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

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

Tesis de
Licenciatura en Ciencias de la Computación

Juan Manuel Pérez

Director: Agustín Gravano
Codirector: Ramiro H. Gálvez
Buenos Aires, 2016

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

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 buena parte de 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.

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 nuestro caso, nos es de interés el *entrainment* sobre las variables acústico-prosódicas, como el tono, intensidad, y otras.

En el presente trabajo, nos proponemos explorar y refinar una métrica del entrainment acústico-prosódico definida en trabajos previos. Analizamos la relación entre los valores obtenidos y las percepciones sociales que terceros tienen sobre las conversaciones, en un corpus de diálogos orientados a tareas en inglés.

Palabras claves: Procesamiento del Habla, Series de Tiempo, Entrainment, Regresión Lineal

A mis amigos.
A mis compañeros de la Facultad, con los que tanto remamos.
A mi familia, que me acompañaron.

Índice general

1.. Introducción **1**
1.1. Sistemas de diálogo . . . 2
1.2. Mimetización . . . 2
1.3. Midiendo la mimetización . . . 3
1.4. Objetivo del estudio . . . 3

2.. Antecedentes **5**
2.1. Descripción del método TAMA . . . 6
2.2. Análisis bivariado . . . 7

3.. Materiales y Método **9**
3.1. Columbia Game Corpus . . . 10
3.1.1. Juego de Objetos . . . 10
3.1.2. Anotaciones sobre comportamiento social . . . 10
3.1.3. Extracción de variables acústico/prosódicas . . . 11
3.2. Modificaciones a TAMA . . . 12
3.2.1. Selección de Ventana . . . 13
3.3. Time plots . . . 13
3.4. Medición del Entrainment . . . 15
3.5. Panel de datos . . . 16
3.6. Análisis de regresión . . . 17

4.. Análisis mediante Regresión Lineal Agrupada **19**
4.1. Modelo clásico de Regresión Lineal . . . 20
4.2. Nuestro modelo de regresión . . . 20
4.3. Resultados sobre *entrainment* . . . 21
4.4. Valor absoluto de *entrainment* . . . 21
4.5. Resultados sobre *unsigned entrainment* . . . 24

5.. Análisis mediante Regresión Lineal con Efectos Fijos **29**
5.1. Modelo de Efectos Fijos . . . 30
5.2. Modelo de efectos fijos *within group* . . . 30
5.3. Resultados . . . 31

6.. Conclusiones y trabajo futuro **35**

7.. Apéndice **37**
7.1. Series de Tiempo . . . 38

7.1.1. Procesos estocásticos 39
7.1.2. Estacionariedad 39
7.1.3. Autocorrelación y autocorrelograma 40

8.. Bibliografía **41**

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 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.

1.2. 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.

1.3. 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*.

1.4. 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. ANTECEDENTES

En esta sección describiremos el método TAMA desarrollado en [KDW+08] en el cual hemos basado la métrica de *entrainment* utilizada en el presente trabajo.

2.1. 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 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. En el Apéndice 7.1 describimos más en detalle los conceptos básicos sobre series de tiempo.

Un problema 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 2.1 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 %. En la Sección 3.2.1, describimos la elección del tamaño de ventana que hicimos en base al corpus utilizado.

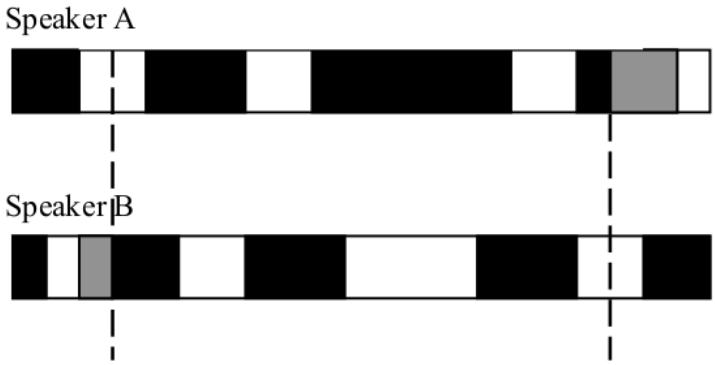

Fig. 2.1: Gráfico de la separación del diálogo en ventanas (Fuente [KDMC08]
)

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:

$$\mu = \sum_{i=1}^{N} f_i d'_i \tag{2.1}$$

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

$$d'_i = \frac{d_i}{\sum_{i=1}^{N} d_i} \tag{2.2}$$

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

Como se ve en la ecuación 2.1, 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 2.1, necesitamos efectuar algún tipo de análisis sobre éstas para obtener una medida del *entrainment*.

2.2. 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 podría 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, utilizaremos 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.

$$r_{AB}(k) = \begin{cases} \dfrac{\sum_{t=k+1}^{n} (A_t - \mu_A)(B_{t-k} - \mu_B)}{\sqrt{\sum_{t=1}^{n} (A_t - \mu_A)^2 \sum_{t=1}^{n} (B_t - \mu_B)^2}} & \text{si } k \geq 0 \\ \dfrac{\sum_{t=-k+1}^{n} (B_t - \mu_B)(A_{t+k} - \mu_A)}{\sqrt{\sum_{t=1}^{n} (A_t - \mu_A)^2 \sum_{t=1}^{n} (B_t - \mu_B)^2}} & \text{si } k < 0 \end{cases} \tag{2.3}$$

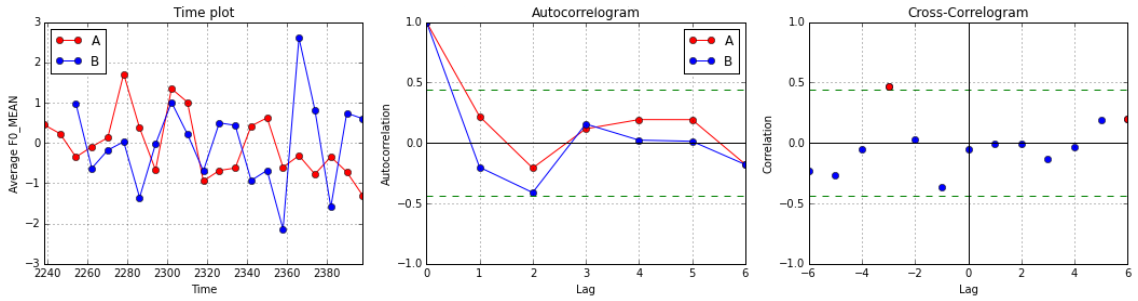

Fig. 2.2: Time-plot producido por TAMA, junto a su autocorrelación y correlación cruzada

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 3.4 comentaremos cómo utilizamos la técnica descripta para la medición del entrainment direccional.

