutores

Tesis presentada para optar al título de
Licenciado en Ciencias de la Computación

Juan Manuel Pérez

Director: Agustín Gravano
Codirector: Ramiro Gálvez
Buenos Aires, 2015

MEDICIÓN DE LA MIMETIZACIÓN ENTRE INTERLOCUTORES UTILIZANDO SERIES DE TIEMPO

El *entrainment* (mimetización) es un fenómeno inconsciente que se manifiesta a través de la adaptación de posturas, forma de hablar, gestos faciales y otros comportamientos entre dos o más interactores. A su vez, la ocurrencia de esta mimetización está fuertemente emparentada con el sentimiento de empatía y compenetración entre los participantes.

En esta tesis, nos proponemos explorar una técnica algorítmica para detectar el entrainment entre variables prosódicas de dos personas. Esta técnica nos permitirá determinar si existe o no convergencia para ciertos parámetros, y ver como está ésto correlacionado con variables sociales tales como la empatía, la compenetración con la tarea, y otras.

Palabras claves: Guerra, Rebelión, Wookie, Jedi, Fuerza, Imperio (no menos de 5).

MEASURING ENTRAINMENT BETWEEN SPEAKERS USING TIME SERIES

In a galaxy far, far away, a psychopathic emperor and his most trusted servant – a former Jedi Knight known as Darth Vader – are ruling a universe with fear. They have built a horrifying weapon known as the Death Star, a giant battle station capable of annihilating a world in less than a second. When the Death Star's master plans are captured by the fledgling Rebel Alliance, Vader starts a pursuit of the ship carrying them. A young dissident Senator, Leia Organa, is aboard the ship & puts the plans into a maintenance robot named R2-D2. Although she is captured, the Death Star plans cannot be found, as R2 & his companion, a tall robot named C-3PO, have escaped to the desert world of Tatooine below. Through a series of mishaps, the robots end up in the hands of a farm boy named Luke Skywalker, who lives with his Uncle Owen & Aunt Beru. Owen & Beru are viciously murdered by the Empire's stormtroopers who are trying to recover the plans, and Luke & the robots meet with former Jedi Knight Obi-Wan Kenobi to try to return the plans to Leia Organa's home, Alderaan. After contracting a pilot named Han Solo & his Wookiee companion Chewbacca, they escape an Imperial blockade. But when they reach Alderaan's coordinates, they find it destroyed - by the Death Star. They soon find themselves caught in a tractor beam & pulled into the Death Star. Although they rescue Leia Organa from the Death Star after a series of narrow escapes, Kenobi becomes one with the Force after being killed by his former pupil - Darth Vader. They reach the Alliance's base on Yavin's fourth moon, but the Imperials are in hot pursuit with the Death Star, and plan to annihilate the Rebel base. The Rebels must quickly find a way to eliminate the Death Star before it destroys them as it did Alderaan (aprox. 200 palabras).

Keywords: War, Rebellion, Wookie, Jedi, The Force, Empire (no menos de 5).

Índice general

1. INTRODUCCIÓN

1.1. Sistemas de diálogo

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, que radica en el intercambio de afecto, emociones, actitudes y otras intenciones de los participantes. La habilidad de los participantes de poder expresar, comprender, y reaccionar de acuerdo a estas señales sociales es necesaria para el entendimiento mutuo y una comunicación exitosa.

Un aspecto particular de la comunicación es el fenómeno de *entrainment*(arrastre, mimetización, efecto camaleón), que comprende la adaptación inconsciente de las variables acústicas/prosódicas(a/p) (por ejemplo, el tono de la voz, la velocidad del habla, etc) de manera dinámica en el transcurso de una o varias interacciones. Este fenómeno ha sido introducido por *Brennan et al*[Bre96] en 1996, y se ha observado que la convergencia de los participantes en estas variables ocurre en conjunto con una interacción más fluída y un mayor sentimiento de simpatía por sus interlocutores [CB99].

Poder medir esta mimetización de los interlocutores no es una tarea fácil, sin embargo. En primer lugar, un diálogo no es una sucesión de turnos, sino que es una serie de tiempo dinámica, llena de interrupciones. Más aún, la mimetización no tiene un carácter instantáneo, sino que se sucede a lo largo de la interacción entre los participantes. Estos factores dificultan ostensiblemente poder modelar este fenómeno.

