

Estudio de Turn-Taking en conversaciones entre humanos y sistemas interactivos de diálogo

Alumna: Claudia A Jul Vidal

Director: Agustín Gravano

Departamento de Computación

Facultad de Ciencias Exactas y Naturales

Universidad de Buenos Aires

Buenos Aires, Noviembre 2013

Índice general

Resumen	2
Agradecimientos	3
1. Introducción	4
2. Materiales	6
2.1. Cuerpo de Datos Let's Go!	6
2.2. Unidad de Análisis	7
2.3. Extracción de atributos	8
2.4. Tipos de cambio de turno	11
3. Indicadores individuales de cesión del turno	13
3.1. Técnicas estadísticas	14
3.1.1. Distribución estadística de los datos	14
3.1.2. Tests estadísticos	15
3.2. Entonación Final	16
3.3. Duración del IPU	18
3.4. Tasa del Habla	20
3.5. Niveles de intensidad y tono	24
3.6. Completitud gramatical	27
3.7. Calidad de la voz	29
4. Señales Complejas de Cesión del Turno	34
4.1. Algoritmo para determinar la presencia de indicadores	35
4.2. Análisis	36
5. Conclusión y Trabajo Futuro	39

Resumen

En los últimos años los sistemas de conversión de texto al habla y de reconocimiento de voz han tenido un marcado crecimiento. Este comportamiento también se vio reflejado en los distintos campos de las tecnologías del habla. Con todo esto los sistemas interactivos de voz han incrementado no sólo sus funcionalidades sino también su complejidad. Con el fin de lograr un comportamiento amigable y natural, este trabajo presenta evidencia de que los indicadores de fin del turno encontrados en trabajos previos para diálogos entre humanos, también están presentes y se comportan de manera similar en diálogos entre sistemas interactivos de voz y sus usuarios. Mediante la descripción de dichos indicadores de fin del turno y su comportamiento, se logra una base teórica para una futura implementación que mejore la interacción entre los sistemas interactivos de voz y sus usuarios.

Agradecimientos

Mucha Gente merece mi gratitud por hacer posible esta tesis.
Primero a mi director Agustín Gravano, por su increíble dedicación.
A mi mamá y mi hermano, por acompañarme a lo largo de esta carrera.
A Martín, gracias por estar siempre a mi lado.
A mis amigos y amigas, esos hermanos que la vida me regalo.
Y en especial a mi papá, te llevo conmigo en mi corazón siempre.

Capítulo 1

Introducción

En los últimos años los sistemas de conversión de texto al habla (text-to-speech, TTS) y de reconocimiento de voz (*automatic speech recognition*, ASR) han tenido un marcado crecimiento. Este comportamiento también se vio reflejado en los distintos campos de las tecnologías del habla. Con todo esto los sistemas interactivos de voz (*interactive voice response*, IVR) han incrementado no sólo sus funcionalidades sino también su complejidad.

Se denomina sistema interactivo de diálogo al sistema informático que interactúa con el ser humano, empleando la comunicación verbal en un lenguaje natural como el castellano o el inglés. Estos sistemas pueden encontrarse en todo tipo de servicios telefónicos de atención al cliente, y en un futuro no muy lejano prometen revolucionar las interfaces usuario-computador.

Sin embargo y aún a pesar del gran crecimiento de este tipo de tecnología, todavía son considerados confusos e intimidantes (Ward et al., 2005; Raux et al., 2006). Esta categorización se debe principalmente a los problemas de coordinación que presenta la dinámica del diálogo entre el usuario y el sistema. Uno de los métodos más utilizados para determinar cuándo el usuario ha finalizado sus dichos y queda a la espera de una respuesta, consiste en esperar por una pausa lo suficientemente larga. Claramente, esta estrategia es raramente usada en un diálogo entre humanos, en el cual se utilizan otros tipos de indicadores que determinan este tipo de comportamiento. Estos indicadores se pueden catalogar entre aquellos visuales y verbales. El estudio realizado en este trabajo se realiza sobre los indicadores verbales entre los cuales podemos encontrar la sintaxis, la prosodia y la acústica.

En trabajos previos se ha mostrado la existencia de indicadores auditivos cuya presencia denota que el hablante ha concluido su discurso y queda en

silencio a la espera de una respuesta. A estos se los conoce como indicadores de cesión del turno, del inglés *turn-yielding cues* (Duncan, 1972; Ford and Thompson, 1996; Gravano, 2009). Dentro de las mismas se puede encontrar: la calidad de la voz, la tasa del habla, la intensidad, la completitud sintáctica, la duración de las frases o unidades prosódicas, la entonación y el tono. A su vez, también se menciona que, en un diálogo entre dos personas, a mayor número de indicadores presentes en los dichos del hablante, mayor es la probabilidad de que al terminar sus dichos el hablante ceda el turno conversacional a su interlocutor (Gravano, 2009). Si se pudiera mostrar que estos indicadores se comportan de manera similar en un diálogo entre un sistema IVR y un usuario, las mismas podrían utilizarse para naturalizar la dinámica del mismo.

Este estudio tratará de mostrar si los indicadores de fin del turno también están presentes en diálogos entre un sistema IVR y su usuario. De obtener un comportamiento similar al mostrado en Gravano (2009), se pasará a ver el comportamiento de estos indicadores de manera conjunta. Este análisis se desarrolla en los capítulos 3 y 4 de este documento. Los datos utilizados para este trabajo se extrajeron de Let's Go!, un sistema IVR utilizado para obtener información sobre recorridos y horarios de líneas de colectivos en Pittsburgh. El capítulo 2 describe cómo se extrajeron y cómo se transformaron los datos de Let's Go!.

Capítulo 2

Materiales

En este capítulo describimos el cuerpo de datos empleado en esta Tesis, así como las anotaciones realizadas y los atributos extraídos.

2.1. Cuerpo de Datos Let's Go!

El cuerpo de datos utilizado para este estudio se extrajo de la utilización del sistema interactivo de diálogo llamado Let's Go!, desarrollado y mantenido por Carnegie Mellon University (Raux and Eskenazi, 2008). Let's Go! es un sistema IVR, conectado a la línea central de la Port Authority Transit (PAT) de la ciudad de Pittsburgh que funciona de lunes a viernes de 7pm-7am y fines de semanas/feriados de 6pm-8am. Let's Go! provee los horarios y recorridos de las distintas líneas de colectivos que circulan por el condado de Allegheny. Para utilizar este servicio se debe llamar al número 412-268-3526. Fuera de los horarios mencionados el servicio es provisto por un grupo de operadores humanos.

Al ser Let's Go! un servicio que se ofrece dentro de los Estados Unidos, el sistema se maneja completamente en inglés y su particularidad es que está pensado para interactuar con un amplio espectro de la población, como personas de avanzada edad o aquellas cuya lengua madre no es el inglés.

Para este trabajo se tomaron 233 llamadas al azar entre los años 2007 y 2009 obtenidas como parte del Spoken Dialogue Challenge (Black and Eskenazi, 2009). El sistema provee entre sus logs de salidas, una transcripción de dichas llamadas en formato HTML, además de una grabación en formato RAW de las mismas. Para estudiar lo que ocurre a lo largo de las conversacio-

nes que tienen los usuarios con el sistema, se debieron tomar estos archivos y extraer de ellos solo los datos necesarios y reescribirlos en una estructura más sencilla de analizar.

Como se mencionó Let's Go! crea un archivo HTML por cada llamada que recibe, un ejemplo de esto puede verse en la figura 2.1. Allí se muestra información relacionada con el diálogo que tuvo el sistema con el usuario que efectuó la llamada. Algunos de estos datos son: la cantidad de turnos que se realizaron a lo largo de la llamada; la duración de la misma; la transcripción de los dichos del sistema (prompt) y lo que interpretó este último de lo dicho por el usuario (speech); entre otros. El diálogo se transcribe agrupando un turno del sistema y el turno del usuario que corresponde a la respuesta del anterior, y así continuamente hasta que este último finaliza la llamada.

LetsGoPublic-20070501-000-gloss

00:08:10.88 on 01-MAY-2007

[\[index\]](#) 001

Duration: 93.03 sec [MAXdelta: 10.19] Turns: 12

The [folder](#). The raw [logfile](#). The [dialog](#) logfile. The [helios](#) logfile. The [session audio](#) file ([raw](#)) The [annotations](#) file.

000 0.2 s	<p>prompt Welcome to the CMU Let's Go bus information system.</p> <p>prompt To get help at any time, just say Help or press zero.</p> <p>prompt What can I do for you?</p> <p>speech FRONT. NEXT .BUS. .FROM. IF [raw] [0.00] [WER:~]</p> <p>parse [no_parse]</p>
001 12.9 s	<p>prompt For example, you can say, when is the next 28X from DOWNTOWN to THE AIRPORT? or I'd like to go from MCKEESPORT to HOMESTEAD tomorrow at 10 a.m..</p> <p>speech BUS. NEXT .BUS. .FROM. FROM EIGHTH AVENUE TO DUQUESNE VILLAGE</p> <p>[raw] [0.09] [WER:~]</p> <p>parse PlaceInformation[2_DeparturePlace] (FROM [stop_name] ([covered_place] ([registered_stop] ([stop_part] (EIGHTH) AVENUE)))) [3_ArrivalPlace] (TO [stop_name] ([covered_place] ([ambiguous_covered_place] ([ambiguous_covered_part] (DUQUESNE)))))))</p>
002 30.1 s	<p>prompt Leaving from EIGHTH AVENUE . Did I get that right?</p> <p>speech YES [raw] [0.89] [WER:~]</p> <p>parse Generic[Generic] ([Yes] (YES))</p>

Figura 2.1: Archivo de Salida - Let's Go! - HTML

2.2. Unidad de Análisis

Definimos como unidad de análisis a la unidad inter-pausal (IPU, del inglés *inter-pausal unit*) una secuencia de palabras de un mismo hablante

rodeadas por silencios mayores a los 50 milisegundos. Asimismo definimos el concepto de *turno*, como una secuencia de IPU de un mismo hablante de manera tal que entre ellos no existen dichos del interlocutor.