3. MATERIALES Y MÉTODO

En esta sección se describe tanto el corpus de diálogo utilizado en el estudio como así también las modificaciones que efectuamos sobre el método TAMA para medir el *entrainment* de las variables acústico-prosódicas.

3.1. Columbia Game Corpus

Empleamos el Columbia Game Corpus [Gra09] que consiste de 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 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'.

3.1.1. 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 3.1. 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.

3.1.2. 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:

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

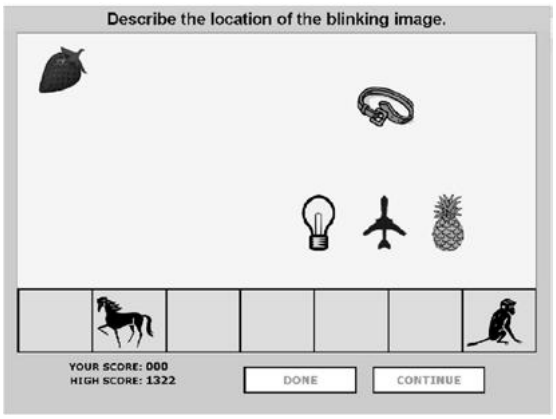

Fig. 3.1: Juego de objetos del Columbia Games

- ¿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*)
- ¿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'.

3.1.3. 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.

Repasemos algunos conceptos que necesitamos para definir las variables acústicas.

- *f0* refiere a la frecuencia fundamental de una onda, que es el recíproco del período de ésta. El *tono* o *pitch* es la percepción que tenemos de la frecuencia fundamental, que nos marca cuán agudos o graves son los sonidos.

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

- *Intensity* refiere al volumen o intensidad de la onda. Ésta mide la amplitud de la onda, y es la percepción de cuán fuerte es el sonido.

- *jitter y shimmer* se refieren, en un intervalo de tiempo, a los desplazamientos de la onda de la verdadera periodicidad y de la amplitud, respectivamente. Están asociadas con la percepción de la calidad de la voz.

- Un *fonema* es la articulación simple de sonidos del habla, tanto de vocales como de consonantes. Ejemplos de fonemas son los sonidos de las letras u, a, s, k en español.

- El *noise-to-harmonics ratio* (abreviado NHR) puede considerarse como una medida de calidad de la voz, que cuantifica la proporción de ruido que hay en ésta.

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

3.2. Modificaciones a TAMA

Al método TAMA descripto en 2.1 le hemos aplicado algunas variaciones, que pasaremos a detallar.

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 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).

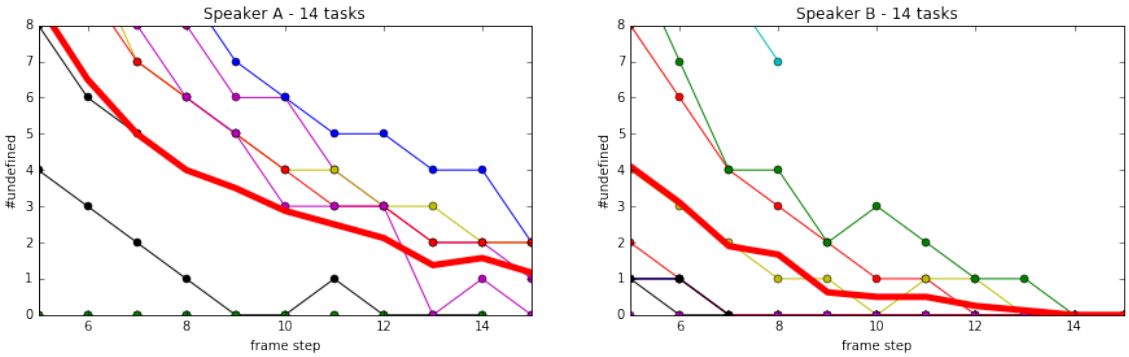

Fig. 3.2: Cantidad de puntos indefinidos en función del step (en segundos) para una sesión en particular, tanto para un interlocutor como para el otro. En rojo se grafica la curva de los promedios

3.2.1. Selección de Ventana

En esta sección discutiremos los parámetros *tamaño de ventana* o *frame size* y el *frame step*. En el trabajo [KDMC08] se menciona una elección de *frame step* y *frame length* de 10s y 20s respectivamente. En el caso de nuestro corpus, queremos buscar los parámetros que mejor se ajustan a éste, manteniendo la superposición del 50 % entre ventanas sucesivas.

¿Qué queremos optimizar? El criterio que elegimos para esto es encontrar un balance entre una ventana 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 hablante que no tienen ninguna participación medible de su parte. Para ver esto, graficamos la cantidad de indefiniciones en función del step tomado, para ver qué forma tenían estas curvas. En la Figura 3.2 podemos ver las indefiniciones en función de los steps para una sesión del corpus. Cada tarea tiene su propia curva, y además graficamos el promedio de todas ellas.

Finalmente, para tener una visión general de lo que ocurría, graficamos una curva promedio de todas las sesiones, que se ilustra en la Figura 3.3. Sobre esta curva aplicamos el "método del codo" para ver si podemos encontrar el valor en el cual la pendiente de las indefiniciones se estanca. Si bien es poco preciso hacer esto, puede observarse que hasta 8-10 segundos hay un fuerte descenso de las indefiniciones, que luego se atenúa. Dado que en general tenemos tareas cortas, preferimos tomar 8 segundos como step, y por ende 16 segundos como largo de ventana.