1.2. 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 se describe a este fenómeno como una "imitación no conciente 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*, 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.

1.3. Midiendo la mimetización

Muchos estudios han examinado la mimetización del habla, listados en [DLSVC14]. Por ejemplo, [LGW+12] propone un método basado en el cálculo de la media de la feature para cada hablante; sin embargo, estos modelos no capturan la esencia dinámica del proceso de *entrainment*.

A la hora de hacer comparaciones razonables entre dos interlocutores, surgen dos problemas [KDW+08]. En primer lugar, las curvas tienen diferentes escalas y deben ser normalizadas (por ejemplo, el *pitch* entre un interlocutor masculino y uno femenino), aunque en algunos casos las comparaciones desnormalizadas tienen sentido (volumen). En segundo lugar, surge el problema del alineamiento. ¿Qué partes del diálogo de un interlocutor deberían compararse con qué otras partes? Un approach de comparar interlocuciones uno a uno es demasiado simple y no captura situaciones de diálogo reales.

Para atacar estos inconvenientes, utilizamos el método TAMA (Time Aligned Moving Average), 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. El entrainment podría entonces pensarse, en primera instancia, como la correlación cruzada entre estas series generadas [Cha13].

2. MÉTODO

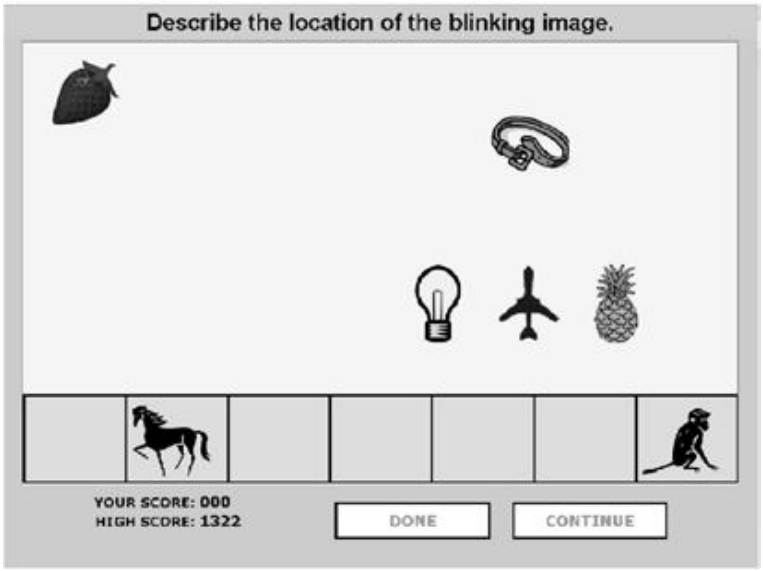


Fig. 2.1: Juego del Columbia Games

2.1. Columbia Game Corpus

Nuestro corpus consiste en doce conversaciones diádicas (i.e., con dos participantes) entre trece 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.

