5.1 Introducción

En el presente capítulo se describen los parámetros propuestos y los experimentos realizados para la obtención de medidas de confianza en el sistema CU Communicator, desarrollado en The Center for Spoken Language Research (CSLR) de la Universidad de Colorado (apéndice A), y en los sistemas de reconocimiento de habla continua realizados en la presente tesis: reconocimiento de nombres deletreados (capítulo 3) y reconocimiento de fechas y horas (capítulo 4).

Atendiendo a la clasificación de las medidas de confianza según su nivel de aplicación, realizada en el capítulo 2, cabe comentar que el estudio realizado sobre el sistema CU Communicator se centrará en los niveles de palabra, concepto semántico y frase. En este estudio, utilizaremos parámetros provenientes del decodificador, del modelo de lenguaje y del módulo de comprensión. Por otro lado, en el reconocedor de fechas y horas trabajaremos principalmente al nivel de palabra, utilizando parámetros tanto del decodificador como del modelo de lenguaje. Las mismas fuentes de parámetros se usarán para obtener medidas de confianza al nivel de frase (secuencia de letras correspondientes a un nombre) en el caso del reconocedor de nombres deletreados. En este último sistema, se analizará también el problema de la detección de nombres fuera del vocabulario de reconocimiento (OOV: Out Of Vocabulary). En todos los casos, para la combinación de los diferentes parámetros considerados y la obtención de un único valor de confianza, utilizaremos una Red Neuronal sencilla, un Perceptrón Multi-Capa (Widrow et al, 1988; Ruck et al, 1990).

En este capítulo no trabajaremos con medidas de confianza al nivel de interacción sistema–usuario, ni utilizaremos parámetros del gestor de diálogo como propone Carpenter (Carpenter et al, 2001). En cualquier caso, en el capítulo 6 se comentarán medidas para la comparación y evaluación de diálogos. Estas medidas serán utilizadas por un lado, con el fin de gestionar la técnica de modelado de usuario introducida, y por otro lado, para comparar varias alternativas de diálogo y elegir la mejor de ellas.

5.2 Medidas de Confianza en el sistema CU Communicator

El sistema CU Communicator utiliza el reconocedor Sphinx (Ravishankar, 1996) para decodificar la señal de habla, y una versión modificada del parser Phoenix (Ward, 1994) para la extracción de la información semántica de la frase reconocida. Más detalles sobre este sistema se pueden consultar en el apéndice A.

5.2.1 Base de datos

La base de datos utilizada para los experimentos realizados sobre este sistema, se obtuvo durante la recogida de datos realizada en el CSLR desde noviembre de 1999 hasta mayo de 2000 (Pellom et at, 2000). Durante este período se recogieron más de 900

llamadas telefónicas obteniendo alrededor de 11.500 frases con más de 30.000 palabras en total.

A la hora de realizar los experimentos, hemos dividido aleatoriamente el conjunto de frases en tres subconjuntos: 66% de las frases para el entrenamiento de la red neuronal utilizada en la combinación de los parámetros (ver apartado 5.2.3), 17% para su validación y el 17% para evaluación. Esta división se ha repetido 6 veces realizando un proceso Round-Robin, de forma que cada vez, se van utilizando unos datos diferentes para entrenar, validar o evaluar la Red Neuronal, consiguiendo que se usen todos los datos para evaluar las medidas de confianza al menos una vez. Los resultados presentados en esta sección son la media de los valores obtenidos en todos los experimentos.

Para los resultados de reconocimiento que presentaremos en el apartado 5.2.6, se ha utilizado un conjunto de datos independiente, obtenido de la evaluación realizada por NIST(National Institute of Standards) en Junio de 2000 (Pellom et al, 2000)(ver apéndice A, apartado A.1.4)

5.2.1.1 Etiquetación automática de los ejemplos

Para poder realizar la experimentación sobre medidas de confianza es necesario clasificar cada ejemplo como correcto (el sistema debe aceptarlo) o incorrecto (el sistema lo debe rechazar). Al nivel de palabra, esta etiquetación se realiza mediante un alineamiento dinámico entre la hipótesis del reconocedor y la frase de referencia, obteniéndose las palabras correctas, insertadas, borradas y sustituidas. En un sistema real, a la salida de reconocedor se dispone únicamente de las palabras correctas, insertadas y sustituidas, por esta razón, trabajaremos siempre con este tipo de palabras, no asignando ningún valor de confianza a las palabras borradas.

Al igual que en el caso anterior, necesitamos etiquetar cada concepto semántico de la base de datos como correcto o incorrecto. Esta etiquetación se ha realizado de forma automática de la siguiente forma (ver figura 5-1). Al analizar la frase de referencia obtenemos una secuencia de conceptos determinada. Esta secuencia se utiliza como base con la que comparar la secuencia de conceptos obtenida de la hipótesis, y así poder calcular los casos de conceptos correctos, sustituidos, borrados e insertados. El proceso es análogo al seguido para el caso de las secuencias de palabras.

Este mecanismo tiene la limitación de que se toma como base el análisis que hace el módulo de comprensión de la frase de referencia. Por esta razón sólo se detectan errores de los conceptos debidos a problemas o fallos en el reconocimiento de palabras, y no a los que puedan surgir por errores en el propio analizador semántico. En cualquier caso, para un sistema real, la evaluación de medidas de confianza debe caracterizar los errores producidos por deficiencias en el reconocimiento y no los problemas del propio analizador. Estos errores deben haber sido resueltos en una etapa de ajuste o depuración anterior.

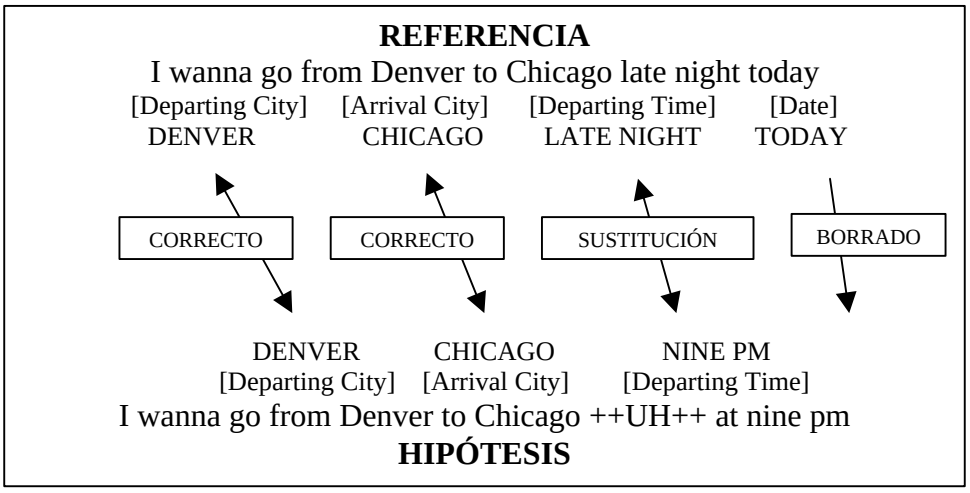

Figura 5-1: Etiquetación automática de conceptos como correctos o incorrectos.

Cuando nos planteamos la aplicación de medidas de confianza al nivel de frase, lo primero que podemos pensar es en la detección de frases que no pertenecen al dominio de la tarea (San-Segundo et al, 2000a). En este caso el mecanismo utilizado para la etiquetación de las frases ha sido de forma manual. Pero no sólo este tipo de frases debe ser detectado y rechazado en un sistema automático: puede ocurrir que frases pertenecientes al dominio de aplicación no puedan ser correctamente interpretadas debido a la gran cantidad de errores de reconocimiento que contienen. Para poder analizar la detección de este tipo de frases, junto con las no pertenecientes al dominio, debemos etiquetarlas de forma lo más automática posible. El mecanismo propuesto por Pao (Pao et al, 1998) consiste en realizar el análisis semántico tanto de la frase de referencia (transcripción) como de las N mejores hipótesis obtenidas del decodificador. Si alguna de las hipótesis genera un conjunto de conceptos (englobados en una plantilla) igual al conjunto generado en la frase de referencia se considera como frase que se debe aceptar, si no es así, se etiqueta como frase a rechazar. Por otro lado Hazen (Hazen et al, 2000b, Hazen et al, 2002) propone otro criterio basado en la calidad de reconocimiento obtenida. En este caso se etiquetan como aceptables aquellas frases cuya referencia coincide con alguna de las cuatro mejores hipótesis, o cuando para la mejor hipótesis tenemos, al menos, dos de cada tres palabras correctas.

A la hora de decidir un criterio para etiquetar las frases, debemos pensar en qué información utiliza el gestor de diálogo para interactuar con el usuario. Esta información es la obtenida a la salida del analizador semántico. Por esta razón nos parece más apropiada la medida propuesta por Pao (Pao et al, 1998) que aquellas basadas en la calidad de reconocimiento (Hazen et al, 2000b). En la presente tesis se propone una modificación del primer criterio (Pao et al, 1998). Esta modificación consiste en una relajación del mismo, de forma que, en lugar de forzar la igualdad en una plantilla completa, se considerará como frase aceptable toda aquella que tenga al menos un concepto semántico igual que en la frase de referencia. Este criterio es más relajado puesto que permite la aceptación de frases con algún error, lo que dificulta en mayor medida las estrategias de rechazo. Si en una misma frase existen conceptos

correctos e incorrectos, estos deberán ser discriminados con las medidas de confianza al nivel de concepto.

5.2.2 Parámetros de medida de confianza

En este apartado, describiremos los parámetros utilizados clasificándolos según el nivel de aplicación: palabra, concepto semántico y frase.

5.2.2.1 Nivel de palabra

En este nivel se pretende la etiquetación de cada palabra con un valor de confianza que nos ofrezca una idea de la certeza con la que se ha reconocido dicha palabra. Los parámetros que hemos utilizado al nivel de palabra son un subconjunto de los parámetros que mejores resultados han dado en la literatura. Estos parámetros se obtienen principalmente del proceso de decodificación y del modelo de lenguaje.

Los parámetros considerados del proceso de decodificación son los siguientes (Chase, 1997a; Chase, 1997b; Kamppari y Hazen, 2000; Macías-Guarasa et al, 2000b):

- **Verosimilitud normalizada**: es el logaritmo de la verosimilitud acumulada a lo largo de la palabra (durante el proceso de reconocimiento), dividido por el número de tramas. En los casos en los que la palabra se ajuste muy bien a los modelos acústicos, esta verosimilitud será mayor, revelando una mayor confianza en el reconocimiento de esa palabra. En las zonas de voz con gran ambigüedad o problemas de ruido, esta verosimilitud será menor.

