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

Juan Manuel Pérez

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

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

1. Introduction

Los sistemas de diálogo humano-computadora son cada vez más frecuentes, y sus aplicaciones comprenden una amplia gama de rubros: desde aplicaciones móviles, motores de búsqueda, juegos o tecnologías de asistencia para ancianos y discapacitados. Si bien es cierto que estos sistemas logran captar la dimensión lingüística de la comunicación humana, tienen un déficit importante a la hora de procesar y transmitir el aspecto superestructural de la comunicación oral, que radica en el intercambio de afecto, emociones, actitudes y otras intenciones de los participantes. Este problema puede verse en cualquier sistema que interactúe sintetizando lenguaje humano, como 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.

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. Un manejo adecuado de estas componentes es lo que, hoy día, distingue una voz humana de una artificial. Esta carencia de habilidad sobre la prosodia conlleva cierta dificultad en la interacción con agentes conversaciones, que suelen ser calificados como "mecánicos" o "extraños" en su forma de comunicarse. [RBL$^+$06,WRWN05]

En pos de mejorar el entendimiento entre agentes conversacionales y sus usuarios, resulta de vital importancia poder entender y modelar las variaciones

prosódicas de la comunicación oral. Esto se traduciría tanto en una mejor apreciación de lo que quiere comunicar el usuario, como en una mayor naturalidad de la voz sintetizada por el agente.

1.1. Mimetización

Se ha observado 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 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* 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[NGH08,GBLH15]. En nuestro caso, nos interesa principalmente el *entrainment* de la prosodia.

1.2. Midiendo la mimetización

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

Un problema común que surge a la hora de calcular estas métricas es el hecho de que las conversaciones no están alineadas en el tiempo, ni se dan en turnos de duración constante. Nos preguntamos entonces qué partes del diálogo de un hablante deberían compararse con qué otras partes de su par. Un enfoque de comparar interlocuciones uno a uno es demasiado simple y no captura situaciones de diálogo reales, mucho más dinámicas y con solapamiento casi constante.

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

1.3. Objetivo del estudio

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

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

2. Materiales y Método

2.1. Columbia Game Corpus

Para el estudio, 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. Cada sesión duró aproximadamente 45 minutos, totalizando 9 horas de diálogos, 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'.

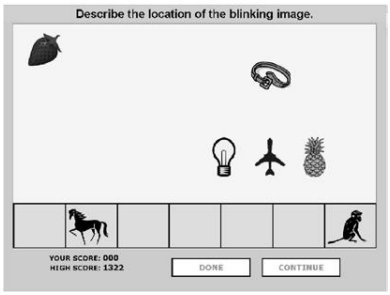

Figura 1. Juego de objetos del Columbia Games

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

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

2.1.1. Anotaciones sociales y acústico/prosódicas 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, listadas en la Figura 1

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.

A su vez, los valores de las variables acústico-prosódicas listadas en la Figura 2 fueron medidos para cada segmento del habla del corpus. Para entender mejor a qué se refieren estas variables, repasamos a continuación algunos conceptos:

- *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.

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

Nombre	Pregunta
contributes-to-completion	¿el sujeto contribuye para el éxito del equipo?
engaged-with-game	¿el sujeto parece comprometido con el juego?
making-self-clear	¿el sujeto se expresa correctamente?
planning-what-to-say	¿el sujeto piensa lo que va a decir?
gives-encouragement	¿el sujeto alienta a su compañero?
difficult-for-partner-to-speak	¿el sujeto le hace difícil hablar a su compañero?
bored-with-game	¿el sujeto está aburrido con el juego?
dislikes-partner	¿al sujeto no le agrada su compañero?

Cuadro 1. Preguntas sobre las percepciones sociales realizadas a los anotadores

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

Cuadro 2. Variables acústico-prosódicas medidas

- *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.

3. Descripción del método TAMA

En [KDW+08] se introdujo un método novedoso para el análisis del *entrainment* acústico-prosódico. Esta técnica consiste, a grandes rasgos, en armar dos series de tiempo para cada uno de los interlocutores y luego utilizar herramientas de análisis sobre las series construídas. Uno de los problemas que resuelve esta técnica es el del alineamiento: si intentásemos comparar cada segmento del habla (*elocución* o *utterance*) con otros, ¿cómo los alinearíamos? Al introducir

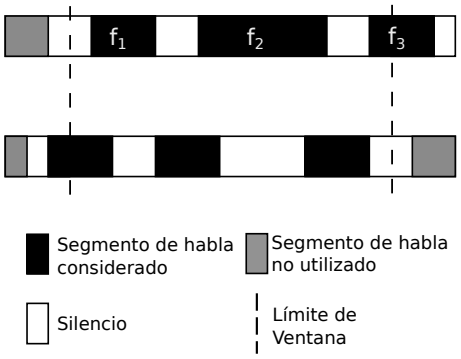

Figura 2. Gráfico de la separación del diálogo en ventanas. Fuente [KDMC08]

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 tengan intersección con la ventana dentro de cada ventana; aquellos que atraviesen los límites de las ventanas serán cortados para que se mantengan dentro de éste. En la Figura 2 se ilustra el proceso: las líneas punteadas marcan los límites de la ventana, los intervalos coloreados en negro los segmentos de habla, y en gris los segmentos cortados.