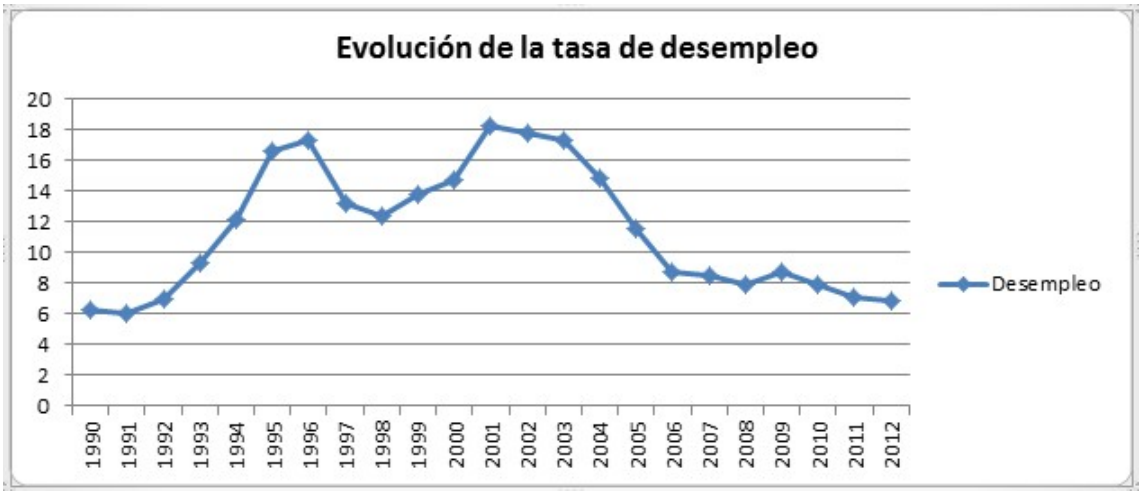


Fig. 2.2: Gráfico de serie de tiempo de la evolución del desempleo en Argentina

2.2. Series de Tiempo

Definición Informal

En términos informales, una serie de tiempo es un conjunto de datos recolectados secuencialmente en el tiempo. Este tipo de datos se dan en varios campos de estudio, mayormente en economía, ciencias de la atmósfera, y otras.

Ejemplos de series de tiempo:

- Volumen de lluvias en sucesivos días de un año
- Precio de acciones en diferentes meses
- Cantidad de habitantes de una ciudad año a año

¿Para qué queremos series de tiempo?

Hay varios motivos por los cuales uno querría efectuar un análisis de una serie de tiempo.

1) Descripción Usualmente, lo primero que se hace al obtener la serie de tiempo es graficarla y obtener las características más notorias de ésta. Por ejemplo, en 2.2 puede notarse que hay una tendencia decreciente del 2003 hasta el 2012. En otras (como en el volumen de lluvias) podrá observarse cierta estacionalidad en la serie.

Si bien ésto no requiere técnicas avanzadas de análisis, es el primer paso fundamental para comprender una serie de tiempo.

2) Explicación Cuando analizamos dos o más series de tiempo, podemos querer ver cómo se comportan en conjunto. Una variación en una serie de tiempo puede producir un cambio en otra. Por ejemplo, podemos intentar buscar como varían en conjunto la temperatura diaria con la cantidad de mL de lluvia caídos.

3) Predicción Dada una serie de tiempo, podemos querer intentar predecir un valor futuro.

4) Control Dado un proceso del que se mide cierto parámetro de calidad, podemos querer ajustar variables de entrada para mantenerla en ciertos valores.

En nuestro caso, nos es de interés 1 y 2.

2.2.1. Procesos estocásticos

Definición 1. Una proceso estocástico es una colección de variables aleatorias $\{X_t\}_{t\in T}$ donde T es un conjunto de puntos de tiempo. En nuestro caso, nos interesa $T = \mathbb{N}$, de manera que el proceso será de la forma $X_1, X_2, \ldots$

Podemos entender un proceso estocástico como un conjunto de variables ordenadas por el tiempo. Llamamos serie de tiempo a una observación de este proceso estocástico. Usualmente sólo tendremos esta instancia, a diferencia de otros problemas estadísticos donde tendremos muchas observaciones.

2.2.2. Estacionariedad

Un concepto importante en series de tiempo es el de estacionariedad. En lenguaje coloquial, una serie de tiempo estacionaria es aquella en la que no observamos cambios sistemáticos de ésta en el tiempo: si tomamos una parte de la serie, y observamos otro parte distinta de la serie, las propiedades de ésta se mantienen.

Ejemplos de series de tiempo estacionarias son las de ruido blanco, y ejemplos de no estacionarias aquellas que tienen una tendencia. (mejorar esto...)

Definición 2. Un proceso estocástico $X_i, i \in \mathbb{N}$ se dice fuertemente estacionario si, para todo conjunto de índices $t_1, \ldots, t_n$ y para un desplazamiento $\tau \in \mathbb{N}$ tenemos que

$$F_{X_{t_1},X_{t_2},\ldots,X_{t_n}} = F_{X_{t_1}+\tau,X_{t_2}+\tau,\ldots,X_{t_n}+\tau}$$

Es decir, que la función de probabilidad se preserva por traslados.