- **Homogeneidad de la palabra en la lista de las 100 mejores hipótesis:** porcentaje de veces que una misma palabra aparece en posición análoga (mismo segmento de voz) en las 100 mejores hipótesis de reconocimiento. A medida que una palabra aparece con mayor frecuencia en las diferentes hipótesis nos da una idea de la mayor confianza en su reconocimiento. Las palabras que primero cambian de una hipótesis a otra, son las que reflejan una mayor incertidumbre en su reconocimiento.

- **Densidad del grafo de palabras:** número de enlaces o transiciones, desde cualquier palabra hasta la palabra considerada, calculadas sobre el grafo de palabras obtenido durante la primera etapa de reconocimiento. Cuando una palabra ha sido reconocida correctamente y con bastante seguridad, esta palabra actúa como un punto de anclaje en el grafo de palabras, de forma que gran cantidad de caminos posibles transcurren a través de ella. Este parámetro está relacionado con el anterior puesto que una palabra con gran confianza, será punto de paso de gran cantidad de caminos posibles, y por tanto, surgirán muchas hipótesis de reconocimiento que contengan dicha palabra.

- **Perplejidad de fonemas:** número medio de modelos de alófono activos (sobreviven a la poda introducida por el Beam Search) a lo largo de las tramas en las que permanece activa la palabra bajo estudio. El número de modelos activos

nos da una medida de la confusión entre los diferentes modelos acústicos a la hora de decodificar las tramas de la palabra analizada. Un valor elevado nos transmite la idea de que son muchos los modelos que están compitiendo sin que haya un claro vencedor.

Los parámetros considerados provenientes del modelo de lenguaje son los siguientes (Uhrik y Ward, 1997; San-Segundo et al, 2000a):

- **Comportamiento Back-Off del modelo de lenguaje:** comportamiento del modelo de lenguaje utilizado para calcular la probabilidad de la palabra en la secuencia ($P(W_j)$ como función de las palabras anteriores: W_{j-1} y W_{j-2})(Katz, 1987). Este comportamiento nos puede dar información sobre la certeza de la palabra: una mayor confianza se puede asignar a palabras que forman secuencias trigram (de tres palabras) contenidas en el modelo de lenguaje, y una menor confianza a comportamientos de orden inferior (bigrams o unigrams). En la tabla 5-1, se pueden ver los valores de confianza (dentro del intervalo 0-1) asignados a cada tipo de comportamiento.

Valor de Confianza	Comportamiento
1,0	$P(W_j)$ como sucesión trigram: $P(W_j,W_{j-1},W_{j-2})$
0,8	$P(W_j)$ como sucesión bigram-bigram: $P(W_j,W_{j-1})$ y $P(W_{j-1},W_{j-2})$
0,6	$P(W_j)$ como sucesión bigram: $P(W_j,W_{j-1})$
0,4	$P(W_j)$ como sucesión unigram-bigram: $P(W_j)$ y $P(W_{j-1},W_{j-2})$
0,3	$P(W_j)$ como sucesión unigram-unigram: $P(W_j)$ y $P(W_{j-1})$
0,2	$P(W_j)$ como unigram: $P(W_j)$
0,1	Palabra desconocida. Nunca se da en la salida del reconocedor.

Tabla 5-1: Asignación del valor de confianza según el comportamiento utilizado en el cálculo de la probabilidad de la palabra en la secuencia.

- **Probabilidad de la palabra en la secuencia $P(W_j)$, obtenida del modelo de lenguaje:** este parámetro nos ofrece información complementaria puesto que palabras con comportamientos iguales pueden tener probabilidades diferentes, con lo que se debe asignar una mayor confianza a las palabras con mayor probabilidad de aparición en el modelo de lenguaje.

A la hora de analizar una palabra concreta, se considerarán como parámetros de confianza tanto el comportamiento y la probabilidad de la palabra considerada, como el de las dos palabras anteriores y las dos posteriores. La razón de utilizar un contexto de dos es porque el modelo de lenguaje utilizado es 3-gram con lo que errores de palabras contextuales pueden hacer que los parámetros de la palabra analizada sean malos sin que esta sea una palabra incorrecta.

5.2.2.2 Nivel de concepto semántico

En el caso en el que las palabras pertenecientes a los conceptos semánticos se obtuviesen correctamente, el funcionamiento del sistema, desde el punto de vista del usuario, sería el mismo aunque algunas otras palabras de relleno (como artículos, pronombres,...) puedan haber sido reconocidas con error. Veamos el ejemplo de la figura 5-2. En este ejemplo tenemos una sustitución de la palabra "that" por "that's" y un borrado de la palabra "is" pero sin embargo el significado semántico es el mismo.

Frase Pronunciada:	THAT IS CORRECT
Frase Reconocida:	THAT'S CORRECT

Figura 5-2: Ejemplo de reconocimiento erróneo pero con el mismo significado.

Dado que los conceptos semánticos son la fuente de información que hace evolucionar el diálogo, las medidas de confianza a este nivel son muy útiles para hacer una gestión del diálogo más eficiente. Veamos el ejemplo presentado en la figura 5-3. Como se puede observar, para los conceptos Ciudad Origen (Departing City) y Ciudad Destino (Arrival City), con muy alta confianza, el sistema hace una confirmación implícita. Por otro lado, el concepto Hora de Salida (Departing Time) tiene una confianza baja con lo que el sistema decide realizar una confirmación explícita de la información.

S: What are you travel plans?
U: *I wanna go from Denver to Chicago late night today.*
(Reconocido: I wanna go from Denver to Chicago at nine pm)
Conceptos y medidas de confianza obtenidas:
[Departing City]: Denver (0.95)
[Arrival City]: Chicago (0.92)
[Departing Time]: nine pm (0.43)

S: You want to go from Denver to Chicago. I understood you wanna leave at nine pm. Is that correct ?
U: *No, I would prefer to leave later, around eleven pm.*
...

Figura 5-3: Ejemplo de gestión del diálogo utilizando medidas de confianza. (S) frase del sistema, (U) frase del usuario.

En este nivel, trabajaremos con parámetros del decodificador, del modelo de lenguaje y del sistema de compresión o analizador semántico. Asociado a cada parámetro se define un código para facilitar su identificación en las tablas de resultados. Los dos primeros parámetros pretenden la incorporación de la confianza de cada palabra en el cálculo de la confianza al nivel de concepto semántico. De esta forma se incorpora el conocimiento del decodificador y del modelo de lenguaje en la etiquetación de confianza a este nivel:

- **Confianza Media de las palabras pertenecientes a la Regla utilizada en la obtención del Concepto (CMRC).** Este parámetro se obtiene realizando la media de la confianza obtenida en el nivel anterior, a lo largo de las palabras utilizadas en la aplicación de la regla que generó el concepto analizado.

- **Confianza Media de las palabras pertenecientes al Valor del Concepto (CMVC).** De forma análoga al anterior, este parámetro se calcula realizando la media de la confianza obtenida en el nivel anterior, a lo largo de las palabras que forman el valor del concepto analizado.

  La razón de hacer esta diferenciación es porque el conjunto de palabras pertenecientes a la regla aplicada puede ser mayor que las que forman el valor del concepto propiamente dicho. Si consideramos la frase de la figura 5-3, podemos ver que "from Denver" es la regla aplicada para obtener la Ciudad Origen (preposición "from" y nombre de ciudad), mientras que el valor de la ciudad origen es únicamente la palabra "Denver".

En esta tesis proponemos un conjunto de parámetros obtenidos exclusivamente del módulo de compresión. Son los siguientes:

- **Número de Palabras contenidas en la Regla aplicada para obtener el concepto (NPR).** Definir una regla de análisis en el analizador semántico Phoenix, es equivalente a imponer una gramática de contexto libre sobre la secuencia de palabras. A medida que se aplican reglas que involucran mayor número de palabras, se pone de manifiesto un mayor ajuste entre la frase y la gramática, lo que redunda en una mayor confianza del concepto obtenido.

- **Número de Palabras contenidas en el Valor del concepto obtenido (NPV).** Siguiendo un razonamiento similar al anterior, el valor de un concepto suele ser una secuencia de palabras características con alta probabilidad en el modelo de lenguaje, lo que refleja también una mayor confianza cuando estas secuencias son más largas.

- **Homogeneidad del concepto en las 100 mejores hipótesis (HC).** De forma análoga al caso de las palabras, este parámetro es el porcentaje de veces que un concepto aparece en las 100 mejores hipótesis de reconocimiento, por ejemplo Ciudad Destino del viaje (Arrival City). Se permite que el concepto tenga diferentes valores, es decir, que en algunos casos la ciudad de destino sea "Boston" y en otros "Austin". Cuanto mayor sea este porcentaje, mayor será la confianza del concepto analizado.

- **Homogeneidad del concepto y su valor en las 100 mejores hipótesis (HCV):** En este caso exigimos que aparezca el concepto con el mismo valor.

  Estos dos últimos parámetros son muy útiles cuando tenemos palabras que tienen el mismo comportamiento semántico (ej: dos nombres de ciudad), y además, tienen un gran parecido acústico como por ejemplo las ciudades Boston y Austin. En estos casos se producen patrones característicos en los que el concepto semántico aparece muchas veces, pero su valor fluctúa bastante. Este tipo de comportamiento nos

refleja una gran confianza de que el concepto especificado es ese, por ejemplo la Ciudad Destino (Arrival City), pero su valor es bastante incierto. La detección de estos patrones permite ajustar y particularizar mejor las confirmaciones o correcciones a realizar sobre esa parte de la información.

Los dos últimos parámetros que describimos a continuación se obtienen a partir de la definición de un modelo de lenguaje al nivel de concepto. Una vez definidas las secuencias de conceptos, podemos entrenar un modelo de lenguaje que caracterice las secuencias que con mayor frecuencia aparecen en los diálogos del dominio de aplicación. En nuestro caso, hemos entrenado un modelo de lenguaje conceptual 3-gram, utilizando las secuencias de conceptos obtenidas como resultado de analizar las frases de referencia (transcripciones manuales) del conjunto de entrenamiento. De forma análoga al caso del nivel de palabra, podemos definir los siguientes dos parámetros:

- **Comportamiento Back-Off del modelo de lenguaje (CML):** comportamiento del modelo de lenguaje utilizado para calcular la probabilidad del concepto en la secuencia ($P(C_j)$ como función de los conceptos anteriores: C_{j-1} y C_{j-2}).

- **Probabilidad del concepto en la secuencia $P(C_j)$, obtenida del modelo de lenguaje (PML).**

  De igual forma que en el apartado anterior, se utilizarán los parámetros del concepto analizado, junto con los parámetros de los dos conceptos anteriores y posteriores (contexto de 5 conceptos).