Para extraer las IPU de los archivos HTML que proporciona Let's Go! se desarrolló un parser que genera un archivo de texto plano por cada IPU del usuario. En estos archivos se utilizó el carácter # para denotar los silencios que delimitan el IPU, figura 2.2. Estos archivos serán utilizados luego junto con el audio para determinar valores acústicos de los IPU.

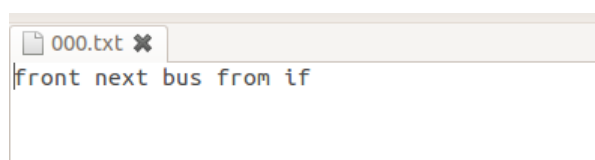


Figura 2.2: Archivo de Salida - Parser - TXT

2.3. Extracción de atributos

Paralelamente a la generación de los archivos de texto mencionados, se debió transformar del formato RAW a WAV a los archivos de audio donde se grabó la llamada. Una vez obtenidos el texto y audio que formaban parte del output del sistema, se utilizó un programa conocido como PRAAT, Boersma and Weenink (2001), para generar un archivo del tipo TEXTGRID donde se reúnen los datos de ambos. PRAAT, del holandés “hablar”, es un software gratuito para el análisis científico del habla, muy usado en lingüística.

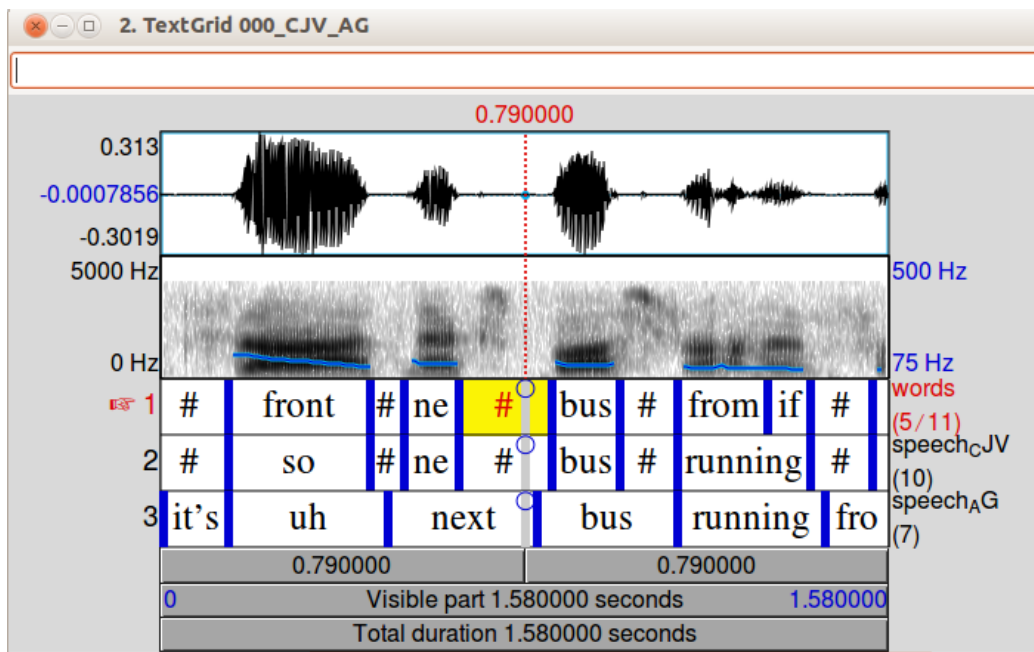


Figura 2.3: Interfaz de PRAAT: Forma de habla, espectrograma y tres niveles de anotaciones manuales

La figura 2.3 muestra la interfaz de PRAAT. Este software permitió trabajar en conjunto con el texto y el audio para poder alinearlos manualmente, generando luego una salida donde se puede apreciar los tiempos, silencios y dichos del hablante correctamente. A lo largo de la alineación manual no solo se trató de hacer coincidir los silencios y palabras con el audio, sino también se corrigieron algunas de las transcripciones hechas por el sistema. Let's Go! puede interpretarse erróneamente lo dicho por el usuario, por lo cual se debieron realizar algunas correcciones en los archivos utilizados para el estudio, en primera instancia por la autora y en una segunda revisión por el director de este trabajo.

```

File type = "ooTextFile"
Object class = "TextGrid"

xmin = 0
xmax = 1.708
tiers? <exists>
size = 3
item []:
  item [1]:
    class = "IntervalTier"
    name = "words"
    xmin = 0
    xmax = 1.708
    intervals: size = 7
    intervals [1]:
      xmin = 0
      xmax = 0.06521432114222425
      text = "# "
    intervals [2]:
      xmin = 0.06521432114222425
      xmax = 0.18215020345267618
      text = "when"
    intervals [3]:
      xmin = 0.18215020345267618
      xmax = 0.35011265258950713
      text = "is"
    intervals [4]:
      xmin = 0.35011265258950713
      xmax = 0.46221890608220595
      text = "the"
    intervals [5]:
      xmin = 0.46221890608220595
      xmax = 0.9751880962126501
      text = "previous"
    intervals [6]:
      xmin = 0.9751880962126501
      xmax = 1.1771682565670671
      text = "bus"

```

Figura 2.4: TextGrid File

Un ejemplo de los archivos generados por PRAAT, *TEXTGRID*, se ve en la figura 2.4. Allí se observan datos como la duración de cada silencio o palabra dicha por el hablante, así también como la transcripción de la IPU. Recordemos que cada archivo representa una IPU del usuario del sistema. PRAAT también fue utilizado para extraer valores como el tono, la intensidad, proporción de frames sonoros, jitter, shimmer y relación ruido-armónico de las distintas IPUs.

Una vez obtenidos estos archivos, se codificaron una serie de scripts en PERL con los cuales se obtuvo la base de datos con los valores de las IPUs que posteriormente se utilizaron para estudiar la presencia de los indicadores de cesión del turno. Entre los atributos de las IPUs que se extrajeron

se encuentran: cantidad de sílabas y duración de la IPU; nivel de intensidad (medido en decibeles); nivel tonal (o *pitch level*, medido en hertz); entonación final; tasa del habla (medida en palabras por segundo y en sílabas por segundo); y *shimmer*, *jitter* y relación ruido-armónico (tres medidas de la perturbación de la onda sonora). El nivel tonal se estimó mediante la frecuencia fundamental de la señal, la cual fue computada con el método de autocorrelación. La entonación final se aproximó mediante la pendiente de un modelo lineal ajustado al nivel tonal, y puede ser ascendente, descendente o sostenida. Todos los atributos acústicos se computaron sobre la totalidad de la IPU, y también sobre los últimos 500, 300 y 200 milisegundos de la misma.

Por otro lado previo a los estudios estadísticos se debió estandarizar los valores del tono. Esto se debe a que los valores del tono de los hombres están en un rango distinto al de las mujeres. El rango del tono de las mujeres es aproximadamente 75-500Hz, y para los hombres 50-300Hz. Para poder comparar los valores de tono de voz sin importar el género del hablante se debió normalizar los datos obtenidos. Para ello definimos la siguiente transformación lineal: $tono_normalizado = K * tono_original + D$ donde K , D son tales que $500K + D = 300$, $75K + D = 50$. De igual manera, se puede aplicar esta transformación a la media y la pendiente del tono de voz: $media_tono_normalizada = K * media_original_tono + D$ y $pendiente_normalizada = K * pendiente_original$, debido a que $(k * f(x))' = k * f'(x)$.

2.4. Tipos de cambio de turno

La base de datos que finalmente se obtuvo consta de un total de 490 *IPUs*. Las 490 *IPUs* fueron catalogadas manualmente como HOLD, SWITCH y UNDEF.

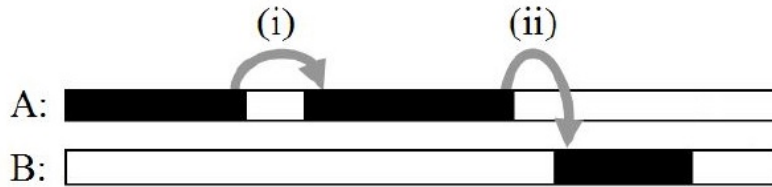


Figura 2.5: Definición de Cambio de Turno: Los segmentos negros representan los dichos y el silencio los blancos; (i) HOLD,(ii) SWITCH, cambio sin superposición.

Se define a una IPU como HOLD cuando es seguida luego de un corto silencio por otra IPU del mismo hablante. En cambio se dice que una IPU es SWITCH, cuando la misma es la última de los dichos del hablante, es decir que luego de un silencio comienza a hablar el interlocutor sin interrumpir. Por último se define como UNDEF a las IPUs que, o bien su audio termina de manera abrupta, o bien corresponde a otras transiciones, como superposiciones, por ejemplo. Para este trabajo sólo se tomaron en cuenta aquellas IPUs catalogadas como HOLD o SWITCH.