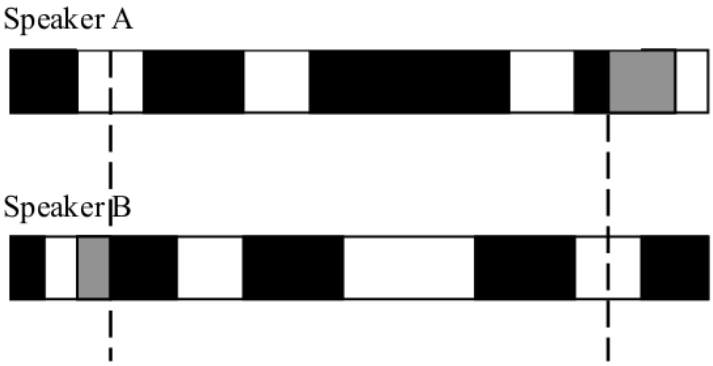
Se derivan como propiedades que, para todo X_t y cualquier desplazamiento τ

$$E[X_t] = E[X_{t+\tau}] \tag{2.1}$$
$$Var[X_t] = Var[X_{t+\tau}] \tag{2.2}$$
$$Cov(X_s, X_t) = Cov(X_{s+\tau}, X_{t+\tau}) \tag{2.3}$$

Las ecuaciones 2.1 y 2.2 nos dicen que tanto la media como la varianza son constantes (no dependen de t), y que la covarianza sólo depende de la diferencia $|s - t|$.

Definición 3. Un proceso se dice débilmente estacionario si cumple 2.1, 2.2, 2.3

A partir de aquí, cuando hablemos de series estacionarias estaremos hablando de series débilmente estacionarias



2.3. Descripción TAMA

Para construir la serie de tiempo de cada interlocutor, dividimos primero el diálogo en ventanas solapadas de igual tamaño [KDW+08]. A la diferencia entre ventana y ventana llamaremos *frame step*, y al tamaño de ventana *frame length*.

Nuestro corpus está anotado de manera que tenemos separadas los intervalos donde los interlocutores hablan (llamaremos a cada uno de éstos locuciones o *utterances*). Para cada una de los frames, calcularemos la media

$$\mu = \sum_{i=1}^{N} f_i dr_i \tag{2.4}$$

$$dr_i = \frac{d_i}{\sum_{i=1}^{N} d_i} \tag{2.5}$$

$$\tag{2.6}$$

donde i itera sobre las locuciones dentro del *frame*, d_i es la duración de la locución y f_i es el valor de la *feature* que estamos midiendo.

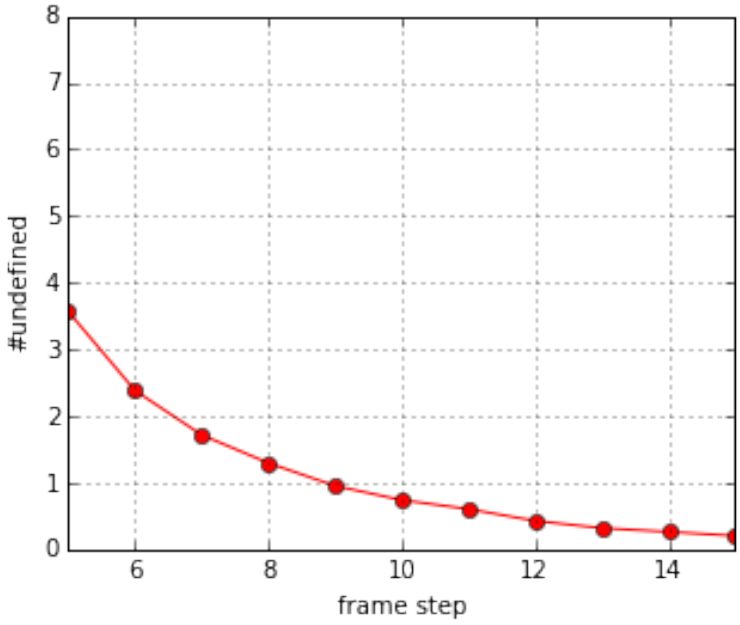
Como se ve en 2.4, el μ que calculamos es una media ponderada por la duración de las locuciones. Así, por ejemplo, al calcular una serie de tiempo sobre el *pitch*, la contribución de interjecciones (usualmente de alto valor) estará disminuída por su breve duración.