Un detalle importante a la hora de entrenar el modelo de lenguaje conceptual es que no se deben considerar los valores de los conceptos. Es decir, el concepto Ciudad Destino, por ejemplo, debe ser considerado como la misma unidad independientemente del valor que tenga asociado.

5.2.2.3 Nivel de frase

El objetivo en este nivel es detectar frases que serán mal interpretadas por el sistema. La razón de este error puede ser debida a fallos en el reconocimiento por problemas en la calidad de la señal, o porque el usuario ha solicitado un servicio que no está considerado entre las posibilidades del sistema (ej: si el usuario desea información del tiempo en un sistema de reserva de billetes de avión) (Uhrik y Ward, 1997; San-Segundo et al, 2000a). En estos trabajos se utilizan únicamente parámetros del modelo de lenguaje para la detección de frases fuera del dominio de aplicación. En otros trabajos (Pao et al, 1998) los autores hacen un estudio bastante completo de parámetros obtenidos del proceso de decodificación, del modelo de lenguaje y del módulo de comprensión para calcular la confianza al nivel de frase.

En este nivel utilizaremos medidas obtenidas del proceso de decodificación, del modelo de lenguaje y del módulo de comprensión. Estas medidas básicamente son las mismas que las propuestas al nivel de concepto pero extendidas a toda la frase.

- **Confianza Media al nivel de Palabra (CMP):** es la media de los valores de confianza obtenidos para cada una de las palabras que componen la frase. De esta forma consideramos la información del proceso de decodificación y del modelo de lenguaje en la etiquetación al nivel de frase.

- **Confianza Media al nivel de Concepto (CMC):** es la media de los valores de confianza obtenidos para cada uno de los conceptos que componen la frase.

- **Porcentaje de Palabras Analizadas Semánticamente (PPAS):** número de palabras que pertenecen a algún concepto o a alguna regla utilizada para obtener algún concepto, dividido por el número de palabras de la frase.

- **Porcentaje de Palabras pertenecientes a la Tarea (PPT):** número de palabras que pertenecen a algún concepto o a alguna regla definida en la tarea (aunque no haya sido utilizada en la frase actual), dividido por el número de palabras de la frase.

  Si bien este último parámetro de forma aislada, como veremos más adelante, puede no tener mucha repercusión sobre las medidas de confianza, su combinación con el porcentaje de palabras analizadas semánticamente (PPAS) nos puede ser de gran ayuda. En situaciones en las que aparezca gran cantidad de palabras pertenecientes a la tarea (porque acústicamente encajen bien con la frase pronunciada, como preposiciones “to” o “from”) pero el porcentaje de palabras involucradas en las reglas o conceptos extraídos (PPAS) sea bajo, se puede deducir que la confianza de la frase en general será muy baja.

- **Porcentaje de Conceptos (PC):** número de conceptos extraídos divido por el número de palabras que componen la frase. Cuando el número de conceptos es muy bajo comparado con el número de palabras contenidas en una frase, este hecho nos revela una baja confianza de la frase.

- **Porcentaje de frases en las 100 Mejores hipótesis con algún Concepto (PMC):** porcentaje de hipótesis de reconocimiento en las que se obtiene algún concepto al ser analizadas semánticamente. Esta medida es la complementaria a la propuesta por Pao (Pao et al, 1998) debido a que en nuestro caso hemos considerado un criterio de clasificación más relajado.

5.2.3 Combinación de parámetros

Para todos los niveles hemos utilizado un Perceptrón MultiCapa para combinar los diferentes parámetros y obtener una única medida de confianza. En la literatura podemos encontrar otras alternativas de combinación como los Árboles de Decisión (Breiman et al, 1984; Pao et al, 1998) o el Análisis Lineal Discriminante (Linear Discriminant Analysis)(Hazen et al, 2000a; Hazen et al, 2000b; Hazen y Bazzi, 2001).

Al nivel de palabra, realizamos una cuantificación de los parámetros de entrada de la red. Para cada una de las entradas consideramos 10 bits, excepto para el comportamiento del modelo de lenguaje en el que únicamente son necesarios 6 bits para

codificar las 6 posibles situaciones. La codificación se ha realizado utilizando intervalos de tamaño variable, de forma que se permita una mayor resolución en rangos con mayor cantidad de datos. Esta distribución se ha realizado teniendo en cuenta los datos del conjunto de entrenamiento. La capa oculta está formada por 30 neuronas y la capa de salida por una única neurona. En el entrenamiento se utiliza el algoritmo de retropropagación para estimar los pesos de la red. En esta fase se fija un valor objetivo de 1 para el caso de una palabra correcta, y 0 para los casos de palabras incorrectas (sustituciones e inserciones).

Al realizar la codificación de los parámetros, las entradas, y por tanto el número de pesos a entrenar, se incrementa considerablemente pero puede ofrecer mejores resultados si se utiliza un número de bits elevado y se dispone de suficientes datos para entrenar correctamente los pesos de la red. En el caso de nivel de palabra, disponemos de 3.480 pesos a entrenar y alrededor de 20.000 ejemplos para entrenarlos. Al nivel de concepto y frase, el número de entradas sería mayor puesto que tenemos un mayor número de parámetros, y la cantidad de datos para entrenar es menor: 10.000 ejemplos para el nivel de concepto y alrededor 6.000 para el caso de frase. Por esta razón, para estos dos últimos niveles no realizaremos la codificación de los parámetros y los aplicaremos directamente a las entradas de la red. Con esta solución, es necesario realizar un preproceso de estos parámetros con el fin de limitar su rango dinámico al intervalo [0,1]. En este caso, el preproceso consiste en un reescalado del parámetro, teniendo en cuenta los valores máximo y mínimo obtenidos del conjunto de entrenamiento. En este punto es importante comentar la siguiente consideración: en los datos de entrenamiento se pueden encontrar ejemplos, que debido a un mal funcionamiento hagan que el valor mínimo o máximo se aleje mucho del resto de valores del parámetro. Este hecho puede provocar que el reescalado haga perder cierta resolución en aquellos rangos con mayor cantidad de ejemplos. Para evitar este problema, los máximos y mínimos utilizados para reescalar los parámetros se obtienen haciendo la media con el 5% de los mayores (cálculo del máximo) o menores (cálculo del mínimo) valores.

5.2.4 Evaluación de las medidas de confianza

Antes de proceder a la presentación de los resultados, definiremos algunos términos importantes y explicaremos el proceso de evaluación utilizado. Veamos las siguientes definiciones:

- *Rechazo Correcto (RC):* porcentaje de casos de error que han sido rechazados correctamente.

- *Rechazo Incorrecto (RI):* porcentaje de casos correctos que han sido rechazados incorrectamente.

- *Error de Clasificación (EC):* porcentaje de ejemplos que se etiquetan incorrectamente, ya sean ejemplos a rechazar o ejemplos que se deben aceptar. El complementario al error de clasificación lo denominaremos Tasa de Clasificación (TC): TC = 100% – EC.

- *Error de Referencia (ER):* el error de referencia queda definido por la distribución inicial de los ejemplos para el estudio de las medidas de confianza (en el caso del estudio al nivel de palabra, esta distribución viene definida por la calidad del sistema de reconocimiento). Al nivel de palabra dispondremos de palabras correctas, insertadas y sustituidas, siendo el Error de Referencia el porcentaje correspondiente a las palabras insertadas y sustituidas. De igual forma el Error de Referencia al nivel de concepto es el porcentaje de conceptos insertados y sustituidos en la secuencia. En el caso de las frases, el Error de Referencia lo define el conjunto de casos etiquetados como frases a rechazar según el criterio comentado en el apartado 5.2.1.1.

Para evaluar las medidas de confianza procedemos de la siguiente forma: a cada uno de los ejemplos del conjunto de test le aplicamos la Red Neuronal, obteniendo un valor de confianza comprendido en el intervalo [0, 1]. Para cada tipo de ejemplo (ejemplo a rechazar o a aceptar) se representa la distribución de casos según la medida de confianza obtenida de la Red Neuronal. Para realizar esta representación se divide el intervalo [0, 1] en 100 segmentos de anchura 0,01. En la figura 5-4 podemos ver un ejemplo de esta representación.

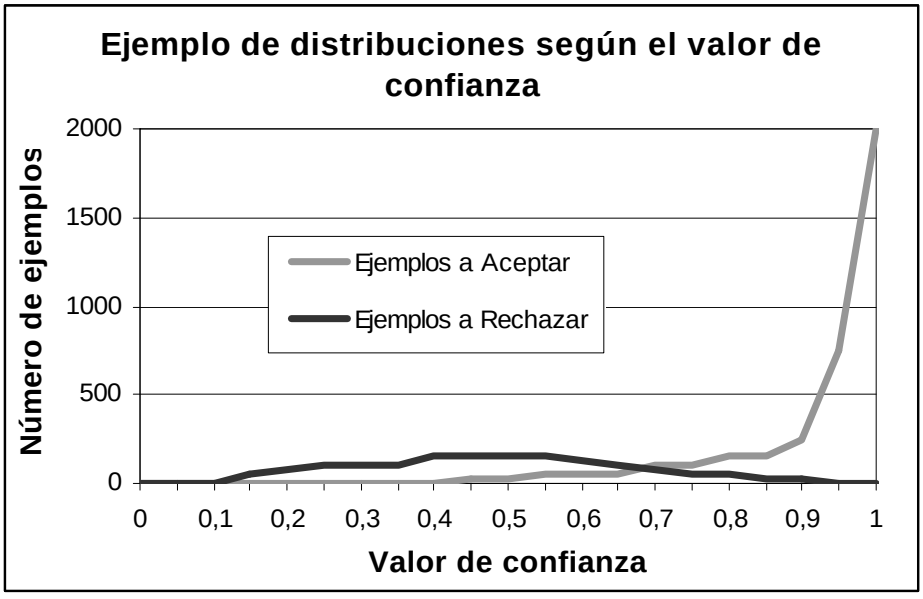

Figura 5-4: Ejemplo de representación del número de casos según el valor de confianza.