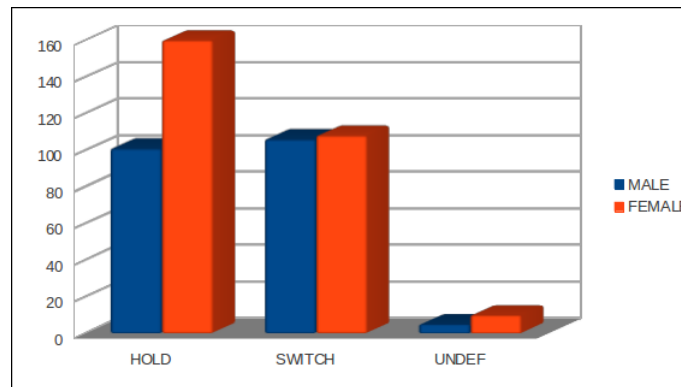


Figura 2.6: Distribución de los Datos

Del cuerpo de datos se extrajeron 490 IPU de las cuales 261 han sido catalogadas como HOLD, 214 como SWITCH y 15 como UNDEF, como se observa en la figura 2.6.

Capítulo 3

Indicadores individuales de cesión del turno

Se define como *indicadores de cesión del turno* (en inglés, *turn-yielding cues*) a las características que pueden ser detectadas en los dichos de un hablante inmediatamente antes del final de sus turnos conversacionales. Estas características se presentan como un cambio en el tono de voz, en la duración y en la calidad de la voz, entre otras que serán explicadas en detalle a lo largo de este capítulo. Se decidió llamarlos indicadores ya que indican el final de un turno.¹ Se ha mostrado en distintos estudios que a mayor número de estos indicadores, mayor es la probabilidad de que el hablante ceda su turno al finalizar sus dichos (Gravano, 2009; Gravano and Hirschberg, 2011; Duncan, 1972; Hjalmarsson, 2011). Estos trabajos han sido realizados siempre dentro del marco de un diálogo entre dos personas. Sin embargo, la presencia de estos indicadores no siempre implica que el oyente tome la palabra: si así lo desea puede permanecer en silencio o bien alentar al interlocutor para que continúe hablando. Si este mecanismo de toma del turno es usado apropiadamente, el oyente podrá comenzar a hablar en respuesta a la presencia de los indicadores de cesión emitidas por el hablante, sin tener necesidad a esperar a un silencio lo suficientemente largo.

Gravano (2009) identificó siete indicadores de cesión del turno en conversaciones colaborativas entre dos personas. Estos son: la entonación; la duración de la IPU; la tasa del habla; la intensidad; el tono de voz; la completitud

¹Por simplicidad, a lo largo de esta tesis muchas veces hablamos de un indicador x para referirnos al indicador dado por un cambio determinado en el valor de la variable x .

sintáctica y la calidad de la voz. A lo largo de este capítulo se tratará de determinar la existencia de estos indicadores en las IPU de los usuarios del sistema IVR. Es decir, se busca evidencia empírica de que los resultados descriptos para los diálogos entre humanos son válidos también para diálogos entre un humano y una computadora.

3.1. Técnicas estadísticas

En esta sección se describen los tests estadísticos empleados para el análisis de los indicadores individuales de sesión del turno.

3.1.1. Distribución estadística de los datos

Previo a realizar los estudios estadísticos se debió verificar si los datos obtenidos poseían una distribución normal. Para ello se utilizó el test estadístico Shapiro-Wilk, el cual es utilizado para contrastar la normalidad de un conjunto de datos. Para aplicarlo se debe partir de una hipótesis nula donde el cuerpo de datos estudiado proviene de una población normalmente distribuida y una hipótesis alternativa donde no lo es. La hipótesis nula es rechazada cuando W es lo suficientemente pequeño.

En la figura 3.1 se puede ver que los valores de W para casi todas las variables es cercano a 0, por lo que no podemos suponer que el cuerpo de datos de este trabajo sigue una distribución normal, para muestras variables de estudio. En consecuencia, deberemos emplear tests estadísticos no paramétricos.

W-VALUE	CUE	W-VALUE	CUE	W-VALUE	CUE
2.20E-016	NUM WORDS	2.87E-016	INTENSITY STDEV 200	2.20E-016	SLOPE PITCH 300
2.20E-016	IPU DURATION	9.54E-016	PITCH MAX 200	2.20E-016	SLOPE STYPCH 300
1.58E-007	INTENSITY MAX	3.41E-015	PITCH MEAN 200	2.20E-016	NOISE TO HARMONICS_RATIO 300
0.001139	INTENSITY MEAN	2.20E-016	PITCH MIN 200	2.20E-016	SAL JITTER 300
1.21E-009	INTENSITY MIN	2.20E-016	PITCH STDEV 200	2.20E-016	SAL SHIMMER 300
0.0194	INTENSITY STDEV	2.20E-016	RATIO VOICED FRAMES 200	2.20E-016	SVL JITTER 300
1.95E-012	PITCH MAX	4.31E-013	SLOPE INTENSITY 200	2.96E-012	SVL SHIMMER 300
3.14E-010	PITCH MEAN	2.20E-016	SLOPE PITCH 200	1.81E-006	INTENSITY MAX 500
1.97E-013	PITCH MIN	2.20E-016	SLOPE STYPCH 200	9.36E-007	INTENSITY MEAN 500
2.20E-016	PITCH STDEV	2.20E-016	NOISE TO HARMONICS_RATIO 200	7.94E-011	INTENSITY MIN 500
2.94E-013	RATIO VOICED FRAMES	2.20E-016	SAL JITTER 200	3.27E-005	INTENSITY STDEV 500
2.20E-016	SLOPE INTENSITY	2.20E-016	SAL SHIMMER 200	1.20E-014	PITCH MAX 500
2.20E-016	SLOPE PITCH	2.20E-016	SVL JITTER 200	2.60E-013	PITCH MEAN 500
2.20E-016	SLOPE STYPCH	1.05E-012	SVL SHIMMER 200	2.20E-016	PITCH MIN 500
2.20E-016	NOISE TO HARMONICS_RATIO	6.31E-007	INTENSITY MAX 300	2.20E-016	PITCH STDEV 500
2.22E-014	SAL JITTER	3.18E-006	INTENSITY MEAN 300	9.73E-013	RATIO VOICED FRAMES 500
2.20E-016	SAL SHIMMER	3.96E-006	INTENSITY MIN 300	0.2363	SLOPE INTENSITY 500
1.83E-012	SVL JITTER	6.63E-014	INTENSITY STDEV 300	2.20E-016	SLOPE PITCH 500
0.000833	SVL SHIMMER	4.31E-016	PITCH MAX 300	2.20E-016	SLOPE STYPCH 500
2.20E-016	VR IPU	2.88E-015	PITCH MEAN 300	2.20E-016	NOISE TO HARMONICS_RATIO 500
2.20E-016	VR LASTWORD	2.20E-016	PITCH MIN 300	2.20E-016	SAL JITTER 500
0.0001465	INTENSITY MAX 200	2.20E-016	PITCH STDEV 300	2.51E-011	SAL SHIMMER 500
0.0004361	INTENSITY MEAN 200	2.20E-016	RATIO VOICED FRAMES 300	2.20E-016	SVL JITTER 500
1.96E-005	INTENSITY MIN 200	9.26E-006	SLOPE INTENSITY 300	1.89E-010	SVL SHIMMER 500

Figura 3.1: W-Values

3.1.2. Tests estadísticos

El objetivo inicial es comparar dos grupos de IPU, aquellas seguidas por un hold (H) y por un switch (S), en forma apareada para los hablantes. Es decir, para cada hablante tenemos un valor para el grupo H y otro para el grupo S, y un test estadístico apareado nos dirá si la diferencia entre ambos grupos es significativa. El test apareado de Wilcoxon es una prueba no paramétrica para comparar las medias de dos muestras apareadas y determinar si existen diferencias entre ellas. Este test requiere que los datos estén apareados y que cada par sea aleatorio e independiente de los demás, pero no requiere que los datos posean una distribución normal. Con este test se obtuvo un resultado significativo para casi todos los indicadores, como veremos en el resto del presente capítulo.

En aquellos casos en que no se obtuvo un resultado significativo, ello puede deberse a lo reducido del tamaño de la muestra, ya que en nuestros datos no todos los hablantes tienen un valor definido tanto para H como para S (o sea, hay hablantes que sólo tienen valores pertenecientes a uno de los dos grupos). Para poder incluir a todos los hablantes en el análisis, se decidió correr también el test estadístico Kruskal-Wallis, una variante no paramétrica de ANOVA. Entonces, en este segundo análisis, comparamos los valores de todas las IPU del grupo H contra todas las del grupo S. Nótese que en este caso estamos violando la suposición de independencia de las muestras, dado que cada hablante aporta no una sino varias IPU.

Por último, llevamos adelante un tercer análisis para acotar la inexactitud del segundo análisis, debido a la violación de la suposición de independencia. En este caso, tomamos un único valor por hablante, descartando el resto de sus datos, asegurando así la independencia en la muestra a analizar. Este muestreo aleatorio es generado 1000 veces, aplicando para cada uno de ellos el test Kruskal-Wallis explicado en el párrafo anterior, y reportando el promedio de los p-valores arrojados en cada iteración.

3.2. Entonación Final

En primer lugar se estudió la entonación final de los IPU. Este es uno de los indicadores más frecuentemente mencionados en la literatura (Duncan, 1972; Ford and Thompson, 1996). Se define entonación final como la variación del nivel tonal del habla sobre la última parte de la frase. Puede ser ascendente, descendente o sostenida.

Gravano (2009) mostró que en aquellas IPU que se encuentran al final de un turno, la entonación presenta o bien una subida o bien una caída abrupta. Ahora veremos si este comportamiento, corroborado en un diálogo entre dos personas, se mantiene en un intercambio entre el sistema Let's Go! y un usuario. Para estimar la entonación final se tomó el *valor absoluto* de la pendiente del tono: un valor absoluto cercano a 0 corresponde a una entonación sostenida, y un valor absoluto lejano a 0, a una entonación ascendente o descendente.