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 mediante el siguiente cálculo:

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

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

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

Una vez obtenidas las series de tiempo respectivas, una posible medida del *entrainment* se puede obtener midiendo cuánto influye una serie sobre otra, considerándolas a ambas como parte de un sistema donde ambas interactúan. Este *entrainment*, entonces, sería direccional: queremos medir cuánto influye el

interlocutor A sobre el interlocutor B y viceversa. Puede darse el caso en que ambos tengan fuerte interacción, en tal caso hablamos de *feedback*.

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

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

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.

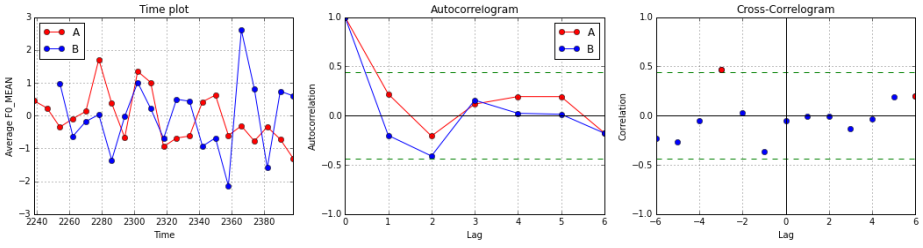

Figura 3. Time-plot producido por TAMA, junto a su autocorrelación y correlación cruzada

3.1. TAMA sobre Columbia Games

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).

Además se modificó el tamaño de la ventana para ajustarlo a nuestro corpus. En vez del step de 10" y tamaño de 20", optamos por 8 y 16 segundos respectivamente luego de efectuar un análisis con el fin de encontrar un balance entre el tamaño de la ventana y la cantidad de indefiniciones. Considerando estos parámetros, se procedió a calcular las dos series de tiempo para cada conversación y cada variable acústico-prosódica. De estas tareas, sólo nos quedamos con aquellas que tengan al menos 5 puntos definidos para cada serie, de manera que tenga sentido poder calcular la correlación cruzada más adelante. Con esto, no sólo nos interesa la duración de la charla, sino cierta calidad de las series generadas.

Finalmente, definimos una primer métrica de *entrainment* $\mathcal{E}^{(1)}_{AB}$ como el valor de $r_{AB}(k)$ con mayor valor absoluto, dado $k \leq 0$. Análogamente lo definimos para $\mathcal{E}^{(1)}_{BA}$. Además, definimos una segunda métrica $\mathcal{E}^{(2)}_{BA}$, como el valor absoluto de la primera, es decir $\mathcal{E}^{(2)}_{BA} = |\mathcal{E}^{(1)}_{BA}|$. 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.

4. Análisis de regresión

Luego de construir las series de tiempo para cada una de las conversaciones (y cada una de las variables acústico-prosódicas) que seleccionamos anteriormente, 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].

Utilizamos *regresión de efectos fijos* [GP99, chap 16], un modelo que ayuda a controlar la heterogeneidad no observada cuando ésta es constante en el tiempo para cada sujeto del sistema. Asumimos que estos factores son inherentes a la conversación entre el hablante y su interlocutor, y por este motivo, definimos los sujetos 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. Resultados

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

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

Cuadro 3. 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 − − −

- *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].

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 dis-entrainment con valores negativos. Se aplicó análisis de regresión sobre esta métrica, y los resultados que dio no fueron significativos. En segundo lugar, construimos una métrica que valora de igual manera el *entrainment* y el dis-*entrainment*, de acuerdo a trabajos previos que sugerían que el segundo fenómeno puede considerarse en algunas circunstancias como un mecanismo de adaptación cooperativa. Al efectuar el análisis de regresión sobre

esta métrica, los resultados fueron significativos y consistentes con la hipótesis planteada de que el *entrainment* se relaciona positivamente con características sociales favorables de la conversación, mientras que lo hace de manera inversa con aquellas negativas.

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

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

A pesar de que los resultados son prometedores, siguen siendo preliminares y su robustez requiere de más validaciones. Como trabajo futuro, proponemos reproducir estos análisis sobre otros corpus de habla, como por ejemplo Switchboard [2]. 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.

Referencias

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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