Sobre el eje x podemos definir un umbral de forma que los casos que obtengan un valor de confianza por debajo de este umbral se etiqueten como ejemplos a rechazar y los que obtengan un valor de confianza mayor se considerarán como ejemplos a aceptar. Para cualquier umbral considerado, podemos calcular el Rechazo Correcto (RC), el Rechazo Incorrecto (RI), el Error de Clasificación (EC) y las contribuciones a este error debido a los dos tipos de ejemplos; ejemplos a rechazar y ejemplos a aceptar. Las fórmulas son las siguientes:

$$RC\ (\%)=100\times\frac{N_{ERRORES\ DEBAJO\ UMBRAL}}{N_{ERRORES}} \qquad RI\ (\%)=100\times\frac{N_{ACIERTOS\ DEBAJO\ UMBRAL}}{N_{ACIERTOS}}$$

$$ECa\ (Contribución\ Aciertos) = 100 \times \frac{N_{ACIERTOS\ \ DEBAJO\ UMBRAL}}{N_{TOTAL}}$$

$$ECe\ (Contribución\ Errores) = 100 \times \frac{N_{ERRORES\ \ ENCIMA\ UMBRAL}}{N_{TOTAL}}$$

$$EC\ (\%) = ECa(Contribución\ Aciertos) + ECe(Contribución\ Errores)$$

donde:

- $N_{ERRORES\ DEBAJO\ UMBRAL}$: nº de casos a rechazar con confianza menor que el umbral.
- $N_{ACIERTOS\ DEBAJO\ UMBRAL}$: nº de casos a aceptar con confianza menor que el umbral.
- $N_{ERRORES\ ENCIMA\ UMBRAL}$: nº de casos a rechazar con confianza mayor que el umbral.
- $N_{ERRORES}$: nº de casos a rechazar.
- $N_{ACIERTOS}$: nº de casos a aceptar.
- N_{TOTAL}: nº total de ejemplos.

Como se puede ver en la figura 5-4, generalmente las dos distribuciones se solapan, lo que impide la definición de un umbral que permita separar ambas distribuciones sin error.

A la hora de evaluar las medidas de confianza calcularemos la evolución del RC según el RI, al ir modificando el umbral de confianza considerado, y el Error Mínimo de Clasificación obtenido a lo largo de esta variación del umbral. Especial énfasis haremos sobre la tasa de Rechazo Correcto para tasas de Rechazo Incorrecto del 2,5% y del 5,0%. Estos valores corresponden con el margen sobre el que oscilará el punto de trabajo en el que queremos que funcione nuestro sistema. Generalmente no es recomendable que el sistema rechace más del 5% de casos correctos (que deberían haber sido aceptados), porque podría ocasionar al usuario cierta sensación de frustración en su interacción. Para estos límites de Rechazo Incorrecto deseamos detectar y rechazar la mayor cantidad posible de errores. Por último como ya hemos comentado, el Error de Referencia vendrá determinado por la distribución inicial de ejemplos.

A la hora de calcular los intervalos de confianza de los resultados (al 95%) utilizaremos la fórmula presentada en el apartado 3.3.1.1, donde la variable **p** es el valor de la medida cuya confianza deseamos evaluar. La aplicación de esta fórmula considera la definición del valor de **n** (número de ejemplos con los que se evalúa). Para el caso de las medidas de Error de Clasificación debemos considerar el total de ejemplos utilizados (tanto ejemplos a aceptar como ejemplos a rechazar) mientras que para las medidas de Rechazo Correcto o Rechazo Incorrecto debemos tener en cuenta sólo el número de ejemplos a rechazar ($N_{ERRORES}$) o el número de ejemplos a aceptar ($N_{ACIERTOS}$) respectivamente. Para los valores de Rechazo Correcto, los márgenes de confianza se muestran en las tablas mientras que para el Error de Clasificación consideraremos que son valores del orden de ±0,5% para el caso de nivel de palabra, ±0,6% en el nivel de concepto, y ±0,7% al nivel de frase (valores calculados según el número de ejemplos **n** y el Error de Referencia **p** en cada nivel).

5.2.5 Detección de errores

En este apartado describiremos los experimentos y resultados obtenidos en la detección de errores para los tres niveles de trabajo considerados.

5.2.5.1 Nivel de palabra

Como hemos comentado anteriormente el objetivo en este nivel es detectar palabras que fueron reconocidas incorrectamente debido a problemas en la calidad de la señal o debido a que la palabra pronunciada no se encontraba incluida en el diccionario de reconocimiento. En la tabla 5-2 se presentan los porcentajes de Rechazo Correcto de errores para tasas de Rechazo Incorrecto del 2,5% y del 5,0% y el Error de Clasificación (con sus dos contribuciones ECa y ECe) para el caso de RI del 5,0%. En esta tabla también se presenta el Error Mínimo de Clasificación y el Error de Referencia. Como se puede observar los parámetros del modelo de lenguaje (ML) son mejores indicadores de la confianza al nivel de palabra que los obtenidos del proceso de decodificación (PD). Utilizando únicamente los parámetros del modelo de lenguaje, el 42,0% de errores de reconocimiento se detectan para un RI del 5%. Estos resultados son similares a los obtenidos anteriormente (San-Segundo et al, 2000a). Utilizando únicamente parámetros del proceso de decodificación sólo podemos detectar el 28,5% para el mismo RI. Los mejores resultados se obtuvieron combinando todos los parámetros. En este caso podemos rechazar más de la mitad de los errores con un Rechazo Incorrecto del 5%. Estos parámetros permiten reducir 6,2 puntos el Error de Clasificación lo que supone una reducción relativa del 32,6%.

Nivel de Palabra (Error de Referencia: 19.0%)						
	Rechazo Correcto (%)		**5,0% RI**			**Mínimo Error de Clasificación**
	2,5% RI	**5,0% RI**	**ECe**	**ECa**	**EC**	
PD	16,9 (±1,0%)	28,5 (±1,2%)	13,7%	4,1%	17,6%	17,5%
ML	28,3 (±1,2%)	42,0 (±1,3%)	11,0%	4,1%	15,1%	15,0%
PD + ML	39,0 (±1,3%)	53,2 (±1,3%)	8,9%	4,1%	13,0%	12,8%

Tabla 5-2: Rechazo Correcto de errores para Rechazos Incorrectos de 2,5% y 5%, considerando parámetros del proceso de decodificación (PD), del modelo de lenguaje (ML) y ambos juntos (PD+ML). También se muestran los Errores de Clasificación para RI del 5,0%, Mínimo Error de Clasificación y el Error de Referencia.

En la figura 5-5, representamos el Rechazo Correcto (RC) vs. Rechazo Incorrecto (RI) para los casos en los que se utilizan los parámetros del proceso de decodificación (PD), del modelo de lenguaje (ML) o todos combinados (PD + ML). Como podemos ver, los parámetros del modelo de lenguaje se comportan mejor que los parámetros del PD para casi todos los valores de RI, sobre todo en valores bajos que es donde se encuentra el punto de trabajo deseado: para el mismo valor de RI conseguimos valores de RC mayores. Otro aspecto a comentar es que los parámetros ML y PD proporcionan

información complementaria de forma que al combinarlos la gráfica resultante mejora considerablemente las dos anteriores.

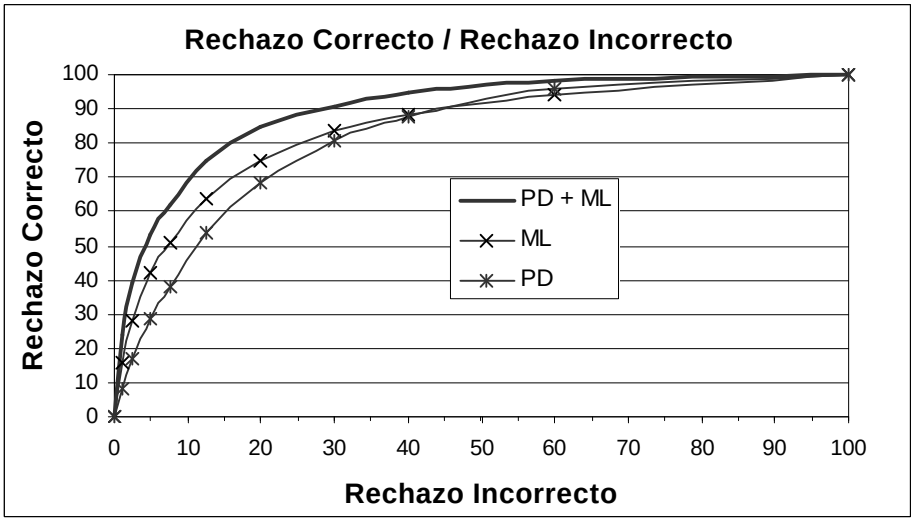

Figura 5-5: Rechazo Correcto (RC) según el Rechazo Incorrecto (RI) para los parámetros del proceso de decodificación (PD) del modelo de lenguaje (ML) o todos combinados (PD + ML).

Estos resultados son comparables a los obtenidos por Wessel y Moreau (Wessel et al, 1999; Moreau et al, 1999). En nuestro caso, obtenemos resultados ligeramente peores debido posiblemente a que en nuestra base de datos tenemos únicamente un 1,4% de palabras fuera del vocabulario de reconocimiento. Es decir, el 98,6% de las palabras están en el vocabulario lo que hace la tarea de rechazo más complicada.

5.2.5.2 Nivel de concepto

Como ya hemos comentado, el objetivo de este nivel es analizar cada concepto de forma independiente y asignarle un valor de confianza entre 0 y 1. Considerando la etiquetación automática de los conceptos (apartado 5.2.1.1), podemos calcular la Tasa de Error Conceptual de forma análoga a la Tasas de Error de palabra siguiendo las fórmulas comentadas en el capítulo 4, apartado 4.1.3. La Tasa de Error Conceptual del sistema es del 27,9% y el porcentaje de conceptos erróneos es del 16,5% (considerando sustituciones e inserciones). En la tabla 5-3 se muestran los porcentajes de Rechazo Correcto de cada parámetro independiente, para tasas de Rechazo Incorrecto del 2,5% y del 5,0% y el Error de Clasificación (con sus dos contribuciones ECa y ECe) para el caso de RI del 5,0%. En esta tabla también se presenta el Error Mínimo de Clasificación y el Error de Referencia. Como se puede deducir de los resultados presentados, las confianzas medias de las palabras en la regla o en el valor del concepto, CMRV y CMVC, son los mejores parámetros considerados en este nivel. Por ejemplo, el 47,1% de los conceptos erróneos fueron detectados para un RI del 5%, valor muy cercano al 50,1% conseguido cuando se combinan todos los parámetros propuestos. De los parámetros que utilizan exclusivamente información del analizador semántico, los que mejores resultados ofrecieron fueron los derivados del modelo de lenguaje conceptual: CML y PML.