3.3. 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

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

Tab. 3.1: Tabla de tareas seleccionadas y sus duraciones

adelante. Con esto, no sólo nos interesa la duración de la charla, sino cierta calidad de las series generadas. En la Tabla 3.1 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 (ver apéndice 7.1). En la Figura 3.4 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

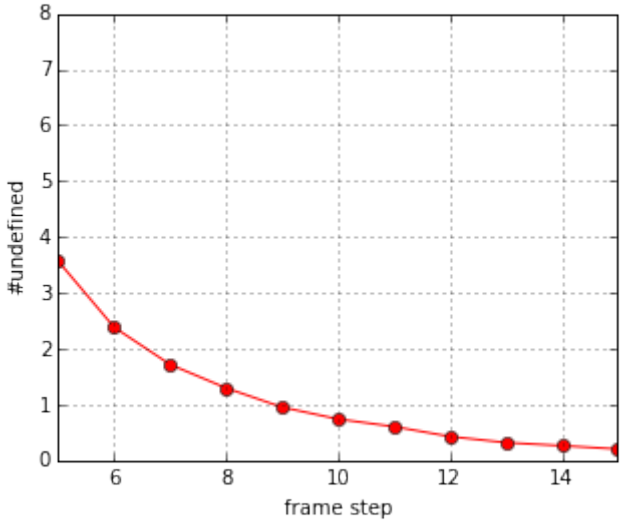

Fig. 3.3: Promedio de cantidad de puntos indefinidos en función del step

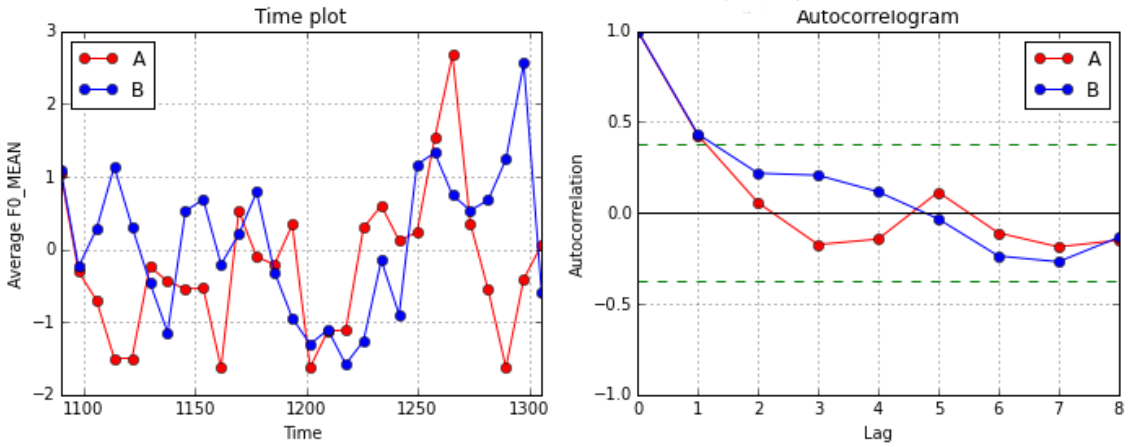

Fig. 3.4: Time-plot generado por el método TAMA, junto a su autocorrelograma

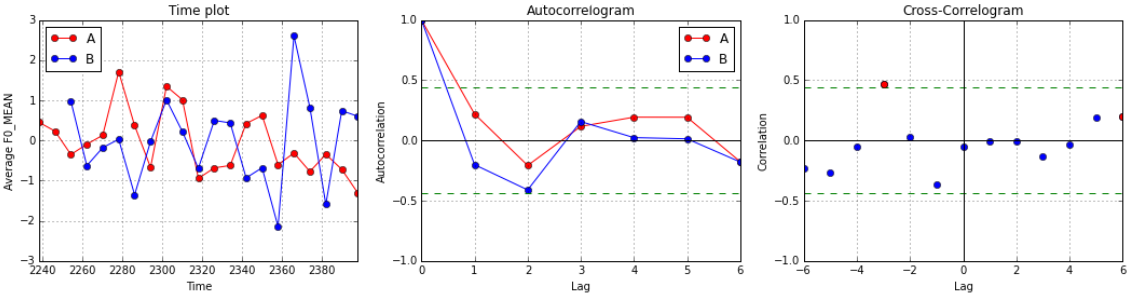

Fig. 3.5: Time-plot generado por el método TAMA, junto a su autocorrelograma y correlograma cruzado

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, como se menciona en la Sección 7.1.3. Esto nos habilita a efectuar el análisis bivariado de las series.

3.4. Medición del Entrainment

Considerando todo lo mencionado en la Sección 2.2, 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 2.3 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 3.5 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:

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

Tab. 3.2: Extracto de la tabla generada para *F0 Mean*

$$\mathcal{E}^{(2)}_{BA} = |\mathcal{E}^{(1)}_{BA}| \tag{3.1}$$

La justificación de la utilización de esta métrica es desarrollada en la Sección 4.4.

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.

3.5. 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.

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 3.3 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 3.3. 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 3.2 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 $speaker = 0$ 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 3.3, no tomamos todas las tareas de todas las sesiones sino aquellas que tienen cierta calidad de diálogo.

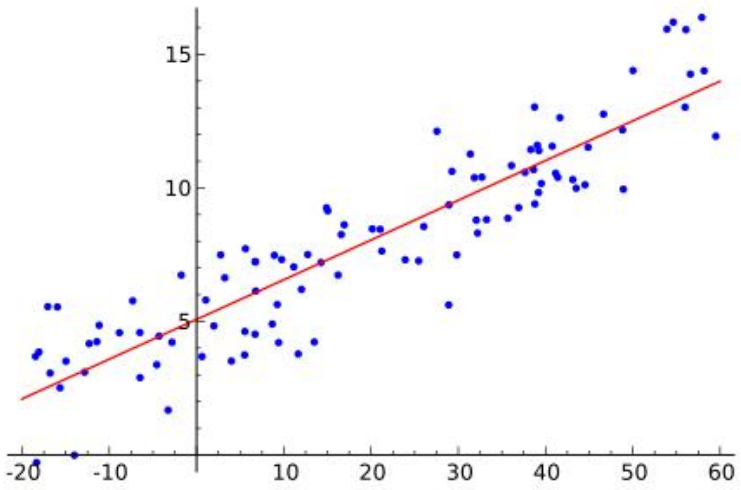

Fig. 3.6: Ejemplo de Regresión Lineal

3.6. 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

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?

Tab. 3.3: Columnas de la tabla generada para ser utilizada en los análisis de regresión lineal

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.

4. ANÁLISIS MEDIANTE REGRESIÓN LINEAL AGRUPADA

En esta sección, mostraremos el primer análisis realizado. Éste consistió en aplicar un modelo de regresión lineal de cada variable social sobre el *entrainment*, sin desagregar los datos por sesión y hablante.

Una variación que usamos esta sección es utilizar como variable independiente el *valor absoluto* del *entrainment*, en base a estudios que sugieren que los interlocutores pueden también *diferenciarse* como un rasgo positivo en la conversación.