La hipótesis nula planteada en todos los tests (H_0) se define de manera tal que la pendiente del tono se mantiene constante entre los distintos IPU, a diferencia de la hipótesis alternativa que determina que varía entre los distintos tipos de IPU. Este estudio se realizó a lo largo de toda la IPU, pero solo se muestra el resultado de los últimos 500 ms y 300ms de las IPU, los cuales se hallan descriptos en las tablas 3.1 y 3.2.

Test Estadístico	P-Valor
Test Wilcoxon	$1,05e^{-005}$
Test Kruskal-Wallis	$6,037e^{-08}$
Test Kruskal-Wallis (1000)	0,18

Tabla 3.1: Resultados Obtenidos - Entonación - Últimos 500ms de la IPU

Test Estadístico	P-Valor
Test Wilcoxon	0,0004
Test Kruskal-Wallis	$2,18e^{-08}$
Test Kruskal-Wallis (1000)	0,14

Tabla 3.2: Resultados Obtenidos - Entonación - Últimos 300ms de la IPU

La hipótesis nula se rechaza en todos los casos donde el p-valor encontrado es menor a 0,05. Así pues el único test estadístico para el cual el resultado obtenido no puede determinar significativamente la diferencia entre los valores de las IPU, es Kruskal Wallis después de 1000 ejecuciones sobre muestras generadas aleatoriamente. A pesar de esto el resto de los resultados bastan para determinar que el comportamiento de la entonación final difiere significativamente en aquellas IPU que son sucedidas por un cambio de hablantes en comparación con aquellas que se encuentran sucedidas por otra IPU del mismo hablante.

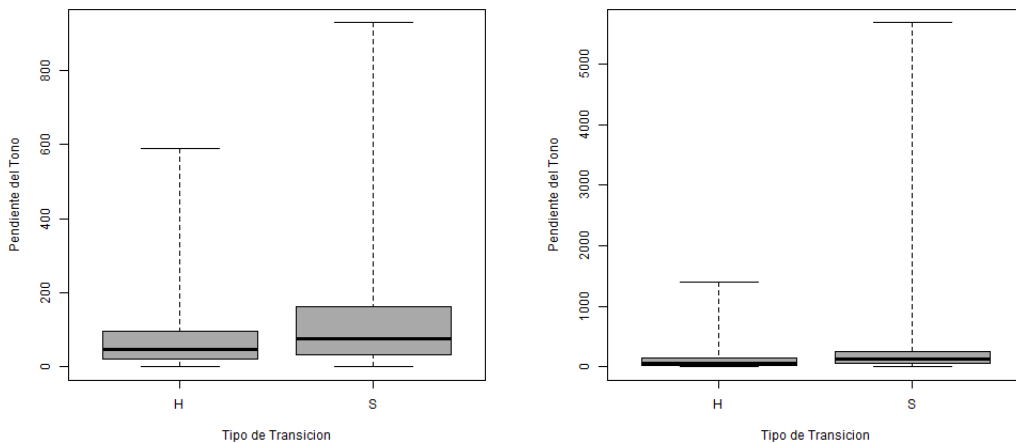


Figura 3.2: BoxPlot - Valor absoluto de la pendiente del tono sobre los últimos 500ms y 300 ms de la IPU

La figura 3.2 muestra que el valor absoluto de la pendiente del tono es mayor antes de una transición. En otras palabras, el tono de voz de una persona cae o asciende abruptamente cuando la misma está finalizando sus dichos,

previo a dejar paso a su interlocutor. Y lo que es más, este comportamiento se mantiene sin importar si su interlocutor es un sistema de interactivo de diálogo u otra persona, pudiendo reconocer así a la entonación final como un indicador de cesión del turno.

3.3. Duración del IPU

Dentro de las características de la unidad inter-pausal que fueron sometidas a estudio se encuentra la duración de la IPU. Cutler and Pearson (1985) afirma que aquellas IPU que se encuentran al final de un turno, tienen mayor duración que aquellas que se encuentran a la mitad del mismo. Esto último fue mostrado empíricamente en Gravano (2009) utilizando como cuerpo de datos a un conjunto de IPU provenientes de un diálogo entre dos personas. Veamos ahora si este comportamiento se repite cuando un usuario utiliza un sistema IVR, como en este caso Let's Go!

La hipótesis nula dice que la duración de las IPU es similar en ambos grupos, y la hipótesis alternativa que las IPU catalogadas como SWITCH poseen una duración distinta a los HOLDS. Los resultados obtenidos tras aplicar la serie de test estadísticos utilizados para este trabajo se detalla en la tabla 3.3.

Test Estadístico	P-Valor
Test Wilcoxon	0,2724
Test Kruskal-Wallis	$4,91e^{-09}$
Test Kruskal-Wallis (1000)	0,009

Tabla 3.3: Resultados Obtenidos - Duración de la IPU en segundos

La hipótesis nula es rechazada en todos los casos donde el p-valor encontrado es menor a 0,05. En este caso los dos tests de Kruskal-Wallis indican que la duración en segundos difiere significativamente en las IPU que se encuentra finalizando un turno. La figura 3.3 muestra que la duración de la IPU es claramente mayor en aquellas IPU catalogadas como SWITCH en comparación a aquellas catalogadas como HOLD.

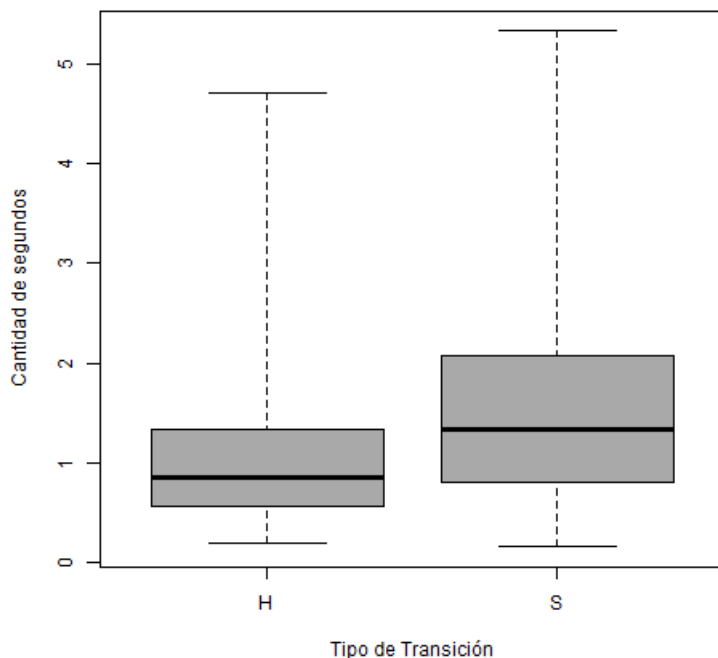


Figura 3.3: Duración de los IPUs en segundos

Por otro lado, la cantidad de sílabas por IPU puede tomarse también como otra medida de duración de la IPU, por lo cual también será analizado su comportamiento en el conjunto de datos estudiado. Sean la hipótesis nula aquella donde la cantidad de sílabas no difiere entre los grupos, y la hipótesis alternativa aquella donde existe una diferencia. Los resultados son detallados en la tabla 3.4

Test Estadístico	P-Valor
Test Wilcoxon	0,1650
Test Kruskal-Wallis	$3,28e^{-10}$
Test Kruskal-Wallis (1000)	0,004

Tabla 3.4: Resultados Obtenidos - Cantidad de sílabas por IPU

La figura 3.4 acompaña los resultados que se hallaron con el estudio estadístico de este indicador. La cantidad de sílabas difiere significativamente

entre las IPU que se encuentran al final de un turno y aquellas que se hallan en medio del mismo. Es más, la gráfica muestra que la diferencia entre la cantidad de sílabas es claramente mayor en las IPU catalogadas como SWITCH en comparación a aquellas catalogadas como HOLD.

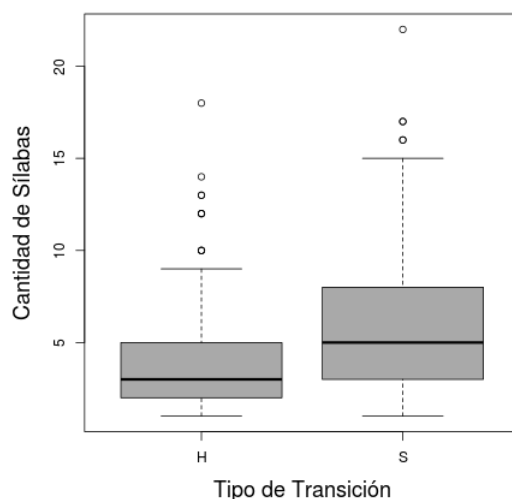


Figura 3.4: Cantidad de Sílabas por IPU

Los resultados obtenidos permiten concluir que la duración de la IPU puede ser utilizada como un indicador de cesión del turno, así sea tomando como medida la cantidad de segundos por IPU o la cantidad de sílabas por IPU.

3.4. Tasa del Habla

Duncan (1972) menciona la existencia de *cierto arrastre en la última sílaba acentuada de la frase*, como una de las características que se presentan cuando el hablante cede su turno. Por su parte Gravano (2009) mostró que, si bien ese arrastre sobre el final de la IPU suele existir en todas las IPU, el mismo se acorta hacia el final de las IPU que se encuentran en el fin del turno. En otras palabras, el hablante tiende a hablar levemente más rápido antes de ceder su turno, sugiriendo así que la tasa del habla de las últimas palabras de una IPU puede tomarse como un indicador de fin del turno.