Nivel de Concepto (Error de Referencia: 16,5%)						
	Rechazo Correcto (%)		5,0% RI			Mínimo Error de Clasificación
	2,5% RI	5,0% RI	ECe	ECa	EC	
CMRC	29,5 (±1,7%)	44,3 (±1,8%)	9,2%	4,2%	13,4%	13,3%
CMVC	30,1 (±1,7%)	45,5 (±1,8%)	9,0%	4,2%	13,3%	13,1%
CMRC + CMVC	**31,0 (±1,7%)**	**47,1 (±1,8%)**	**8,7%**	**4,2%**	**12,9%**	**12,8%**
NPR	5,3 (±0,8%)	12,3 (±1,2%)	14,5%	4,2%	18,7%	16,5%
NPV	7,2 (±0,8%)	14,8 (±1,3%)	14,0%	4,2%	18,2%	16,5%
HC	7,6 (±0,8%)	12,8 (±1,2%)	14,4%	4,2%	18,6%	16,5%
HCV	5,9 (±0,8%)	12,3 (±1,2%)	14,5%	4,2%	18,7%	16,5%
CML	25,4 (±1,6%)	37,3 (±1,8%)	10,3%	4,2%	14,5%	14,2%
PML	24,7 (±1,6%)	36,1 (±1,8%)	10,5%	4,2%	14,7%	14,0%
Resto	**29,3 (±1,7%)**	**40,1 (±1,8%)**	**9,9%**	**4,2%**	**14,1%**	**13,5%**
TODOS	**35,9 (±1,8%)**	**50,1 (±1,9%)**	**8,2%**	**4,2%**	**12,4%**	**12,0%**

Tabla 5-3: Rechazo Correcto de conceptos erróneos para Rechazos Incorrectos de 2,5% y 5%. También se muestran los Errores de Clasificación para RI del 5.0%, Mínimo Error de Clasificación y el Error de Referencia.

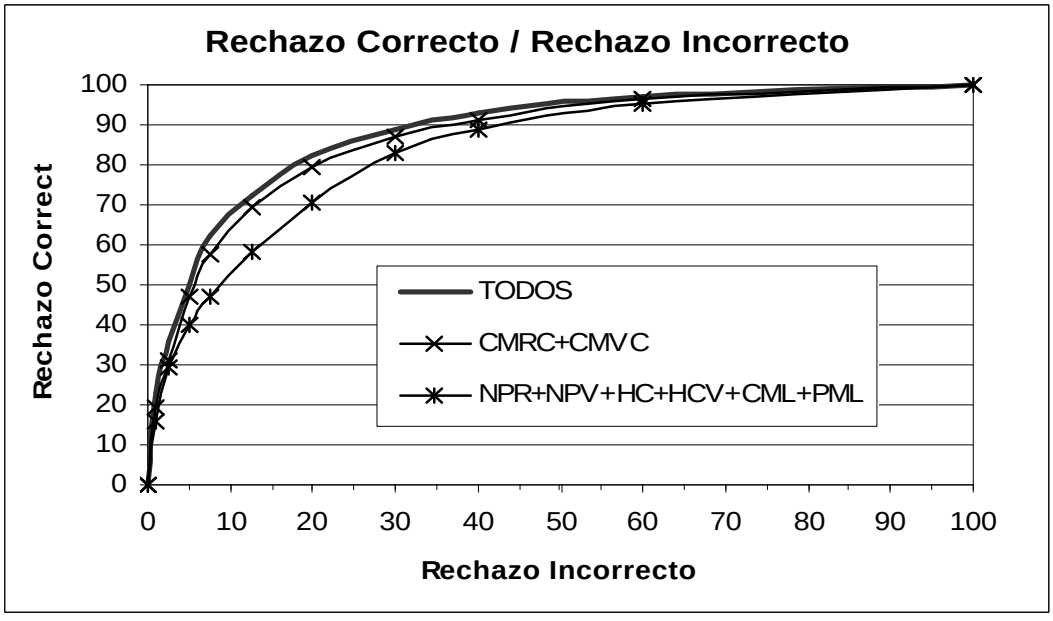

Figura 5-6: RC vs. RI para las confianzas medias al nivel de palabra (CMRC + CMRV), el resto de parámetros (NPR+NPV+HC+HCV+CML+PML) y combinándolos todos juntos.

Como muestra la figura 5-6, el mismo comportamiento comentado anteriormente se mantiene para valores de RI mayores.

5.2.5.3 Nivel de frase

A este nivel vamos a considerar dos problemas por separado. En primer lugar haremos un análisis del poder de discriminación de los parámetros para la detección, únicamente, de frases que no pertenecen al dominio de aplicación, y posteriormente, plantearemos la detección conjunta de frases fuera del dominio de aplicación y frases que no disponen de ningún concepto semántico correcto.

1.- Detección de frases NO pertenecientes al dominio de aplicación

En este primer subapartado presentamos los resultados obtenidos de la detección de frases que no pertenecen al dominio de aplicación. Para poder hacer los experimentos, hubo que etiquetar manualmente el conjunto de frases como pertenecientes o no, al dominio "Communicator". En la tabla 5-4 presentamos los porcentajes de Rechazo Correcto de frases fuera de dominio, para Rechazos Incorrectos del 2,5% y del 5,0% y el Error de Clasificación (con sus dos contribuciones ECa y ECe) para el caso de RI del 2,5%. Como se puede ver, para valores de Rechazo Incorrecto bajos, los parámetros extraídos exclusivamente del analizador semántico, funcionan mejor que lo derivados del proceso de decodificación o del modelo de lenguaje. Por ejemplo, considerando los parámetros del analizador semántico, conseguimos rechazar el 53,1% de las frases fuera del dominio para un RI del 5%. Mientras que utilizando información del proceso de decodificación y del modelo de lenguaje se detecta el 49,8%. Estos resultados mejoran estudios anteriores sobre la misma tarea (San-Segundo et al, 2000a).

Nivel de Frase: detección de frases fuera del dominio (Error de Referencia: 4,8%)						
	Rechazo Correcto (%)		**2,5% RI**			**Mínimo Error de Clasificación**
	2,5% RI	**5,0% RI**	**ECe**	**ECa**	**EC**	
CMP	**41,1 (±4,1%)**	**49,8 (±4,2%)**	**2,8%**	**2,4%**	**5,1%**	**4,2%**
CMC	18,3 (±3,2%)	37,2 (±4,0%)	3,9%	2,4%	6,3%	4,8%
PPAS	21,3 (±3,4%)	35,7 (±4,0%)	3,8%	2,4%	6,2%	4,8%
PPT	20,4 (±3,4%)	34,6 (±4,0%)	3,8%	2,4%	6,2%	4,8%
PC	34,8 (±4,0%)	47,3 (±4,2%)	3,1%	2,4%	5,5%	4,2%
PMC	36,0 (±4,0%)	47,6 (±4,2%)	3,1%	2,4%	5,5%	4,7%
CMC+PPAS+PPT+PC+PMC	**44,3 (±4,1%)**	**53,1 (±4,2%)**	**2,7%**	2,4%	**5,1%**	**4,1%**
TODOS	**46,1 (±4,2%)**	**53,4 (±4,2%)**	**2,6%**	2,4%	**5,0%**	**4,0%**

Tabla 5-4: Rechazo Correcto de frases fuera del dominio para Rechazos Incorrectos de 2,5% y 5% considerando cada parámetro por separado. También se muestran los Errores de Clasificación para RI del 2,5%, Mínimo Error de Clasificación y el Error de Referencia.

De nuevo, los mejores resultados se obtienen combinando todos los parámetros. En este caso podemos detectar el 53,4% de las frases fuera del dominio con un RI del 5%. Considerando que sólo tenemos un 4,8% de frases fuera del dominio, con estas medidas, reducimos 0,8 puntos el error de clasificación (16,7% relativo). Los mejores parámetros del analizador semántico han sido el Porcentaje de frases en las 100 Mejores hipótesis con algún Concepto (PMC) y el Porcentaje de Conceptos (PC). En este último caso conseguimos reducir 0,6 puntos el error de clasificación con este único parámetro.

FRASE FUERA DEL DOMINIO DE APLICACIÓN

REFERENCIA: *IS THERE ANY DISCOUNT FOR STUDENT CONFERENCE TRAVEL*

HIPÓTESIS: *<s> EASTERN ANY DISCOUNT THIRFTY_CAR TO TRAVEL </s>*

<s> EASTERN ANY DISCOUNT [car_rental_company] TO TRAVEL </s>

Figura 5-7: Ejemplo de frase fuera del dominio de aplicación cuyo reconocimiento genera una hipótesis con algún concepto válido en la tarea.

En la figura 5-8, representamos la gráfica de Rechazo Correcto vs. Rechazo Incorrecto. Al aumentar el RI, el parámetro Confianza Media por Palabra (CMP) funciona mejor que el resto de parámetros (CMC, PPAS, PPT, PC, PMC). En estos casos, el reconocimiento erróneo de una frase puede dar lugar a conceptos válidos en el dominio de aplicación aunque la frase sea incorrecta. En esta situación, la información del proceso de decodificación y del modelo de lenguaje (considerada en el valor de confianza obtenido para cada palabra) resulta muy útil para detectar frases fuera del dominio de aplicación. En la figura 5-7 podemos ver un ejemplo.

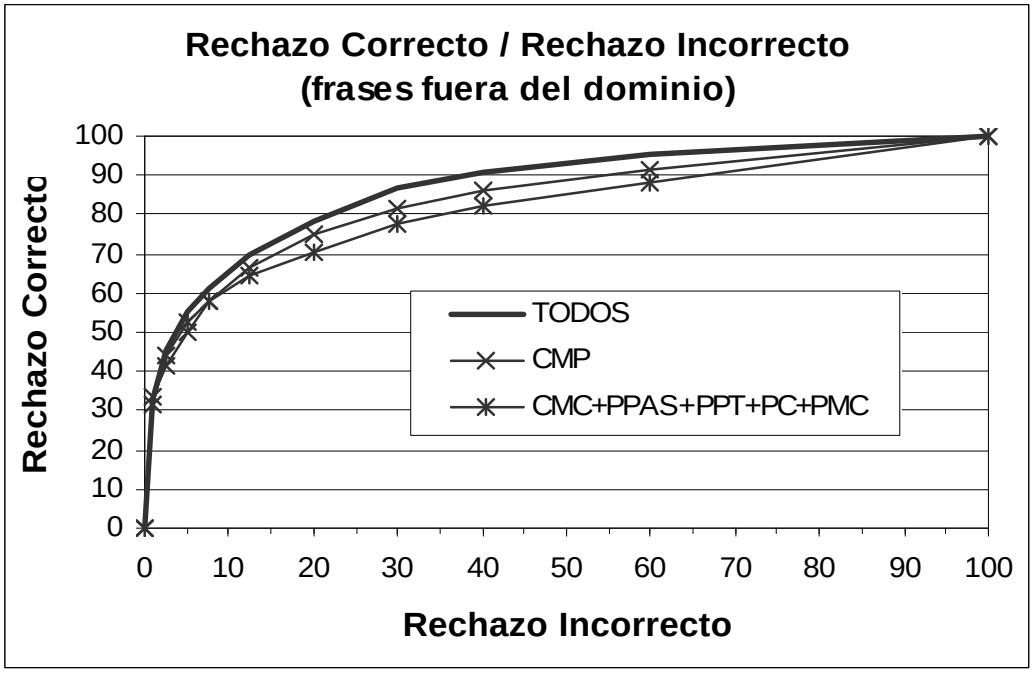

Figura 5-8: Rechazo Correcto vs. Rechazo Incorrecto (RI) para la detección de frases fuera del dominio de aplicación.