4.1. Modelo clásico de Regresión Lineal

En el modelo clásico de regresión lineal, tenemos un conjunto de valores fijos $X_1, X_2, X_3, \ldots, X_n$, que son llamados variables independientes o variables explicativas. Asociado a cada uno de estos valores fijos, tenemos variables aleatorias $Y_1, \ldots, Y_n$. Asumimos, además, que nuestras variables son de la forma

$$Y_i = E[Y|X_i] + u_i \tag{4.1}$$

donde a u_i se la conoce como la perturbación estocástica de la variable. Se asume que el conjunto de $u_1, u_2, \ldots, u_n$ son variables aleatorias idénticamente distribuídas $N(0, \sigma)$

Asumiendo que $E[Y|X_i]$ es una función lineal de X_i; es decir, que existen $\beta_1, \beta_2 \in \mathbb{R}$ que cumplen

$$E[Y|X_i] = \beta_1 + \beta_2 X_i \tag{4.2}$$

obtenemos que

$$Y_i = \beta_1 + \beta_2 X_i + u_i \tag{4.3}$$

Nuestro objetivo es poder entonces conseguir estimadores $\widehat{\beta_1}, \widehat{\beta_2}$ que nos permitan analizar y predecir el comportamiento conjunto de estas variables. A su vez, nos interesa analizar la significancia estadística de estos estimadores: más concretamente, a qué nivel de certeza podemos afirmar que son distintos de 0.

4.2. Nuestro modelo de regresión

Dada una variable acústico-prosódica (por ejemplo, el pitch o la intensidad), queremos investigar la relación entre entrainment y las distintas variables sociales medidas. Sea V una variable social de las enumeradas en la Tabla 3.3. Sean $E_1, \ldots, E_n$ los valores de entrainment para el set de datos que definimos en la Sección 3.5, y sean $V_1, V_2, \ldots V_n$ los valores de la variable social de cada conversación.

Sobre estas variables es que planteamos nuestro modelo de regresión lineal clásica, para analizar qué relación hay tomando como variable "fija" al *entrainment*, y como variable dependiente a la variable social. El problema, entonces, es hallar estimadores $\widehat{\beta_1}, \widehat{\beta_2} \in \mathbb{R}$ de modo que

$$V_i \simeq \widehat{\beta_1} + \widehat{\beta_2} E_i \tag{4.4}$$

Para ello, calculamos los estimadores y efectuamos un análisis de significancia sobre $\widehat{\beta_2}$ para verificar que sea estadísticamente distinto de 0. Esto fue realizado con las funciones que provee el lenguaje R[1].

[1] https://www.r-project.org/

Uno esperaría que un alto *entrainment* se relacione con un alto valor de ciertas variables sociales, por ejemplo la compenetración con el juego o el ayudar a terminarlo. En términos de la ecuación 4.4, esperamos que $\widehat{\beta}_2 \geq 0$. De manera inversa, cuando las variables sociales tienen un carácter negativo de la conversación, esperamos que $\widehat{\beta}_2 \leq 0$.

El modelo de regresión que usamos en este primer análisis se denomina agrupado o *pooled* ya que no distinguimos entre datos provenientes de distintos "grupos" [GP99] y calculamos la regresión lineal agrupando todos los datos disponibles agrupados.

Un problema que surge con este tipo de regresión es que niega todo tipo de *heterogeneidad* de los datos: estos pueden provenir de interlocutores más o menos empáticos, o cuya interacción en el juego se vio influída por factores no medidos en el experimento. Todo esto es descartado, aún cuando puede afectar seriamente el resultado obtenido.

En el siguiente capítulo ahondaremos un poco más en cómo definimos los grupos en nuestro trabajo.

4.3. Resultados sobre *entrainment*

Los resultados de este análisis no fueron interesantes ya que dieron muy pocos valores de $\widehat{\beta}_2$ significativos. En la Figura 4.1 puede verse el gráfico de regresión lineal del *entrainment* contra distintas variables sociales, tomando como variable acústico-prosódica a *F0 Mean*. En las Tablas 4.1 y 4.2 podemos observar las pendientes obtenidas de la regresión, con las estimaciones obtenidas y sus valores de significancia para todas las variables sociales. Se observa que no sólo las pendientes tienen un muy bajo valor absoluto, sino que además ni siquiera tienen los signos que esperábamos en un principio.

En base a lo arrojado por este análisis de regresión, intentamos introducir variaciones en el modelo planteado. La primera, es cambiar la variable explicativa por el valor absoluto del entrainment.

4.4. Valor absoluto de *entrainment*

En la Sección 3.4, definimos una primer métrica del *entrainment* como el valor de la correlación cruzada (en un sentido de los lags) con mayor valor absoluto. Esto puede dar, como resultado, valores positivos entre 0 y 1 a los cuales consideramos como *entrainment*; o bien valores negativos entre -1 y 0, estos considerados como dis-*entrainment*: la divergencia de las variables acústico-prosódicas medidas a través del tiempo.

Este fenómeno de dis-*entrainment* o antimimicry [CB99] refiere al proceso por el cual uno de los hablantes no imita al otro sino más bien todo lo contrario, acentúa alguna diferencia. Si bien estudios de larga data como [BGT73] o [DJ69] lo emparentan con una connotación negativa, [HPH14] y [LBGH15] sugieren que puede entenderse este fenómeno como una conducta de adaptación cooperativa. No sólo eso, sino que este fenómeno de mimetización complementaria podría ser incluso más prevalente que la mimetización a secas [LBGH15].

En base a esto es que decidimos probar alguna medida que capture positivamente el fenómeno de la anti-mimetización de igual manera que con el *entrainment* antes definido. Es decir, esperamos que cuando tengamos o bien *entrainment* o *entrainment* complementario ocurra que tenemos valores altos de variables sociales de carácter positivo. Mutatis mutandis con las variables sociales de connotación negativa.