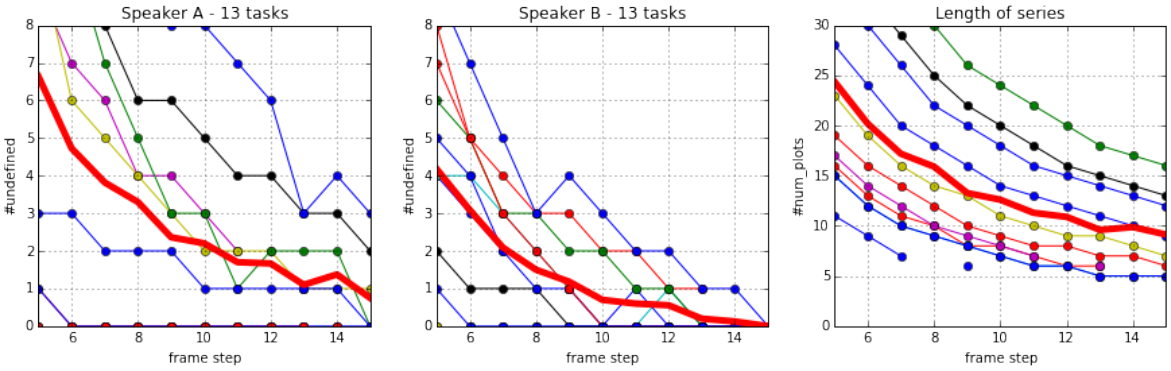
La serie de tiempo constará entonces de la secuencia de medias calculadas con 2.4 para cada uno de los frames.

2.4. Selección de Ventana

En [KDMC08] se menciona una elección de *frame step* y *frame length* de 10s y 20s respectivamente. En el caso de nuestro corpus, quisimos buscar los parámetros que mejor se ajustaban a éste, manteniendo la superposición del 50 % entre ventanas sucesivas. Con lo que nos queda que $FL = 2 * FS$

¿Qué queremos optimizar? La métrica que elegimos para ésto es encontrar un balance entre un frame no tan grande (para no suavizar en exceso la curva) y que nos reduzca considerablemente la cantidad de indefiniciones; es decir, aquellas ventanas que tomamos





en un interlocutor que no tienen ninguna interacción de su parte. Para ver ésto, graficamos la cantidad de indefiniciones en función del step tomado.

Dentro del rango de $FS \in \{5'', 6'', \ldots, 15''\}$, graficamos para cada sesión, tarea y cada interlocutor las curvas de indefiniciones. A su vez, para mayor claridad, graficamos una curva que promedie todas las tareas de una sesión.

Para tener una visión general de lo que ocurría en todas las sesiones, graficamos una curva promedio de todas las sesiones. En ésta puede observarse que hasta $8'' - 10''$ hay un fuerte descenso de las indefiniciones, que luego se atenúa. Dado que en general tenemos tareas cortas, preferimos tomar $8''$ como step, y $16''$ como largo de ventana.

OBS: podríamos cambiar ésto a un boxplot!