2.- Detección de frases fuera del dominio de aplicación ó frases sin ningún concepto correcto.

A continuación comentamos los resultados obtenidos en la detección conjunta tanto de frases no pertenecientes al dominio como frases que sí pertenecen, pero que por problemas de reconocimiento, no contienen ningún concepto correcto en comparación con la frase de referencia. De este último tipo de frases tenemos casos en los que no se puede extraer ningún concepto, o casos en los que todos los conceptos extraídos son erróneos. Es en esta última situación donde el rechazo de la frase es muy importante para evitar que el gestor de diálogo acepte como válidos alguno de los conceptos y guíe el diálogo por caminos que diverjan de los objetivos del usuario. Los resultados presentados en la tabla 5-5 y en la figura 5-9 muestran cómo, ahora sí, los parámetros provenientes del analizador semántico son mucho mejores que el parámetro CMP (confianza media de las palabras de la frase). Para un Rechazo Incorrecto del 5% el CMP consigue detectar más del 53% de las frases erróneas pero los parámetros semánticos superan el 68%. Como en situaciones anteriores, los mejores resultados se obtienen cuando se combinan todos los parámetros. En este caso, se detecta más del 76% de las frases para un RI del 5%. Este aumento tan importante al combinar todos los parámetros pone de manifiesto la complementariedad de los mismos. En cuando a los parámetros considerados, cabe comentar que la Confianza Media de los Conceptos de la frase (CMC) es, junto con el Porcentaje de Palabras Analizadas Semánticamente (PPAS), los dos mejores parámetros.

Nivel de Frase: detección de frases fuera del dominio y frases sin ningún concepto correcto (Error de Referencia: 22.1%)						
	Rechazo Correcto (%)		**5,0% RI**			**Mínimo Error de Clasificación**
	2,5% RI	**5,0% RI**	**Ece**	**ECa**	**EC**	
CMP	**42,7 (±1,9%)**	**53,2 (±1,9%)**	**10,3%**	**3,9%**	**14,2%**	**13,9%**
CMC	55,9 (±1,9%)	59,7 (±1,9%)	8,9%	3,9%	12,8%	11,1%
PPAS	55,1 (±1,9%)	59,3 (±1,9%)	9,0%	3,9%	12,9%	11,0%
PPT	51,6 (±1,9%)	55,3 (±1,9%)	9,9%	3,9%	13,8%	12,2%
PC	19,5 (±1,5%)	22,5 (±1,6%)	17,1%	3,9%	21,0%	19,1%
PMC	33,8 (±1,8%)	43,3 (±1,9%)	12,5%	3,9%	16,4%	16,4%
CMC+PPAS+PPT +PC+PMC	**62,3 (±1,9%)**	**68,7 (±1,8%)**	**6,9%**	**3,9%**	**10,8%**	**10,3%**
TODOS	**66,6 (±1,8%)**	**76,1 (±1,7%)**	**5,3%**	**3,9%**	**9,2%**	**9,0%**

Tabla 5-5: Rechazo Correcto de frases fuera del dominio y frases sin conceptos correctos para Rechazos Incorrectos de 2,5% y 5% considerando cada parámetro por separado. También se muestran los Errores de Clasificación para RI del 5,0%, Mínimo Error de Clasificación y el Error de Referencia.

Como podríamos prever, el hecho de considerar la validez de los conceptos como criterio para rechazar una frase en un sistema automático, hace que la medida de confianza de cada uno de estos conceptos, y en particular la media a lo largo de la frase, presente un gran poder de discriminación.

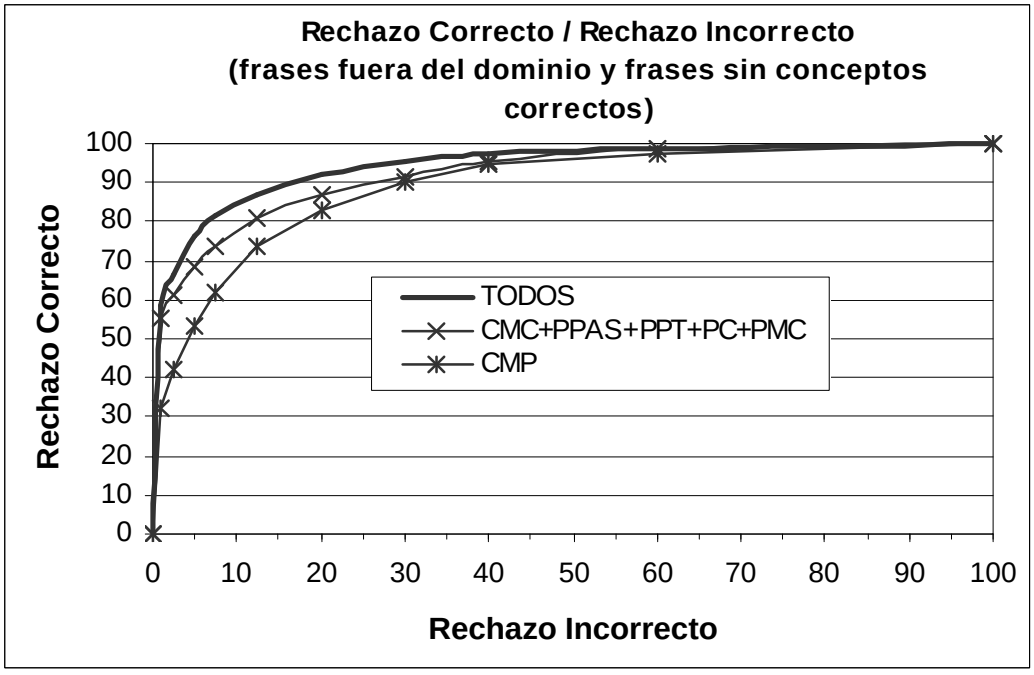

Figura 5-9: Rechazo Correcto vs. Rechazo Incorrecto (RI) para la detección de frases fuera del dominio de aplicación y frases sin ningún concepto correcto.

Considerando una distribución inicial en la que el 22,1% de las frases deben ser rechazadas, el poder de discriminación de estos parámetros permite reducir 12,1 puntos el error de clasificación (54,8% relativo).

5.2.6 Aplicación de medidas de confianza para la recuperación de errores

Cuando se utilizan las medidas de confianza para rechazar palabras mal reconocidas, conceptos erróneos o frases incorrectas, estamos reduciendo la tasa de error pero a costa de perder cierta información. Una posibilidad para recuperar parte de esta información es combinar varias hipótesis de uno o varios reconocedores. En este apartado, propondremos varios métodos para la combinación de varias hipótesis de reconocimiento utilizando las medidas de confianza al nivel de palabra como heurístico de combinación.

5.2.6.1 Métodos para la combinación de hipótesis

Los métodos considerados para la combinación de hipótesis de reconocimiento son los siguientes:

- **Reordenación de las hipótesis (FLCR: Flat List Confidence Rescoring).**

  Para cada hipótesis obtenida del banco de reconocedores, calculamos la confianza media por palabra a lo largo de la frase. La hipótesis con mayor confianza media será la seleccionada como la mejor hipótesis.

- **Reordenación basada en un grafo de palabras (WGCR: Word Graph Confidence Rescoring).**

  En este caso la idea principal es construir un grafo de palabras con las diferentes hipótesis y buscar el camino a lo largo del grafo, con la mayor confianza media por palabra. Esta búsqueda puede producir una nueva hipótesis, diferente de las anteriores. El objetivo fundamental de esta opción es ir cogiendo de cada hipótesis, las partes que tengan mayor confianza. Este método consiste en dos pasos: generación del grafo de palabras y búsqueda del mejor camino en el grafo. En el siguiente apartado se describen dos algoritmos diferentes para la generación de grafos de palabras a partir de varias hipótesis de reconocimiento. En cuanto a la búsqueda del mejor camino, utilizaremos un algoritmo de programación dinámica donde el heurístico considerado es la *Confianza Media Acumulada*, es decir, la confianza media desde el nodo inicial hasta el nodo actual.

Considerando estas mismas ideas, los dos métodos FLCR y WGCR, se pueden extender al nivel de concepto. En este caso los denominaremos FLCR y CGCR (Concept Graph Confidence Rescoring). En el apartado 5.2.6.3 presentaremos los resultados tanto al nivel de palabra como al nivel de concepto.

5.2.6.2 Algoritmos para la generación de un grafo de palabras a partir de varias hipótesis de reconocimiento.

Los métodos considerados para la generación de un grafo de palabras a partir de varias hipótesis de reconocimiento son los siguientes:

1.- Algoritmo basado en la segmentación temporal de las palabras.

Este algoritmo se basa en los límites temporales de las palabras de cada hipótesis para construir un grafo a partir de ellas (Singh et al, 2001). El primer paso de este algoritmo es etiquetar cada una de las palabras pertenecientes a las hipótesis con su valor de confianza, y sus marcas de inicio y fin. En la figura 5-10, podemos ver un ejemplo en el que se muestra la frase de referencia y tres posibles hipótesis. En esta figura las palabras están representadas por transiciones entre nodos.

El siguiente paso es unir los nodos de las diferentes hipótesis para construir un grafo de palabras. En primer lugar se unen los nodos inicio y fin de las secuencias, para posteriormente unir los nodos de todas las palabras situadas en las mismas posiciones en la frase. Es decir, para todas las palabras que comiencen/terminen en las mismas tramas, se deben unir los nodos anteriores/posteriores a dicha palabra (por ejemplo, las palabras I, I'M y I comienzan en la trama 0 y las tres terminan en la trama 10, luego los nodos

anteriores y posteriores de las tres palabras se deben unir). Una vez generado un primer grafo, se eliminan las transiciones paralelas quedándonos con la transición de mayor confianza.