Int Max	$\widehat{\beta_2}$	Std. Error	t value	Pr(>|t|)
contributes_to_successful_completion	-0.0558	0.1400	-3.985368E-01	0.6906
making_self_clear	0.1454	0.1475	9.854566E-01	0.3255
engaged_in_game	0.0647	0.1178	5.494021E-01	0.5833
planning_what_to_say	0.0864	0.1689	5.115176E-01	0.6095
gives_encouragement	-0.0699	0.1486	-4.706984E-01	0.6383
difficult_for_partner_to_speak	-0.0053	0.1304	-4.057895E-02	0.9677
bored_with_game	0.0083	0.1326	6.230464E-02	0.9504
dislikes_partner	-0.0937	0.1129	-8.305546E-01	0.4072
Int Mean	$\widehat{\beta_2}$	Std. Error	t value	Pr(>|t|)
contributes_to_successful_completion	-0.1541	0.1395	-1.104898E+00	0.2705
making_self_clear	-0.1212	0.1475	-8.214940E-01	0.4123
engaged_in_game	0.0967	0.1176	8.221595E-01	0.4119
planning_what_to_say	-0.2808	0.1677	-1.673747E+00	0.0957
gives_encouragement	0.0092	0.1485	6.184983E-02	0.9507
difficult_for_partner_to_speak	-0.0579	0.1302	-4.443657E-01	0.6572
bored_with_game	-0.0797	0.1324	-6.019681E-01	0.5479
dislikes_partner	-0.0937	0.1127	-8.311282E-01	0.4069
F0 Mean	$\widehat{\beta_2}$	Std. Error	t value	Pr(>|t|)
contributes_to_successful_completion	0.1295	0.1405	9.218693E-01	0.3577
making_self_clear	-0.0394	0.1486	-2.648899E-01	0.7914
engaged_in_game	0.1147	0.1182	9.700134E-01	0.3332
planning_what_to_say	0.0855	0.1698	5.034510E-01	0.6152
gives_encouragement	0.2108	0.1488	1.416825E+00	0.1580
difficult_for_partner_to_speak	-0.0886	0.1310	-6.762104E-01	0.4997
bored_with_game	-0.0518	0.1333	-3.884336E-01	0.6981
dislikes_partner	-0.2228	0.1126	-1.978537E+00	0.0492
F0 Max	$\widehat{\beta_2}$	Std. Error	t value	Pr(>|t|)
contributes_to_successful_completion	-0.1848	0.1387	-1.332296E+00	0.1842
making_self_clear	-0.3568	0.1450	-2.460911E+00	0.0147
engaged_in_game	-0.1325	0.1169	-1.133286E+00	0.2584
planning_what_to_say	-0.1801	0.1676	-1.074476E+00	0.2839
gives_encouragement	-0.2067	0.1472	-1.404358E+00	0.1617
difficult_for_partner_to_speak	0.0583	0.1296	4.497919E-01	0.6533
bored_with_game	0.3085	0.1301	2.370678E+00	0.0187
dislikes_partner	0.0996	0.1122	8.876897E-01	0.3757

Tab. 4.1: Tablas con los resultados de la regresión pooled sobre el *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	Pr(>|t|)
contributes_to_successful_completion	-0.0531	0.1378	-3.854086E-01	0.7003
making_self_clear	0.0235	0.1456	1.611797E-01	0.8721
engaged_in_game	0.0028	0.1161	2.384333E-02	0.9810
planning_what_to_say	0.0359	0.1664	2.154899E-01	0.8296
gives_encouragement	-0.0687	0.1463	-4.697588E-01	0.6390
difficult_for_partner_to_speak	0.0323	0.1284	2.519307E-01	0.8013
bored_with_game	-0.0936	0.1304	-7.178558E-01	0.4737
dislikes_partner	-0.1472	0.1108	-1.328068E+00	0.1856
Fonemas/seg	$\widehat{\beta_2}$	Std. Error	t value	Pr(>|t|)
contributes_to_successful_completion	-0.0593	0.1431	-4.143729E-01	0.6790
making_self_clear	-0.0093	0.1512	-6.156528E-02	0.9510
engaged_in_game	0.1173	0.1202	9.755995E-01	0.3304
planning_what_to_say	0.1062	0.1726	6.151568E-01	0.5391
gives_encouragement	-0.0158	0.1520	-1.037459E-01	0.9175
difficult_for_partner_to_speak	-0.0170	0.1333	-1.275075E-01	0.8987
bored_with_game	-0.1324	0.1352	-9.786726E-01	0.3289
dislikes_partner	-0.0502	0.1155	-4.350299E-01	0.6640
Sílabas/seg	$\widehat{\beta_2}$	Std. Error	t value	Pr(>|t|)
contributes_to_successful_completion	-0.1920	0.1426	-1.347147E+00	0.1794
making_self_clear	-0.2043	0.1505	-1.356908E+00	0.1763
engaged_in_game	0.0054	0.1205	4.493146E-02	0.9642
planning_what_to_say	-0.0520	0.1728	-3.009701E-01	0.7637
gives_encouragement	-0.0407	0.1520	-2.677948E-01	0.7891
difficult_for_partner_to_speak	0.0909	0.1332	6.827016E-01	0.4956
bored_with_game	0.0178	0.1356	1.314632E-01	0.8955
dislikes_partner	0.0460	0.1155	3.980614E-01	0.6910
Jitter	$\widehat{\beta_2}$	Std. Error	t value	Pr(>|t|)
contributes_to_successful_completion	-0.1813	0.1385	-1.309210E+00	0.1919
making_self_clear	0.0281	0.1469	1.913441E-01	0.8484
engaged_in_game	0.1072	0.1168	9.176696E-01	0.3599
planning_what_to_say	-0.1635	0.1675	-9.759850E-01	0.3302
gives_encouragement	-0.0380	0.1476	-2.575414E-01	0.7970
difficult_for_partner_to_speak	-0.0411	0.1295	-3.175009E-01	0.7512
bored_with_game	-0.0164	0.1317	-1.247555E-01	0.9008
dislikes_partner	-0.0308	0.1122	-2.747907E-01	0.7837
Shimmer	$\widehat{\beta_2}$	Std. Error	t value	Pr(>|t|)
contributes_to_successful_completion	-0.0299	0.1407	-2.122533E-01	0.8321
making_self_clear	0.1098	0.1484	7.400752E-01	0.4601
engaged_in_game	-0.0214	0.1185	-1.806035E-01	0.8569
planning_what_to_say	0.0283	0.1698	1.668648E-01	0.8676
gives_encouragement	-0.1702	0.1489	-1.143038E+00	0.2543
difficult_for_partner_to_speak	-0.0035	0.1311	-2.645931E-02	0.9789
bored_with_game	0.0431	0.1332	3.232737E-01	0.7468
dislikes_partner	-0.0299	0.1136	-2.631785E-01	0.7927

Tab. 4.2: Tablas con los resultados de la regresión agrupada sobre el *entrainment* 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 $p < 0{,}01$