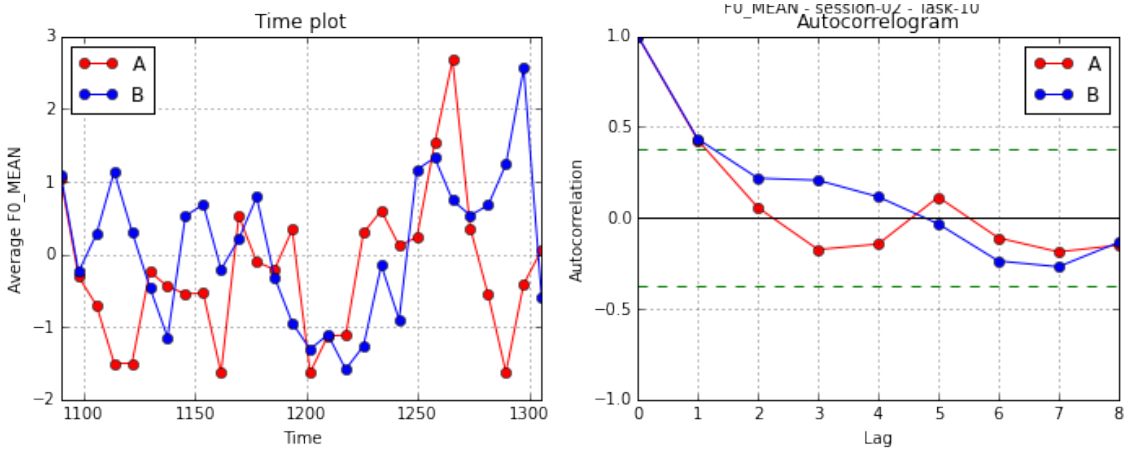


Fig. 2.3: Time-plot producido por TAMA, junto a su autocorrelación

	session	speaker	task	entrainment	bored_with_game	engaged_in_game	gives_encouragement	making_self_clear	planning_what_to_say
0	1	0	10	0.581475	0	5	5	5	3
1	1	0	12	-0.569677	1	5	5	5	3
2	1	0	13	0.533701	2	4	5	4	3
3	1	1	10	-0.917101	0	5	2	3	0
4	1	1	12	0.467112	0	5	4	2	4
5	1	1	13	-0.602364	0	5	4	3	1
6	2	0	3	0.520696	0	4	5	5	2
7	2	0	4	-0.241060	0	5	4	4	4
8	2	0	7	0.743719	0	5	4	5	3
9	2	0	8	0.147362	0	5	4	2	2

Fig. 2.4: Tabla de tareas seleccionadas y sus duraciones

2.5. Time plots

Usando la técnica descripta, generamos dos series de tiempo para cada tarea. Como antes mencionamos, la ventana elegida es de 16″ con un step de 8″ 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 a/p. En ese caso, y a diferencia de [KDW+08], construimos las series sin ese punto, y sin interpolarlo tampoco. (¿por qué no estamos interpolando en vez de dejar los puntos vacíos?)

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 (¿podemos justificar un poco más ésto?). Con ésto, no sólo nos interesa la duración de la charla, sino cierta calidad de las series generadas. En 2.4 pueden verse las tareas que tuvimos en consideración, a la vez que su duración.

Los autocorrelogramas de las series bajan rápidamente a cero, un indicio (necesario pero no suficiente) de que las series son estacionarias *¿necesitamos hacer algún chequeo más fuerte de ésto? En [KDMC08] ni se calientan en hacerlo*

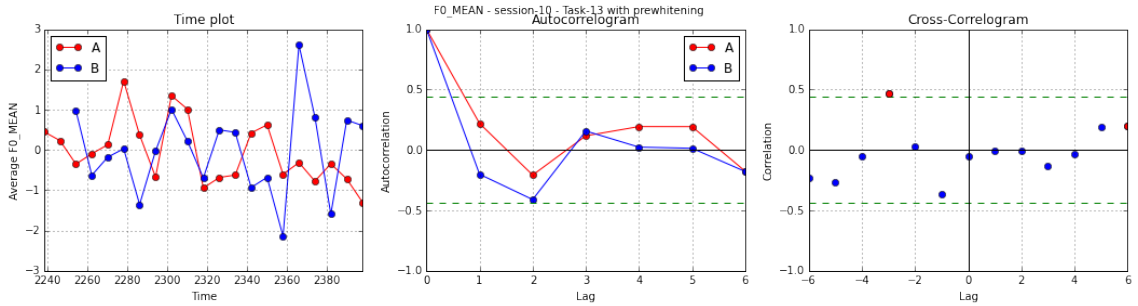


Fig. 2.5: Time-plot producido por TAMA, junto a su autocorrelación y correlación cruzada

2.6. Análisis Bivariado

Para medir cuánto se "mimetizan" las dos series, utilizaremos la función de correlación cruzada (c.c.f), que mide cuánto se parecen la serie X e Y aplicando un desplazamiento k, dándonos un valor entre -1 y 1 (similar a la correlación de la estadística clásica).