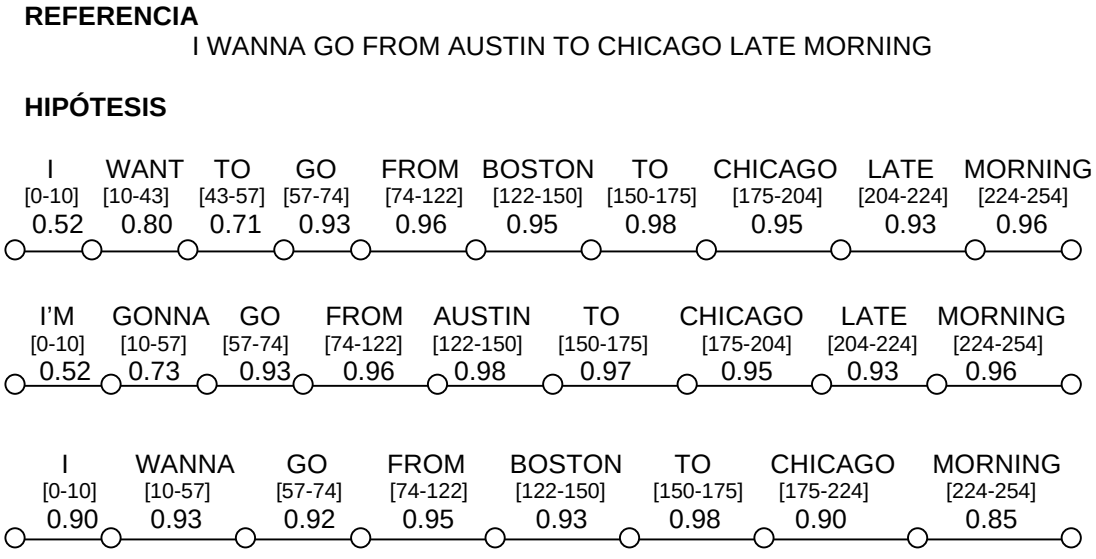

Figura 5-10: Ejemplo de frase de referencia y de tres posibles hipótesis. Cada palabra se etiqueta con la confianza obtenida y las marcas de trama inicio y final.

En la figura 5-11 podemos ver el grafo resultado del ejemplo de la figura 5-10. En el grafo resultante podemos obtener la hipótesis final sin más que seguir las flechas marcadas con línea más gruesa. En este caso podemos ver cómo la nueva hipótesis resultante encaja perfectamente con la frase de referencia.

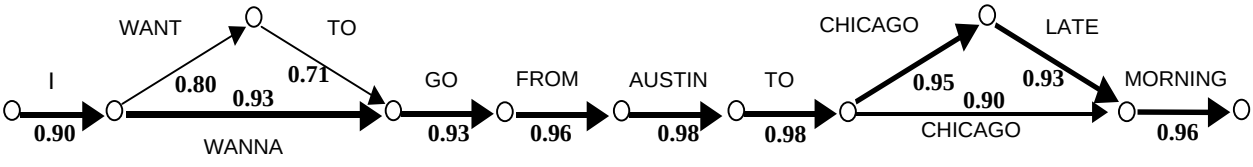

Figura 5-11: Grafo resultado del ejemplo presentado en la figura 5-10.

En esta descripción del algoritmo, hemos impuesto la unión de nodos cuando las marcas temporales (en este caso en tramas) coinciden exactamente. Podríamos relajar esta condición y permitir ciertas variaciones entre marcas. En el apartado 5.4.6 podremos ver algunos experimentos sobre el reconocedor de fechas y horas.

El inconveniente de este algoritmo es que requiere disponer de las marcas temporales para delimitar cada palabra de la secuencia. Estas marcas pueden no estar disponibles cuando se combinan hipótesis procedentes de varios decodificadores con interfaces de salida diferentes. En estos casos es necesario recurrir al siguiente algoritmo.

2.- Algoritmo basado en el alineamiento de texto entre hipótesis.

El punto de partida es el mismo que el comentado en el algoritmo anterior: varias hipótesis de reconocimiento con las palabras etiquetadas con su valor de confianza. A la hora de unir nodos entre hipótesis diferentes, en lugar de considerar el criterio de marcas temporales, debemos hacer un alineamiento de texto entre pares de hipótesis. Este alineamiento es el mismo que el que se realiza para comparar una hipótesis con la frase de referencia y obtener así las palabras correctas, sustituidas, insertadas y borradas.

Una vez alineadas todas las hipótesis dos a dos se procede a la unión de nodos entre ellas. En primer lugar se unen los nodos inicio y fin de las secuencias, para posteriormente unir los nodos iniciales de todas las palabras situadas en las mismas posiciones en la frase. En la unión de estos nodos se deben tener en cuenta las siguientes restricciones:

- Las palabras a unir deben ser iguales. La razón de forzar esta condición es que si permitimos unir palabras diferentes se pueden definir caminos en el grafo con incoherencias gramaticales.

- La segunda restricción es que no podemos unir palabras que tengan diferentes comportamientos cuando se alinean con otras hipótesis. Por ejemplo, en la figura 5-12, no podríamos unir la palabra “TO” de las hipótesis **B** y **C**, porque existe una hipótesis **A**, donde estas dos palabras fueron alineadas con diferentes palabras: en un caso se produjo una sustitución con la palabra “GO” y en el otro caso eran la misma palabra y fueron unidas.

  Esta restricción tiene como objetivo principal la generación de un grafo de palabras sin bucles ya que estos impedirían el cálculo del mejor camino a través del grafo mediante un algoritmo de programación dinámica.

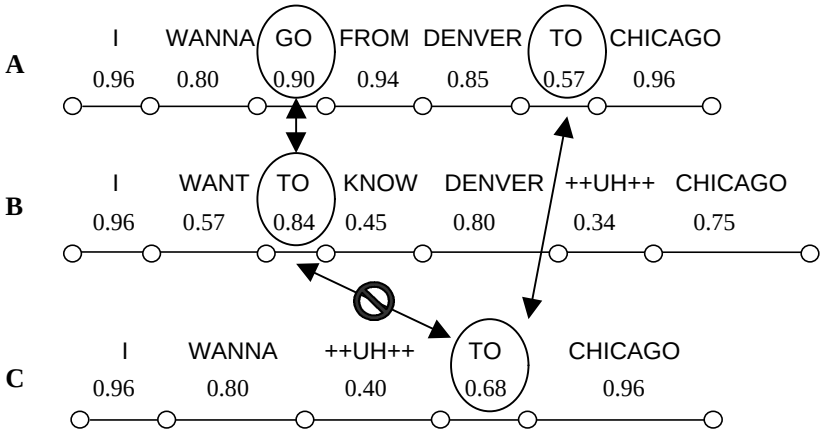

Figura 5-12: Ejemplo de alineamiento entre hipótesis que no genera una unión de nodos por no cumplirse la segunda restricción.

En la figura 5-13 podemos ver el grafo resultante para este ejemplo.

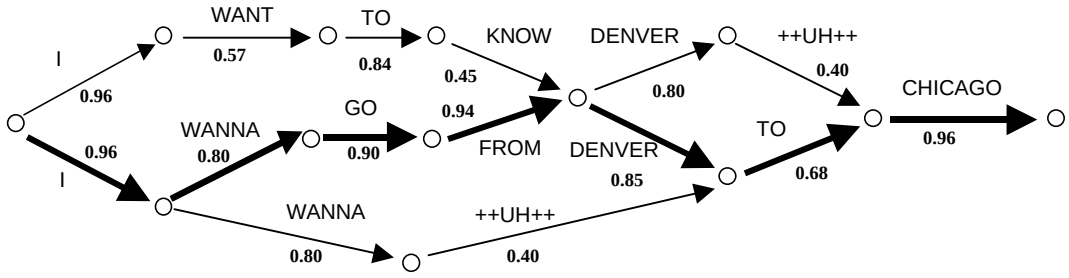

Figura 5-13: Grafo resultado del ejemplo presentado en la figura 5-12.

Este algoritmo es similar al denominado ROVER (Recognizer Output Voting Error Reduction) (Fiscus, 1997). En ROVER se construye una Red simple de Transcripciones entre Palabras (Word Transcription Network: WTN) definiendo nodos de interconexión de hipótesis entre TODAS las palabras. En la figura 5-14, podemos ver el grafo resultante para el ejemplo anterior aplicando ROVER.

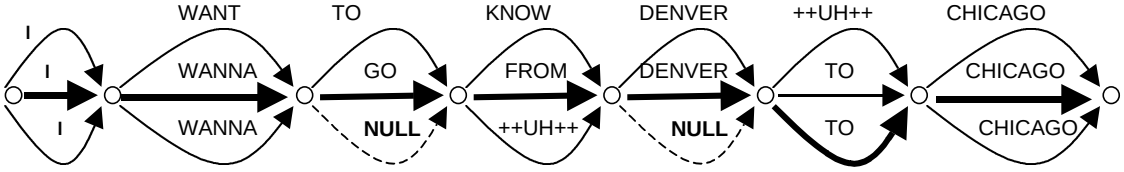

Figura 5-14: Grafo obtenido con el algoritmo ROVER del ejemplo presentado en la figura 5-12.

El alineamiento de las hipótesis se realiza de igual forma, la única diferencia es que se exige la unión de todos los nodos entre palabras. Para resolver el problema de los borrados e inserciones de palabras cuando se comparan hipótesis de diferente longitud, se consideran transiciones nulas (NULL).

ROVER permite generar un grafo con mayor flexibilidad y mayor número de caminos posibles (el grafo de la figura 5-13 está incluido en el grafo de la figura 5-14), pero tiene el inconveniente de que puede dar lugar a frases con incoherencias gramaticales. Este problema se agudiza cuando el número de hipótesis a combinar es reducido (menor de 5). En el caso de que se combine un número de hipótesis mayor, el peso estadístico de las posibles incoherencias será pequeño en comparación con las secuencias de palabras más frecuentes en el modelo de lenguaje. En el trabajo de Schwenk (Schwenk et al, 2000), los autores presentan un análisis detallado de la incorporación de la probabilidad del modelo de lenguaje en el algoritmo ROVER. En este trabajo consiguen reducciones relativas del error del 5% cuando se combinan pocas hipótesis (hasta 5 hipótesis). En el caso de un mayor número de hipótesis las mejoras son menores debido a que la información del modelo de lenguaje aparece de forma implícita en las estadísticas de las secuencias de palabras.

Generalmente en los experimentos realizados siempre utilizaremos el primero de los algoritmos porque ofrece mejores resultados como veremos a continuación. Por otro lado, a la hora de combinar secuencias de conceptos en lugar de palabras, no podemos

utilizar las marcas temporales para delimitar los conceptos puesto que no siempre hay una relación estricta entre la secuencia de conceptos y las marcas temporales de las palabras. En este caso es obligado utilizar el segundo algoritmo que realiza un alineamiento de hipótesis basado en texto.