En este trabajo, definimos la tasa del habla como la cantidad de sílabas por segundo y se calculó el valor sobre toda la IPU y sobre la última palabra de la misma. En primera instancia veamos qué ocurre con la tasa del habla sobre toda la IPU. Sea la hipótesis nula donde la tasa del habla se comporta igual en ambos grupos y la hipótesis alternativa aquella donde la tasa del habla se comporta de manera diferente en el grupo de las IPU. Los resultados obtenidos se detallan en la tabla 3.5, de allí se desprende que la diferencia entre la tasa del habla entre las IPU de los distintos conjuntos es significativa.

Test Estadístico	P-Valor
Test Wilcoxon	$2,232e^{-06}$
Test Kruskal-Wallis	0,058
Test Kruskal-Wallis (1000)	0,51

Tabla 3.5: Resultados Obtenidos - Sílabas por segundo de toda la IPU

La figura 3.5 muestra que esta diferencia se debe a que la media de la tasa de habla es levemente menor en las IPU previas a que el hablante cede su turno a su interlocutor. Antes de sacar conclusiones con este resultado veremos qué ocurre con el comportamiento de la tasa del habla sobre la última palabra de la IPU.

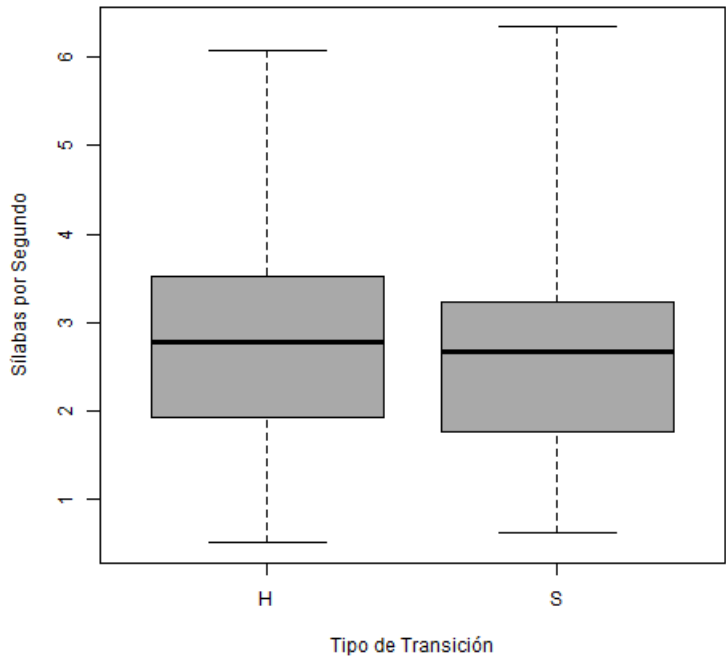


Figura 3.5: Distribución de la tasa del habla - Toda la IPU

Se define como hipótesis nula, en este caso, a la hipótesis que indica que el valor de la tasa del habla en la última palabra de la IPU se mantiene constante entre los grupos de IPU. Por otra parte la hipótesis alternativa dice que el valor de la tasa del habla de la última palabra varía entre las IPU catalogadas como SWITCH y aquellas catalogadas como HOLD. Los resultados obtenidos se muestran en la tabla 3.6, donde además se observa que hay suficiente información para determinar que la diferencia de la tasa del habla en la última palabra es significativa entre las IPU de los distintos grupos.

Test Estadístico	P-Valor
Test Wilcoxon	0,0026
Test Kruskal-Wallis	0,001
Test Kruskal-Wallis (1000)	0,10

Tabla 3.6: Resultados Obtenidos - Sílabas por segundo de la última palabra de la IPU

Es más, en el cuerpo de datos estudiado, la tasa del habla de la última palabra de una IPU SWITCH es mayor que en aquellas IPUs catalogadas como HOLD (ver figura 3.6).

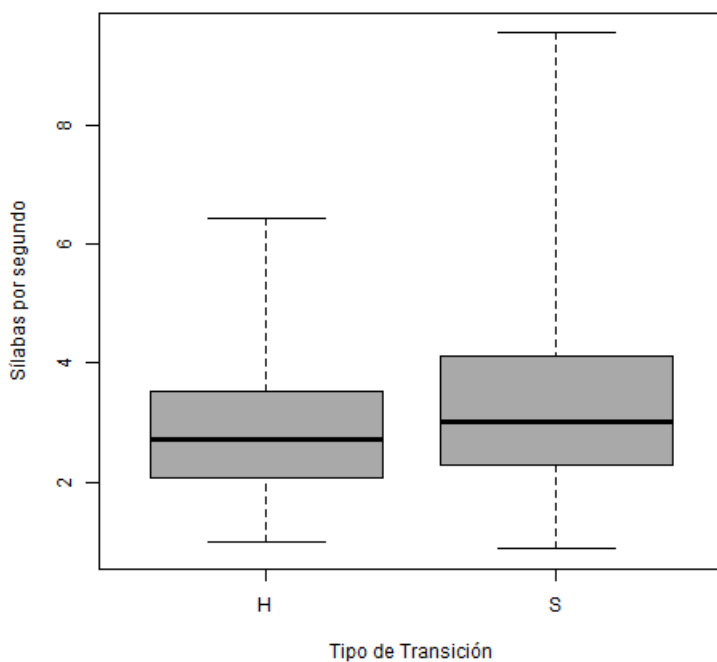


Figura 3.6: Distribución de la tasa del habla - Última palabra

Lo mostrado anteriormente coincide con lo estudiado en Gravano (2009), donde se encontró que, (a) la última palabra de las IPUs que finalizan un turno tienden a ser pronunciadas con mayor velocidad que las últimas palabras de un HOLD. Sin embargo (b) esta medida decrece cuando se analiza la

IPU en su totalidad donde aquellas catalogadas como SWITCH poseen una tasa del habla menor que las HOLD. Concluyendo así que la tasa del habla puede ser utilizada como otro indicador de fin del turno de una IPU también en un diálogo entre un sistema IVR y su usuario.

3.5. Niveles de intensidad y tono

En la teoría de prosodia se menciona que el tono y la intensidad de la voz tienden a disminuir a medida que transcurren los dichos del hablante. Esta disminución se marca aún más al llegar al final de un turno (Pierrehumbert and Hirschberg, 1990). Esto fue confirmado en Gravano (2009), donde además se exhibió que tanto la intensidad como el tono de voz pueden ser tomados como indicadores de fin de turno para los diálogos entre seres humanos.

Siendo H_0 la situación donde la media del tono de voz se mantiene constante entre los conjuntos de IPU, y la hipótesis alternativa aquella donde dicha media varía entre los grupos, se decidió aplicar el test estadístico Kruskal-Wallis. Los resultados obtenidos tras correr los tests estadísticos se detallan en la tabla 3.7. Allí se ve que se cuenta con pruebas para determinar que la diferencia en el tono de voz entre las IPU de distintas categorías, SWITCH y HOLD, es significativa.

Test Estadístico	P-Valor
Test Wilcoxon	$2,825e^{-11}$
Test Kruskal-Wallis	0,0002
Test Kruskal-Wallis (1000)	0,31

Tabla 3.7: Resultados Obtenidos - Media del Tono de voz de la IPU

La figura 3.7 permite agregar a lo anteriormente enunciado que dicha diferencia no solo existe en la muestra analizada, sino que permitió agregar el dato que el tono de voz decae en las IPU que son sucedidas por los dichos del interlocutor. Esto muestra así que al igual que en un diálogo entre personas, al momento de relacionarse con un sistema IVR, el ser humano tiende a disminuir el tono de su voz al llegar al final de su turno.

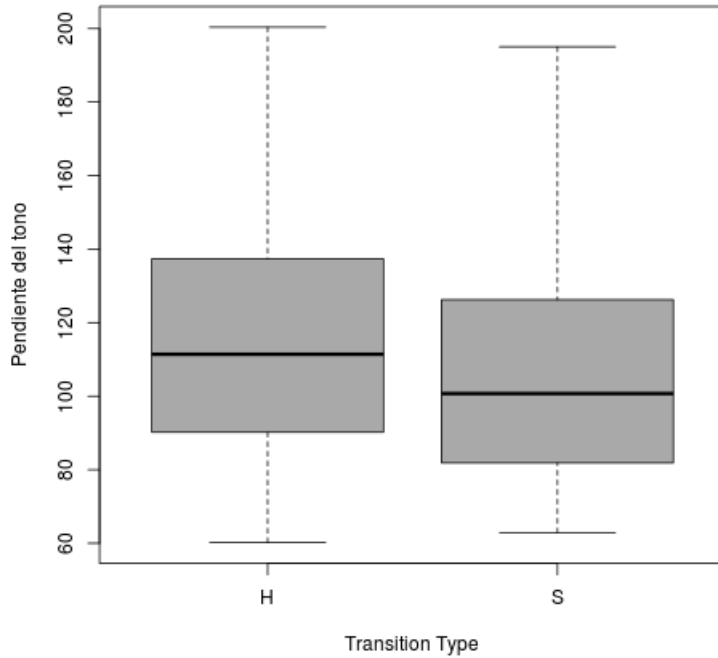


Figura 3.7: Distribución del tono de voz medio

Veamos ahora cómo se comporta la intensidad de la voz en un diálogo entre un usuario y un sistema IVR. Se define H_0 como la hipótesis que determina que la intensidad de la voz se mantiene constante para ambos grupos de IPU y a la hipótesis alternativa como aquella donde la intensidad de la voz varía de acuerdo al grupo de IPU de donde se la mida. Al aplicar los distintos tests estadísticos los resultados obtenidos se detallan en la tabla 3.8. Nuevamente los p-valores encontrados permiten determinar que la diferencia en el valor medio de la intensidad de la voz es significativa entre las IPU de los distintos grupos.