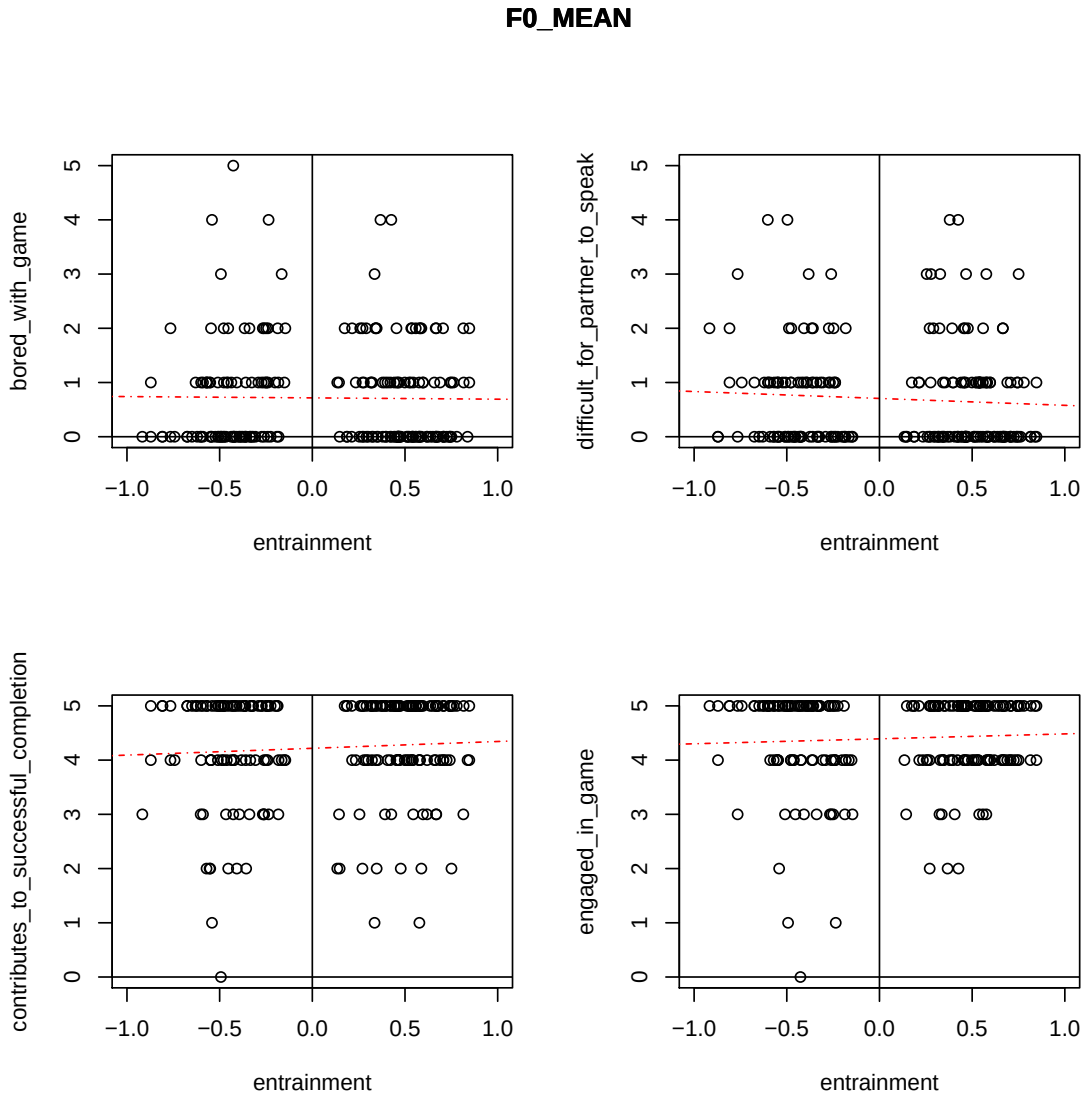

Fig. 4.1: Gráfico de los pares entrainment-variable a/p, junto a la regresión lineal obtenida para *F0_MEAN*

Con este fin, en vez de utilizar sólo el valor de *entrainment* como variable explicativa, efectuamos el mismo análisis pero utilizando la métrica $\mathcal{E}^{(2)}_{AB}$ definida la Sección 3.4, que es el valor absoluto de la métrica anterior. Usar esto permite captar y valorar el *entrainment* complementario de la misma manera que el "positivo" y valorar su relación con las variables sociales medidas. A esta nueva métrica la llamaremos *unsigned entrainment*

4.5. Resultados sobre *unsigned entrainment*

Utilizando esta variable explicativa, los resultados son bastante distintos. En las tablas 4.3 y 4.4 podemos observar que hubo al menos un resultado significativo para todas las variables acústico-prosódicas, exceptuando *Fonemas/seg*.

Casi todos los resultados significativos y positivos de $\widehat{\beta}_2$ son respecto de variables socia-

les de carácter positivo, como *making-self-clear*, *engaged-with-game* y *gives-encouragement*; la notable excepción es *difficult-for-partner-to-speak*, que tiene un carácter negativo pero a su vez $\widehat{\beta}_2 > 0$ en varios casos. El único caso significativo donde $\widehat{\beta}_2 < 0$ es para *bored-with-game*, que era algo justamente esperado.

Habiendo reformulado anteriormente nuestra hipótesis, estos resultados dan indicio de que el valor absoluto del *entrainment* se relaciona con las variables sociales medidas, de manera positiva para aquellas favorables para la conversación, y de manera inversa para aquellas contrarias. Sin embargo, consideramos que en esta asociación influyen factores no medidos dentro de cada conversación, por lo cual planteamos un segundo análisis que contemple esta *heterogeneidad* para analizar mejor cómo interactúan el *entrainment* con los rasgos sociales.