Podemos aproximar la c.c.f. mediante la fórmula de la correlación cruzada muestral.

$$r_{AB}(k) = \begin{cases} \dfrac{\sum_{t=|k|+1}^{n} (A_t - \mu_a)(B_{t-|k|} - \mu_B)}{\sqrt{\sum_{t=1}^{n} (A_t - \mu_a)^2 \sum_{t=1}^{n} (B_t - \mu_b)^2}} & \text{si } k \geq 0 \\ \dfrac{\sum_{t=|k|+1}^{n} (B_t - \mu_b)(A_{t-|k|} - \mu_A)}{\sqrt{\sum_{t=1}^{n} (B_t - \mu_b)^2 \sum_{t=1}^{n} (A_t - \mu_a)^2}} & \text{si } k < 0 \end{cases} \tag{2.7}$$

Podemos ver que, si $k \geq 0$, lo que hacemos es, a grandes rasgos, calcular la correlación de Pearson entre A_{t+k} e B_t. Si $k < 0$, lo hacemos entre A_t e B_{t+k}.

Para cada tarea, calculamos un correlograma cruzado para $k \in \{-6, -5, \ldots, 0, \ldots, 6\}$. Los valores de $k \geq 0$ los podemos considerar como aquellos en los cuales nos estamos fijando si B se mimetiza con A, y aquellos $k \leq 0$ al revés. Luego, definimos

$$A \to B = max\{r_{AB}(k), k \leq 0\} \tag{2.8}$$

$$B \to A = max\{r_{AB}(k), k \geq 0\} \tag{2.9}$$

$$\tag{2.10}$$

$A \to B$ es el entrainment direccional de A hacia B, que mide cuánto se mimetiza B a A. (explicar un poco más ésto)

2.7. Armado de Tabla

Para condensar todos nuestros datos, armamos una tabla por cada variable a/p. Esta tabla contiene información definida para cada interlocutor, tarea y sesión de nuestro corpus.

1. session: número de sesión
2. speaker: 0 si corresponde al interlocutor A; B en otro caso
3. task: número de tarea
4. count: La cantidad de puntos definidos que tiene la serie
5. entrainment: Si $speaker = 0$, es $A \to B$; $B \to A$ en otro caso
6. best_lag: el lag del cross-correlogram donde se logra el *entrainment*
7. tama_mean: el promedio de la variable

Además, agregamos las variables sociales (relativas al interlocutor) para cada fila:

1. contributes_to_successful_completion
2. making_self_clear
3. engaged_in_game
4. planning_what_to_say
5. gives_encouragement
6. difficult_for_partner_to_speak
7. bored_with_game
8. dislikes_partner

El corpus original cuenta con más variables pero éstas son las únicas que tomaremos en cuenta (citar algo acá!)

En el corpus original, cada variable estaba replicada por cada interlocutor, y por sí o por no, de manera que teníamos:

1. $conversation\_awkward\_A\_yes$
2. $conversation\_awkward\_A\_no$
3. $conversation\_awkward\_B\_yes$
4. $conversation\_awkward\_B\_no$

Ésto nos da una tabla de 210 filas, y 21 columnas. Para cada sesión y speaker, podemos pensar que tenemos una serie de tiempo donde el tiempo es cada tarea, y los datos son el 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 interlocutor en una sesión, el tiempo son las tareas, y las mediciones son los entrainments

	session	speaker	task	entrainment	bored_with_game	engaged_in_game	gives_encouragement	making_self_clear	planning_what_to_say
0	1	0	10	0.581475	0	5	5	5	3
1	1	0	12	-0.569677	1	5	5	5	3
2	1	0	13	0.533701	2	4	5	4	3
3	1	1	10	-0.917101	0	5	2	3	0
4	1	1	12	0.467112	0	5	4	2	4
5	1	1	13	-0.602364	0	5	4	3	1
6	2	0	3	0.520696	0	4	5	5	2
7	2	0	4	-0.241060	0	5	4	4	4
8	2	0	7	0.743719	0	5	4	5	3
9	2	0	8	0.147362	0	5	4	2	2

Fig. 2.6: Panel de datos

Bibliografía

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

[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.

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

[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.

[LGW+12] Rivka Levitan, Agustín Gravano, Laura Willson, Stefan Benus, Julia Hirschberg, and Ani Nenkova. Acoustic-prosodic entrainment and social behavior. In *Proceedings of the 2012 Conference of the North American Chapter of the Association for Computational Linguistics: Human language technologies*, pages 11–19. Association for Computational Linguistics, 2012.

[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.