Test Estadístico	P-Valor
Test Wilcoxon	$1,987e^{-05}$
Test Kruskal-Wallis	0,009
Test Kruskal-Wallis (1000)	0,23

Tabla 3.8: Resultados Obtenidos - Intensidad media de la voz a lo largo de la IPU

Finalmente, la figura 3.8 se ve que la intensidad decae en las IPU que se encuentran al final de los dichos del hablante, poniendo así a la intensidad de la voz como otro de los indicadores de cesión del turno.

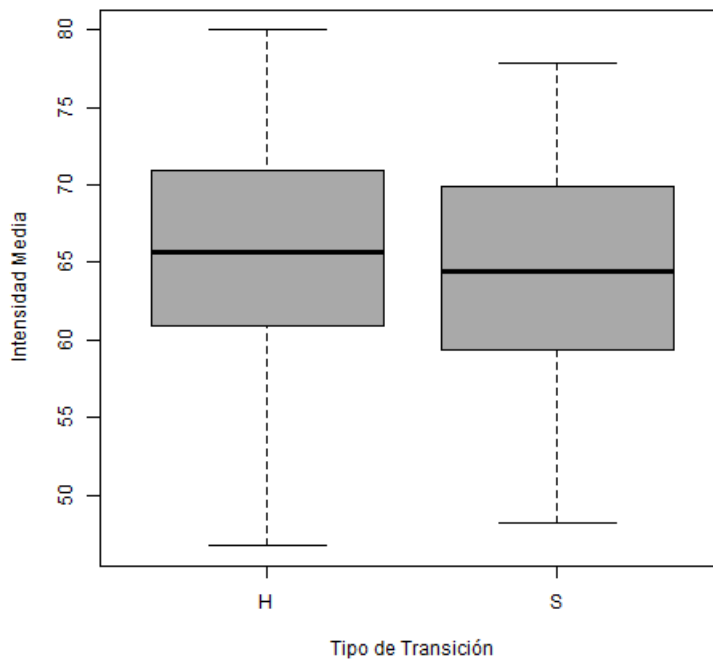


Figura 3.8: Distribución de la intensidad

3.6. Completitud gramatical

Tradicionalmente se dice que la oración es una secuencia de palabras que tiene sentido completo y autonomía sintáctica aunque puede ser imprecisa. Esta definición trata de reflejar el hecho de que la oración es el fragmento más pequeño del discurso que comunica una idea completa y posee independencia, pudiéndola sacar del contexto y seguir comunicando. En la literatura del tema se menciona que la completitud gramatical de la oración puede ser tomada como un indicador de cesión del turno (Duncan, 1972; Ford and Thompson, 1996; Sacks et al., 1974; Wennerstrom and Siegel, 2003).

Gravano (2009) propone una técnica automática para determinar la completitud gramatical de una IPU con un 80% de aciertos. Con ello, se presentó evidencia empírica del rol de la completitud como indicador de cesión del turno. En este trabajo se optó por hacer un análisis manual de este indicador. El proceso utilizado se detalla a continuación.

En una primera instancia se transcribió el habla de los usuarios de Let's Go!, de manera tal que la IPU a rotular como completa o incompleta se presentara acompañada del turno actual hasta la IPU objetivo y el turno completo anterior. A esto llamamos un *token* a etiquetar. Dado que este trabajo se concentra en los dichos de los usuarios de un sistema IVR, el turno completo anterior es lo mencionado por el sistema y la IPU a rotular es la última parte de la respuesta del hablante hasta una pausa. A este par no se le agrega ninguna información del futuro, para poder solo analizar la IPU por sí misma. La figura 3.9 muestra tres tokens de ejemplo. Se presentaron 490 tokens a los etiquetadores en un documento impreso. El orden en el que aparecen los tokens en el documento es completamente aleatorio y no tiene relación con las distintas conversaciones que se transcribieron.

El documento generado fue distribuido a 3 etiquetadores que debieron analizar cada IPU siguiendo la siguiente directiva: *“Siendo A el sistema IVR y B el usuario; determinar para cada token, si lo dicho por B a este punto puede ser interpretado como una respuesta completa, a lo que el interlocutor A dijo en el turno o segmento previo”*. Es decir que se replicó el sistema de etiquetado que se define en (Gravano, 2009). A manera de ejemplo, los tokens de la figura 3.9 fueron rotulados de la siguiente manera: (001) incompleto, (002) completo, (003) completo.

```

-----
[001] A: For example, you can say, when is the next 28X from downtown to airport? Or I' d like to go from
      McKeesport to Homestead tomorrow at 10 a.m.
      B: four a
-----
[002] A: Where do you want to go?
      B: uh children's hospital
-----
[003] A: What can I do for you?
      B: when is the next bus from oakland to downtown
-----

```

Figura 3.9: Extracto del documento a etiquetar

Una vez etiquetadas todas las IPU, se procedió a medir el nivel de acuerdo entre los etiquetadores, a través de la medida Kappa de Fleiss (Fleiss, 1971). Dicha medida permite ver el nivel de acuerdo por sobre el azar entre un número fijo de etiquetadores que asignan una categoría a un número de ítems. Al calcular la medida Kappa sobre la muestra de 490 tokens con un número de 3 etiquetadores, se obtuvo un resultado de 0.602. Este valor permite considerar como *sustancial* al nivel de acuerdo entre los etiquetadores. Una vez catalogadas las IPU de acuerdo a su completitud gramatical se puede verificar si existe una relación entre esta categorización y la finalización de la cesión de turno.

En un primer análisis, la distribución del cuerpo de datos estudiado muestra que las IPU que finalizan los dichos del hablante son en su mayoría completos sintácticamente. En comparación, aquellas IPU que se encuentran sucedidas por otros IPU del mismo hablante, que son compuestos en su mayoría por oraciones incompletas (ver figura 3.10).

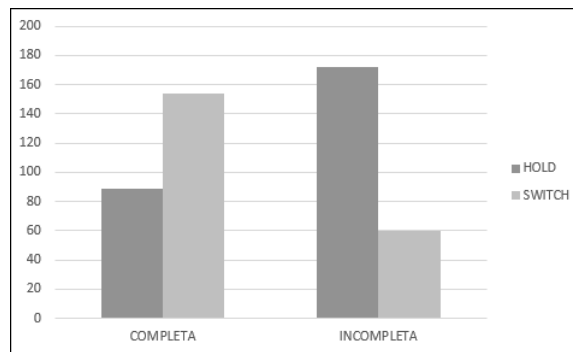


Figura 3.10: Distribución de la Completitud Gramatical

Sin embargo, estos resultados no bastan para concluir un resultado, por lo cual se optó por someter a la muestra a un estudio estadístico. En este caso, se optó por aplicar un test χ^2 . Los valores obtenidos, $\chi^2 = 107.25$, p-valor $< 2,2e^{-16}$, muestran que los resultados son significativos estadísticamente.

Con todo lo anterior, se puede concluir que, al igual que en un diálogo entre personas, la completitud gramatical es un claro indicador de cesión del turno en un sistema IVR.

3.7. Calidad de la voz

En diversos estudios se presenta a la calidad de la voz como otra de las características que tiene un comportamiento particular cuando el hablante finaliza su cesión. Ogden (2003, 2004) presenta a una voz crepitante (en inglés, *creaky voice*), a los suspiros, o a sonidos sin voz como la exhalación, como un indicador de finalización de turno. Para verificar este fenómeno, en Gravano (2009) se analizaron tres atributos mensurables automáticamente, como el jitter, shimmer y la relación ruido-armónico (NHR, del inglés *noise-to-harmonics ratio*).

El *jitter* es una medida de variabilidad de la frecuencia que no tiene en cuenta los cambios voluntarios de la frecuencia fundamental. Nos permite detectar un cambio indeseado y abrupto de la propiedad de una señal. El *shimmer* permite medir la amplitud de la perturbación de la frecuencia, cuantificado así los pequeños lapsos de inestabilidad de la señal vocal. Finalmente la *relación ruido-armónico* se utilizó para estimar el nivel de ruido en la voz humana. Las mediciones obtenidas en el trabajo mencionado permitieron concluir que estas características elevan notoriamente su valor, previo a finalizar un turno, pudiéndose tomar así a la calidad de la voz como un indicador de fin del turno en los diálogos entre personas. En los párrafos siguientes se tratará de determinar cómo se comportan las características mencionadas en un diálogo entre un sistema IVR y su usuario.

Comencemos viendo el comportamiento del shimmer a lo largo de los últimos 500ms de las IPU pertenecientes al cuerpo de datos extraído de Let's Go!. Siendo H_0 la hipótesis donde el shimmer posee un valor constante sin importar el grupo de IPU que se estudie y la hipótesis alternativa aquella donde el valor varía entre los grupos. Los resultados obtenidos se detallan en la tabla 3.9, donde los p-valores muestran que la diferencia en el valor del shimmer es significativa a lo largo de los últimos 500 ms de las IPU

pertenecientes dependiendo de la categoría a la que pertenecen.

Test Estadístico	P-Valor
Test Wilcoxon	0,005
Test Kruskal-Wallis	$3,9e^{-05}$
Test Kruskal-Wallis (1000)	0,05

Tabla 3.9: Resultados Obtenidos - Shimmer a lo largo de los últimos 500ms de la IPU

Esto también se ve en la figura 3.11 donde el valor que toma el shimmer en las IPU sucedidas por una cesión del turno es mayor a aquellas IPU catalogadas como HOLD.