Int Max	Estimate	Std. Error	t value	Pr(>|t|)
contributes_to_successful_completion	-0.1851	0.3852	-4.806290E-01	0.6313
making_self_clear	1.2502	0.3977	3.143535E+00	0.0019
engaged_in_game	0.3906	0.3233	1.207901E+00	0.2285
planning_what_to_say	0.1613	0.4651	3.467699E-01	0.7291
gives_encouragement	0.6711	0.4066	1.650711E+00	0.1003
difficult_for_partner_to_speak	-0.4136	0.3578	-1.155966E+00	0.2490
bored_with_game	0.0760	0.3650	2.081799E-01	0.8353
dislikes_partner	-0.4139	0.3098	-1.335942E+00	0.1830
Int Mean	Estimate	Std. Error	t value	Pr(>|t|)
contributes_to_successful_completion	0.4632	0.4021	1.151825E+00	0.2507
making_self_clear	0.6432	0.4237	1.517942E+00	0.1305
engaged_in_game	0.7620	0.3355	2.271214E+00	0.0242
planning_what_to_say	0.0213	0.4869	4.365416E-02	0.9652
gives_encouragement	0.5379	0.4267	1.260484E+00	0.2089
difficult_for_partner_to_speak	0.6644	0.3729	1.781913E+00	0.0762
bored_with_game	-0.1525	0.3819	-3.992020E-01	0.6902
dislikes_partner	0.4595	0.3241	1.417936E+00	0.1577
F0 Mean	Estimate	Std. Error	t value	Pr(>|t|)
contributes_to_successful_completion	0.2325	0.3918	5.933090E-01	0.5536
making_self_clear	0.1139	0.4141	2.749784E-01	0.7836
engaged_in_game	0.8316	0.3251	2.558367E+00	0.0112
planning_what_to_say	-0.0011	0.4733	-2.364724E-03	0.9981
gives_encouragement	0.4261	0.4153	1.025888E+00	0.3061
difficult_for_partner_to_speak	0.0088	0.3652	2.411215E-02	0.9808
bored_with_game	-0.8747	0.3664	-2.387296E+00	0.0179
dislikes_partner	-0.2077	0.3162	-6.569472E-01	0.5119
F0 Max	Estimate	Std. Error	t value	Pr(>|t|)
contributes_to_successful_completion	0.6593	0.4004	1.646479E+00	0.1012
making_self_clear	0.5479	0.4240	1.292389E+00	0.1977
engaged_in_game	0.5883	0.3369	1.746412E+00	0.0822
planning_what_to_say	0.1313	0.4864	2.700229E-01	0.7874
gives_encouragement	0.4879	0.4266	1.143779E+00	0.2540
difficult_for_partner_to_speak	0.0605	0.3753	1.610613E-01	0.8722
bored_with_game	-0.3913	0.3807	-1.027693E+00	0.3053
dislikes_partner	0.1286	0.3252	3.953927E-01	0.6930

Tab. 4.3: Tablas con los resultados de la regresión pooled sobre el absolute value *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	Pr(>|t|)
contributes_to_successful_completion	0.3012	0.3668	8.212853E-01	0.4124
making_self_clear	0.6912	0.3850	1.795341E+00	0.0741
engaged_in_game	0.3573	0.3083	1.159121E+00	0.2477
planning_what_to_say	-0.6433	0.4411	-1.458137E+00	0.1463
gives_encouragement	0.9573	0.3843	2.490692E+00	0.0135
difficult_for_partner_to_speak	0.2473	0.3417	7.237456E-01	0.4700
bored_with_game	0.1127	0.3478	3.239352E-01	0.7463
dislikes_partner	0.1021	0.2964	3.445479E-01	0.7308
Fonemas/seg	$\widehat{\beta_2}$	Std. Error	t value	Pr(>|t|)
contributes_to_successful_completion	0.4244	0.4041	1.050214E+00	0.2948
making_self_clear	0.4874	0.4266	1.142560E+00	0.2545
engaged_in_game	0.3743	0.3401	1.100468E+00	0.2724
planning_what_to_say	0.0675	0.4891	1.379776E-01	0.8904
gives_encouragement	0.3814	0.4294	8.882870E-01	0.3754
difficult_for_partner_to_speak	-0.2884	0.3768	-7.652857E-01	0.4450
bored_with_game	-0.1753	0.3836	-4.570691E-01	0.6481
dislikes_partner	-0.0253	0.3271	-7.746968E-02	0.9383
Sílabas/seg	$\widehat{\beta_2}$	Std. Error	t value	Pr(>|t|)
contributes_to_successful_completion	0.2603	0.3843	6.774699E-01	0.4989
making_self_clear	0.4636	0.4050	1.144625E+00	0.2537
engaged_in_game	0.6655	0.3205	2.076166E+00	0.0391
planning_what_to_say	0.2054	0.4641	4.425111E-01	0.6586
gives_encouragement	0.7216	0.4054	1.780238E+00	0.0765
difficult_for_partner_to_speak	0.7099	0.3548	2.000706E+00	0.0467
bored_with_game	-0.7740	0.3603	-2.148100E+00	0.0329
dislikes_partner	-0.2765	0.3099	-8.921180E-01	0.3734
Jitter	$\widehat{\beta_2}$	Std. Error	t value	Pr(>|t|)
contributes_to_successful_completion	0.2577	0.3726	6.916180E-01	0.4899
making_self_clear	0.2344	0.3936	5.955030E-01	0.5522
engaged_in_game	0.5409	0.3118	1.734767E+00	0.0843
planning_what_to_say	-0.3620	0.4496	-8.053115E-01	0.4216
gives_encouragement	0.3403	0.3954	8.608149E-01	0.3903
difficult_for_partner_to_speak	-0.1179	0.3473	-3.395874E-01	0.7345
bored_with_game	-0.0256	0.3533	-7.251562E-02	0.9423
dislikes_partner	-0.1089	0.3010	-3.618293E-01	0.7178
Shimmer	$\widehat{\beta_2}$	Std. Error	t value	Pr(>|t|)
contributes_to_successful_completion	0.2636	0.3712	7.101116E-01	0.4784
making_self_clear	-0.0986	0.3925	-2.512623E-01	0.8019
engaged_in_game	0.3814	0.3118	1.223352E+00	0.2226
planning_what_to_say	-0.7361	0.4457	-1.651521E+00	0.1001
gives_encouragement	0.6756	0.3918	1.724198E+00	0.0862
difficult_for_partner_to_speak	0.4077	0.3450	1.181785E+00	0.2386
bored_with_game	-0.5326	0.3501	-1.521345E+00	0.1297
dislikes_partner	-0.2428	0.2996	-8.104076E-01	0.4186

Tab. 4.4: Tablas con los resultados de la regresión agrupada sobre absolute value entrainment 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 $p < 0{,}01$