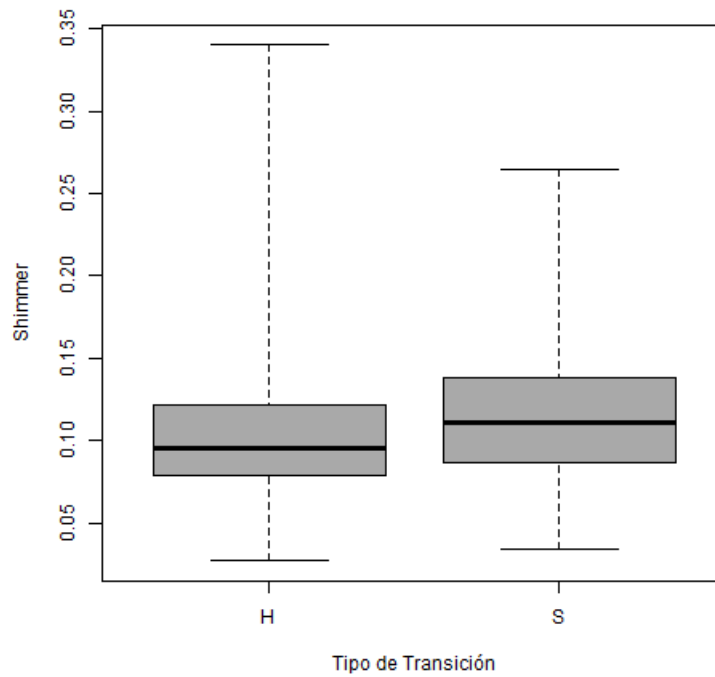


Figura 3.11: Distribución del shimmer en los IPU

En cuanto al jitter ocurre una situación similar como se observa en la

figura 3.12.

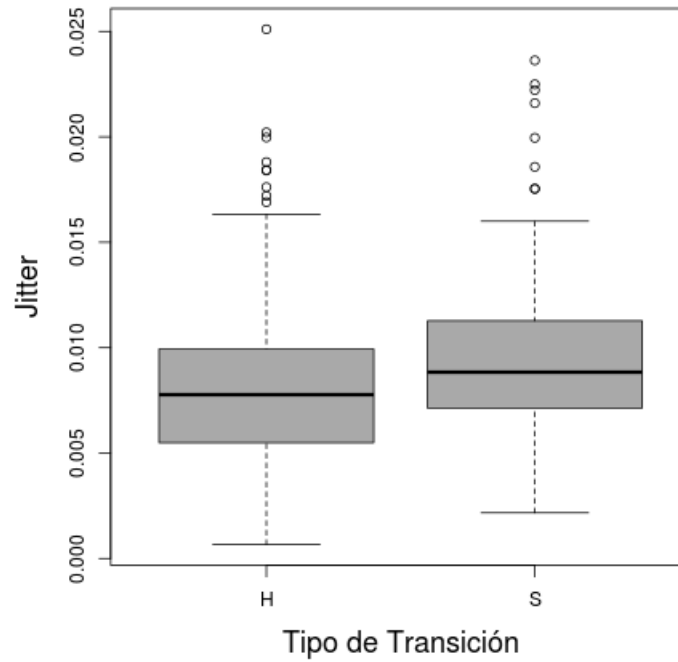


Figura 3.12: Distribución del Jitter en las IPU

Tomando hipótesis similares, se obtuvieron los resultados detallados en la tabla 3.10. De allí se desprende que la diferencia entre los valores del jitter de los grupos es significativa, de forma tal que el jitter incrementa su valor en las IPU catalogadas como SWITCH.

Test Estadístico	P-Valor
Test Wilcoxon	0,0003
Test Kruskal-Wallis	$9,8e^{-06}$
Test Kruskal-Wallis (1000)	0,10

Tabla 3.10: Resultados Obtenidos - Jitter de la IPU

Finalmente veamos qué ocurre con la relación ruido-armónico en los dis-

tintos grupos de IPU. Sea la hipótesis H_0 , donde se considera que el valor de la relación ruido-armónico se mantiene constante entre los dos grupos de IPU. La hipótesis alternativa de este estudio es aquella donde el valor difiere entre los grupos de IPU. Los resultados se detallan en la tabla 3.11, donde se muestra que la diferencia entre los valores del ruido-armónico es significativa entre las IPU de las distintas categorías.

Test Estadístico	P-Valor
Test Wilcoxon	0,0001
Test Kruskal-Wallis	0,0002
Test Kruskal-Wallis (1000)	0,18

Tabla 3.11: Resultados Obtenidos - ruido-armónico de la IPU

Además la figura 3.13 muestra que la media de los valores que toma la relación ruido-armónico en las IPU catalogadas como SWITCH es considerablemente mayor en comparación al valor de la media de aquellas IPU que se encuentran sucedidos por otras IPU del mismo hablante. Podemos así definir a la relación ruido-armónico como un indicador de fin del turno.

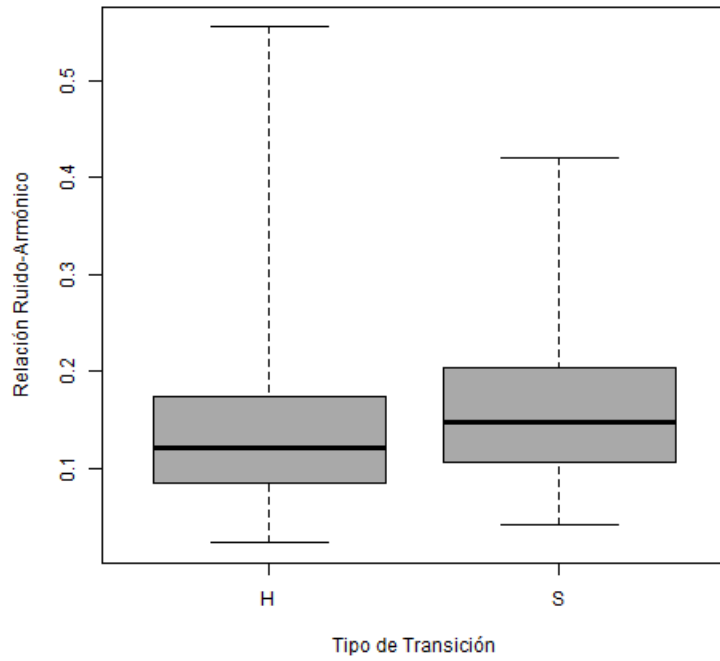


Figura 3.13: Distribución de la Relación Armónico-Ruido

Como conclusión, las diferentes medidas de la calidad de la voz estudiadas pueden ser tomadas como otro indicador de cesión del turno también cuando se trata de un diálogo entre un usuario y un sistema IVR.

Capítulo 4

Señales Complejas de Cesión del Turno

Hasta aquí se mostró la existencia de distintos indicadores acústico-prosódicos y sintácticos que suelen preceder a un fin del turno de un hablante. Lo que se intentará mostrar en este capítulo es si estos indicadores se presentan en conjunto en las IPU, generando señales complejas de fin del turno. El objetivo de estas señales es permitir predecir si la IPU estudiada es la última previo a una cesión del turno.

Para corroborar la existencia de dichas señales primeramente se debió determinar qué indicadores están presentes en una IPU. Para ello se realizó un estudio discreto de los indicadores de fin del turno similar al que se plantea en (Gravano, 2009). La tabla 4.1 muestra las variables elegidas para determinar la presencia o ausencia de cada indicador. Decimos que un indicador está presente en la IPU de un hablante cuando, para alguna de las variables que modelan al indicador, la distancia del valor observado a la media en todas las IPU SWITCH del hablante es mayor a la distancia del mismo valor a la media en todas las IPU HOLD del hablante.

El estudio de señales complejas se realizó solo con aquellos hablantes que contaban con más de 2 IPU de cada tipo (SWITCH, HOLD); en caso contrario este cálculo perdería valor estadístico y no serviría para los fines buscados.

Indicador	Variables Numéricas
Entonación	Pendiente del tono durante los últimos 300ms de la IPU Pendiente del tono durante los últimos 500ms de la IPU
Duración de la IPU	Duración en segundos de la IPU Cantidad de sílabas en la IPU
Tasa del Habla	Sílabas sobre segundos de la IPU Sílabas sobre segundos de la última palabra
Intensidad	Media de la intensidad de la IPU
Tono	Media del tono de la IPU
Calidad de la voz	Jitter sobre todo la IPU Shimmer sobre todo la IPU Relación ruido-armónico sobre todo la IPU
Complejidad	Complejidad Gramatical de la IPU

Tabla 4.1: Variables seleccionadas para determinar la presencia de los diferentes indicadores de fin del turno

4.1. Algoritmo para determinar la presencia de indicadores

Todos los indicadores fueron analizados a través de este algoritmo excepto aquel que habla sobre la complejidad gramatical de una IPU, dado que su valor es binario por definición.

Sean c un indicador y v una variable que modela c ; por ejemplo, cantidad de sílabas por segundo modela al indicador tasa del habla. Se define v_u al valor que tiene la variable v para la IPU u . El algoritmo 1 determina si un indicador está presente en una IPU u . Cabe aclarar que el cálculo de la media de las variables v se resuelve excluyendo a la IPU actual; a este método se lo conoce como *dejar uno afuera* (*leave-one-out*).

Algoritmo 1 Algoritmo para determinar la presencia o ausencia de un indicador de sesión del turno c en una IPU u

$presente \leftarrow falso$;
para toda variable v que modela al indicador c **hacer**
 $v_S \leftarrow$ media de v para las IPUs catalogadas como SWITCH (excl. u);
 $v_H \leftarrow$ media de v para las IPUs catalogadas como HOLD (excl. u);
 $v_u \leftarrow$ valor de la variable v para la IPU u ;
 si $|v_u - v_S| < |v_u - v_H|$ **entonces**
 $presente \leftarrow verdadero$;
 fin si
fin para
devolver $presente$

4.2. Análisis

Una vez determinada la cantidad de indicadores que están presentes en una IPU, se trató de ver cómo se comportan los indicadores de manera conjunta. Para ello se buscó mostrar las primeras diez combinaciones de indicadores tanto para las IPUs catalogadas como SWITCH como aquellas catalogadas como HOLD. Estas combinaciones se pueden ver en la tabla 4.2.

Como se esperaba, la combinación con mayor número de apariciones para las IPUs sucedidas por un final de turno, es aquella que cuenta con la presencia de todos los indicadores. Dentro de las siguientes nueve posiciones del ranking para las IPUs catalogadas como SWITCH, cuatro de ellas poseen seis de los siete indicadores estudiados. Para las IPUs catalogadas como HOLD el comportamiento es muy distinto, la primera combinación que aparece en el ranking solo cuenta con cuatro indicadores, y exceptuando la tercera posición de la lista el resto de las combinaciones no cuentan con más de cuatro indicadores presentes. Es decir, las IPUs que se encuentran finalizando el turno del hablante poseen un mayor número de indicadores en comparación con las IPUs que son sucedidas por otras IPUs del mismo hablante.

Para reforzar lo mostrado se calculó el porcentaje de IPUs SWITCH y HOLD que poseen 0, 1, 2, 3, ..., ó 7 indicadores. En esta oportunidad no importa el orden en el que aparecen los indicadores, sino cuántos de estos están presentes en una IPU. El resultado se muestra en la tabla 4.3, de la cual se desprende que la media de la cantidad de indicadores de las IPUs catalogadas

SWITCH		HOLD	
Indicadores	Cantidad	Indicadores	Cantidad
1234567	17	1.2..6.	14
1.34567	14	1.3456.	13
1.3.567	14	1.3....	13
123.567	12	1.3.567	12
123..67	9	1.3.56.	9
12..567	9	1234567	8
1.3.56.	9	1.345..	8
1.3..67	9	1.34...	8
1234..7	8	..3.56.	8
12.4567	7	123456.	7

TOTAL	108	TOTAL	100

Tabla 4.2: Top 10 de las combinaciones de los indicadores presentes en los IPU. 1: Entonación; 2: Duración; 3: Tasa del Habla; 4: Intensidad; 5: Tono; 6: Calidad de la voz; 7: Completitud Gramatical

como SWITCH ronda el número 6. En cambio la media de cantidad de indicadores para las IPU catalogadas como HOLD, ronda los 3 y 4 indicadores. Más a aún, se puede ver también que la muestra analizada no contiene IPU catalogadas como SWITCH que posean uno o ningún indicador estudiado.

# Indicadores	SWITCH	HOLD
0	0 (0%)	2 (100%)
1	0 (0%)	14 (100%)
2	7 (19.44%)	29 (80.56%)
3	17 (19.54%)	70 (80.46%)
4	46 (42.20%)	63 (57.80%)
5	79 (60.77%)	51 (39.23%)
6	47 (66.20%)	24 (33.80%)
7	17 (68%)	8 (32%)
Total	213 (44.94%)	261 (55.06%)

Tabla 4.3: Distribución de la cantidad de indicadores para cada tipo de transición

Con los valores de la misma tabla se realizó la figura 4.1, donde se puede ver que, a mayor número de indicadores, mayor es la probabilidad de que la IPU sea SWITCH. También se ve que los indicadores presentes en las IPU Hold tiene una pendiente negativa, cuando en cambio el comportamiento es completamente opuesto en las IPU catalogadas previamente como SWITCH. Concluimos así que el número de indicadores presente en las IPU puede ser utilizado también como una señal de que el usuario del sistema IVR está concluyendo su turno.

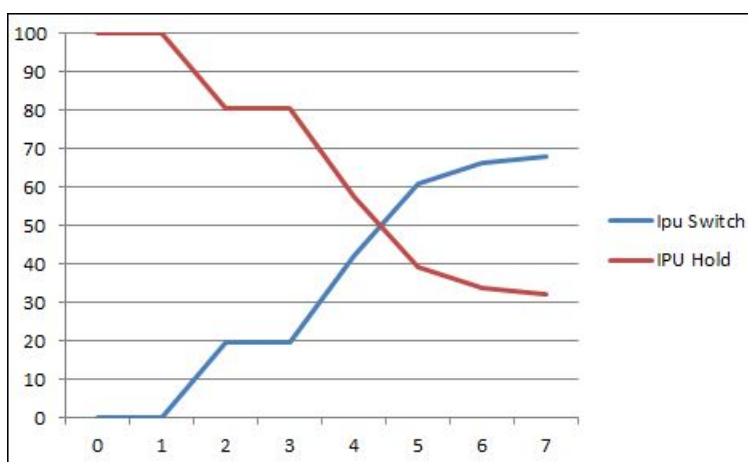


Figura 4.1: Porcentaje de SWITCH vs. HOLD según la cantidad de indicadores presentes

Capítulo 5

Conclusión y Trabajo Futuro

Los estudios realizados en las secciones previas se han basado en los resultados obtenidos en (Gravano, 2009). Allí se mostró la existencia de siete indicadores auditivos mensurables que permiten predecir un final del turno en un diálogo colaborativo entre seres humanos. Lo que se ha mostrado en este trabajo es que los mismos indicadores están presentes y se comportan de manera similar en los diálogos donde interrelacionan un sistema IVR y su usuario, que también puede catalogarse como colaborativo dado que ambos se interrelacionan para llegar a un objetivo común.

Dichos indicadores pueden ser estudiados en conjunto, creando una señal compleja. Estas también puede predecir un final del turno dado que existe una relación entre el número de indicadores que están presentes en una IPU y la clasificación de la misma, es decir si es SWITCH o HOLD. Esta relación indica que a mayor número de indicadores presentes, mayor la probabilidad que dicha IPU sea la última del hablante previo a ceder su turno. De esta manera los indicadores de fin del turno pueden ser utilizados dentro de un sistema IVR permitiéndole al mismo predecir el fin del turno del usuario. La implementación de este análisis en tiempo real de los dichos del usuario forma parte de los trabajos futuros que se desprenden de esta Tesis. De poder predecir con éxito el fin del turno del hablante estos indicadores podrían llevar a una mejora en la fluidez en el diálogo entre el usuario y el sistema IVR, permitiendo cambiar la categorización de “confusos” e “intimidantes” que actualmente poseen los sistemas IVR.

Bibliografía

- Antoine, R., Alan, B., and Eskenazi, M. (2003). Lets go!:improving spoken dialog systems.
- Black, A. and Eskenazi, M. (2009). The spoken dialog challenge.
- Boersma, P. and Weenink, D. (2001). Praat:doing phonetics by computer.
- Cutler, A. and Pearson, M. (1985). On the analysis of prosodic turn-taking cues. In Johns-Lewis, editor, *Intonation in Discourse*, pages 139–155. Croom Helm, London.
- Duncan, S. (1972). Some signals and rules for taking speaking turns in conversations. *Journal of Personality and Social Psychology*, pages 283–292.
- Fleiss, J. (1971). Measuring nominal scale agreement among many raters. *Psychological Bulletin*, 76(5):378–382.
- Ford, C. and Thompson, S. (1996). Interactional units in conversation: Syntactic, intonational and pragmatic resources for the management of turns. In Schegloff, E., Ochs, E., Schegloff, E., and Thompson, S., editors, *In Interaction and Grammar*, pages 134–184. Cambridge University Press.
- Gravano, A. (2009). *Turn Taking and Affirmative Cue Words in Task-Oriented Dialogue*. PhD thesis, Columbia University.
- Gravano, A. and Hirschberg, J. (2011). Turn-taking cues in task-oriented dialogue. *Computer Speech and Language Vol. 25(3)*, pages 601–634.
- Hjalmarsson, A. (2011). The additive effect of turn-taking cues in human and synthetic voice. *Speech Communication*, 53(1):23–35.

- Norskog, L. (2012). Sox - sound exchange.
- Ogden, R. (2003). Proceedings of the xvth international congress of phonetic sciences. In *Creaky voice as a resource for the management of turn-taking in Finnish talk-in-interaction*.
- Ogden, R. (2004). Sound patterns in interaction. In *Non-modal voice quality and turn-taking in Finnish*. Elizabeth Couper-Kuhlen and Cecilia Ford.
- Pierrehumbert, J. and Hirschberg, J. (1990). The meaning of intonational contours in the interpretation of discourse. In *Intentions in Communication*, pages 271–311. P.H. Cohen, J. Morgan, M. E. Pollack.
- Raux, A., Bohus, D., Black, B. L. A., and Eskenazi, M. (2006). Doing research on a deployed spoken dialogue system: One year of let’s go! experience. In *Proceedings of Interspeech*.
- Raux, A. and Eskenazi, M. (2008). Optimizing end pointing thresholds using dialogue features in a spoken dialogue system. In *Proceedings of SIGdial*.
- Sacks, H., Schegloff, E. A., and Jefferson, G. (1974). A simplest systematics for the organization of turn-taking for conversation. In *Language*, volume 50, pages 696–735. Linguistic Society of America.
- Ward, N., Rivera, A., Ward, K., and Novick, D. (2005). Root causes of lost time and user stress in a simple dialog system. In *Proceedings of Interspeech*.
- Wennerstrom, A. and Siegel, A. F. (2003). Keeping the floor in multiparty conversations: Intonation, syntax, and pause. In *Discourse Processes*, pages 77–107. Routledge.