5. ANÁLISIS MEDIANTE REGRESIÓN LINEAL CON EFECTOS FIJOS

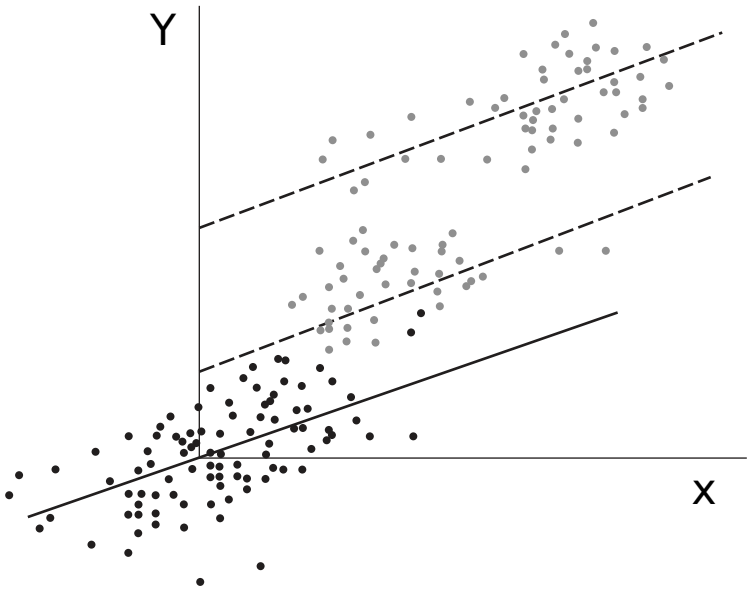

Fig. 5.1: Ejemplo de datos de diferentes sujetos

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 .

5.1. 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.

5.2. 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.

5.3. 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. En las Tablas 5.2 y 5.3 podemos ver el test de coeficientes con las variables sociales significativas resaltadas en distintas tonalidades según su nivel de significancia. Una versión simplificada tabla la podemos ver en la tabla 5.1 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].

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

Tab. 5.1: 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

Tab. 5.2: 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

Tab. 5.3: 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 $p < 0{,}01$

6. CONCLUSIONES Y TRABAJO FUTURO

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 disentrainment 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 [1]. 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.

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

7. APÉNDICE

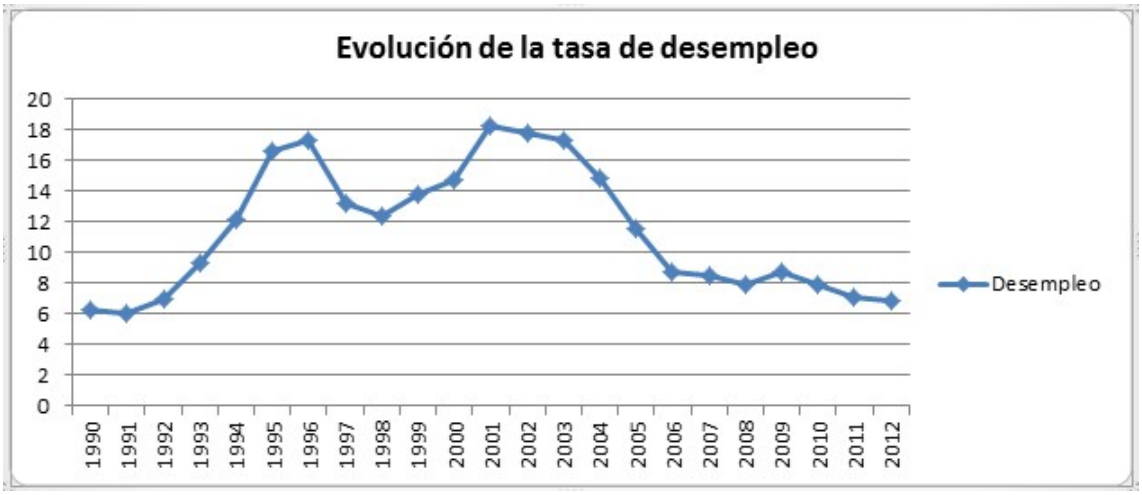

Fig. 7.1: Gráfico de serie de tiempo de la evolución del desempleo en Argentina

7.1. 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, como por ejemplo Economía, Ciencias de la Atmósfera, y otros.

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 [Cha13].

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 7.1 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 esto 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.

7.1.1. 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.

7.1.2. 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.

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

$$E[X_t] = E[X_{t+\tau}] \tag{7.1}$$
$$Var[X_t] = Var[X_{t+\tau}] \tag{7.2}$$
$$Cov(X_s, X_t) = Cov(X_{s+\tau}, X_{t+\tau}) \tag{7.3}$$

Las ecuaciones 7.1 y 7.2 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 7.1, 7.2, 7.3

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

7.1.3. Autocorrelación y autocorrelograma

Una herramienta importante para el análisis de las series de tiempo es la función de autocorrelación [Cha13, p22]. Esta función mide la correlación entre las observaciones a diferentes distancias o lags. Estos coeficientes son de ayuda para analizar el modelo probabilístico de la serie de tiempo.

Dado un conjunto observaciones $x_1, \ldots, x_N$, la fórmula de la función de autocorrelación muestral se define como:

$$r_k = \frac{\sum_{t=1}^{N-k} (x_t - \bar{x})(x_{t+k} - \bar{x})}{\sum_{t=1}^{N} (x_t - \bar{x})^2} \tag{7.4}$$

Una ayuda importante a la hora de evaluar estos coeficientes es graficar r_k en función de k: a este gráfico lo llamaremos autocorrelograma. En [Cha13, p25] se enumeran consejos a la hora de interpretar este gráfico. Vale destacar que el hecho de que r_k descienda rápidamente a 0 es un indicio de estacionariedad de la serie, requisito indispensable para efectuar el análisis bivariado del presente trabajo.

8. BIBLIOGRAFÍA

Bibliografía

[BGT73] Richard Y Bourhis, Howard Giles, and Henri Tajfel. Language as a determinant of welsh identity. *European Journal of Social Psychology*, 3(4):447–460, 1973.

[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.

[DJ69] James M Dabbs Jr. Similarity of gestures and interpersonal influence. In *Proceedings of the annual convention of the American Psychological Association*. American Psychological Association, 1969.

[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.

[HPH14] Patrick GT Healey, Matthew Purver, and Christine Howes. Divergence in dialogue. *PloS one*, 9(6):e98598, 2014.

[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.

[LBGH15] Rivka Levitan, Štefan Benuš, Agustın Gravano, and Julia Hirschberg. Acoustic-prosodic entrainment in slovak, spanish, english and chinese: A cross-linguistic comparison. In *16th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, page 325, 2015.

[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.