5.2.6.3 Experimentos al nivel de palabra

La primera idea que nos planteamos fue utilizar estas técnicas, descritas anteriormente, para combinar las N mejores hipótesis de reconocimiento obtenidas de un mismo reconocedor. Para los resultados presentados en este apartado y en el siguiente, se ha utilizado un conjunto de datos independiente del utilizado para entrenar y evaluar las medidas de confianza. Este conjunto se ha obtenido de la evaluación realizada por NIST en Junio de 2000 (Pellom et al, 2000)(ver apéndice A, apartado A.1.4). En este caso tenemos 1.264 frases con un total de 3.014 palabras.

Considerando que las N hipótesis son posibles salidas del reconocedor, se obtuvieron los valores de confianza de todas las palabras de cada hipótesis. En la tabla 5-6 se presentan la Tasa de Error de Reconocimiento para el sistema base, y la tasa obtenida combinando las N mejores hipótesis (N=2, 4, 8 y 16) utilizando FLCR y WGCR.

Como podemos observar, a medida que aumenta el número de hipótesis la tasa de error también aumenta. Las causas que justifican este comportamiento son las siguientes:

- Al aumentar el número de hipótesis consideradas, la suposición de que cada una de ellas corresponde con una salida posible del reconocedor es menos cierta. El hecho de tener que generar varias hipótesis distintas obliga al reconocedor a considerar ciertas variaciones en la secuencia de palabras que pueden llegar a alejarse bastante del mejor camino. En este caso se producen ciertos errores un tanto artificiales que pueden no responder a problemas concretos del reconocedor considerado, sino a la necesidad de tener que generar varias hipótesis.

- Por otro lado, las medidas de confianza obtenidas se han entrenado considerando los errores que aparecen en la primera hipótesis. Por tanto, estas medidas no modelarán correctamente los errores que aparecen en la secuencia enésima.

Para el cálculo de las medidas de confianza que estamos utilizando al nivel de palabra, se han utilizado parámetros del proceso de decodificación y del modelo de lenguaje. Estas mismas fuentes de información se han utilizado en el proceso de decodificación de la señal de voz. En estas circunstancias podemos pensar que estas medidas poco podrán hacer para mejorar la tasa de reconocimiento reordenando las hipótesis que el propio reconocedor nos ha ofrecido de forma ordenada. Si analizamos en detalle este hecho, no se puede afirmar rotundamente esta posibilidad puesto que en este caso estamos introduciendo un nuevo conocimiento. Este conocimiento es el obtenido del modelado del error resultante al entrenar la Red Neuronal. Esta red puede aprender ciertos patrones de error que permitan detectar y/o corregir algunos errores. A pesar de ello, parece lógico pensar que la mejora que podemos obtener, utilizando un

subconjunto de parámetros tan reducido en comparación con los involucrados en el proceso de reconocimiento, sea muy pequeña o nula como ocurre en nuestro caso.

Tasa de Error obtenida al combinar las N mejores hipótesis de reconocimiento con FLCR y WGCR		
Tasa de Error de Referencia (Reconocimiento): 27,2%		
Número de Hipótesis	**FLCR**	**WGCR**
2	29,2%	29,2%
4	30,8%	31,2%
8	32,2%	32,6%
16	33,5%	33,6%

Tabla 5-6: Resultados de Tasa de Error combinando las N mejores hipótesis de reconocimiento con los métodos FLCG y WGCR.

En la tabla 5-7 se presentan los resultados para 5 hipótesis. En esta tabla se muestran también las mínimas tasas de error que se podrían conseguir realizando una combinación ideal de estas 5 hipótesis. Esta tasa de error mínima se ha calculado de forma diferente para el caso del FLCR y el WGCR. En el primer caso se seleccionará directamente la hipótesis con menor error (conocida la frase de referencia). En el caso del WGCR, la situación de combinación ideal se ha simulado considerando que la confianza de una palabra correcta es 1 y la de una palabra incorrecta 0. En el caso del WGCR, se presentan las tasas de error mínimas para los dos algoritmos de generación de grafos comentados en el apartado 5.2.6.2.

Tasa de Error utilizando FLCR y WGCR para la combinación de las 5 mejores hipótesis de reconocimiento			
Tasa de Error de Referencia: 27,2%			
	FLCR	**WGCR**	
		Alineamiento temporal	**Alineamiento de texto**
Utilizando Medidas de Confianza	31,0%	31,4%	31,6%
Caso Ideal	21,3%	19,4%	20,6%

Tabla 5-7: Tasa de Error utilizando los métodos FLCR y WGCR para combinar las 5 mejores hipótesis del reconocedor y las Tasas de Error mínimas que se podrían conseguir con una combinación ideal de las 5 hipótesis.

Como se puede observar para el caso ideal, el método WGCR permite obtener potencialmente menores tasas de error, siendo el alineamiento temporal el mejor criterio para generar el grafo de palabras a partir de varias hipótesis. La segunda conclusión que podemos sacar de la tabla 5-7 es que en las 5 mejores secuencias de palabras

disponemos de mucha información para reducir considerablemente la Tasa de Error. Dado que los parámetros utilizados (del proceso de decodificación y del modelo de lenguaje) no permiten reducir esta tasa, se debería recurrir a fuentes de conocimiento adicionales a las utilizadas por el reconocedor como modelos acústicos adaptados al sexo del hablante, modelos de lenguaje más potentes o redes semánticas adaptadas al dominio de la aplicación. En los siguientes experimentos trabajaremos con hipótesis de reconocimiento obtenidas de decodificadores diferentes: uno de ellos utiliza modelos independientes del sexo y otro utiliza modelos adaptados a la voz femenina.

A la hora de combinar hipótesis de diferentes reconocedores hemos realizado experimentos con cuatro configuraciones distintas. El sistema de referencia (Pellom et al, 2000) evaluado por NIST, comenzaba ejecutando dos reconocedores en paralelo: uno con modelos independientes del sexo y otro con modelos adaptados a la voz femenina. Después de 500 tramas de voz (5 segundos), el sistema seleccionaba uno de los reconocedores que mantenía activo el resto del diálogo, desactivando el otro. En este trabajo hemos considerado ambos reconocedores funcionando en paralelo constantemente. A la hora de combinar las hipótesis hemos utilizado por un lado la verosimilitud acumulada a lo largo de la frase, y por otro lado, las medidas de confianza desarrolladas en la presente tesis (utilizando los métodos FLCR y WGCR). Los resultados se presentan en la tabla 5-8.

De estos resultados podemos deducir que las medidas de confianza son de gran utilidad para reducir la Tasa de Error mediante la combinación de hipótesis de diferentes reconocedores, siendo mejor heurístico que la verosimilitud acumulada.

Tasa de Error conseguida combinando hipótesis de diferentes reconocedores	
Método utilizado	**Tasa de Error**
Referencia	27,2%
Verosimilitud acumulada	26,2%
Confianza: FLCR	24,2%
Confianza: WGCR	**23,4%**

Tabla 5-8: Tasa de Error obtenida combinando hipótesis de diferentes reconocedores.

Considerando el método WGCR propuesto en esta tesis, podemos alcanzar una reducción significativa de 3,8 puntos (14% relativo) en la Tasa de Error (la banda de fiabilidad de estos experimentos es de un 3%). Este método funciona mejor que el FLCR. En estos experimentos, la diferencia entre ambos métodos no es muy grande (0,8 puntos) porque el número de hipótesis combinadas es pequeño, sólo dos, y la longitud media por frase también es pequeña: 2,4 palabras por frase. En estos casos la generación de un grafo es bastante parecido a considerar las hipótesis por separado. Las Tasas de Error mínimas (considerando medidas de confianza ideales) fueron 19,9% y 18,6% para FLCR y WGCR respectivamente. Otro detalle importante a comentar es que utilizando únicamente dos hipótesis, frente a las 5 utilizadas en la tabla 5-7, la Tasa de Error

mínima (caso ideal) es menor para ambos métodos. En este caso estamos utilizando una nueva fuente de información: modelos acústicos adaptados a la voz femenina.

La siguiente idea que probamos fue combinar varias hipótesis de los dos reconocedores. Los resultados obtenidos se muestran en la tabla 5-9.

Tasa de Error combinando varias hipótesis de reconocedores diferentes, Referencia: 27,2%,		
Número de hipótesis	**FLCR**	**WGCR**
1	24,2%	23,4%
2	**23,3%**	**22,9%**
4	29,4%	26,4%
8	30,5%	28,0%
16	31,4%	28,7%

Tabla 5-9: Tasa de Error obtenida combinando varias hipótesis de diferentes reconocedores.

Cuando el número de hipótesis consideradas es bajo, conseguimos mejorar ligeramente la Tasa de Error. La introducción de conocimiento más potente, como son los modelos acústicos adaptados a la voz femenina, ha permitido reducir ligeramente la tasa de error al combinar varias hipótesis de los mismos reconocedores. Pero este valor aumenta al incrementar por encima de dos el número de hipótesis combinadas. La causa de este efecto es la misma que la comentada anteriormente; la suposición de que las N cadenas son ejemplos de reconocimiento válidos empieza a ser falsa cuando el número de hipótesis aumenta. Considerando los resultados obtenidos para el caso del WGCR, conseguimos una reducción de 4,3 puntos en la Tasa de Error (16% relativo) sobre la tasa de referencia.

5.2.6.4 Experimentos al nivel de concepto semántico

Sobre el mismo corpus utilizado en el apartado anterior, nos planteamos la idea de reducir, utilizando medidas de confianza, la Tasa de Error al nivel de Concepto (CER: Concept Error Rate). En este corpus la CER de referencia es de 25,1%. Para este caso, hemos planteado dos alternativas:

- La primera opción es considerar la mejor frase obtenida de combinar las dos mejores hipótesis de los dos reconocedores (uno con modelos acústicos independientes del sexo y otro con modelos adaptados a la voz de mujer) mediante el método WGCR. Una vez obtenida la mejor frase se pasa por el analizador semántico. En este caso la CER obtenida finalmente ha sido 24,0%, lo que supone una reducción relativa de 4,4%.

- En la segunda opción se analizaron semánticamente las dos mejores hipótesis de los dos reconocedores, se calcularon medidas de confianza para cada uno de los

conceptos obtenidos y se utilizaron dichas medidas como heurístico para la combinación de las secuencias de conceptos mediante los métodos FLCR y CGCR (Concept Graph Confidence Rescoring, análogo al WGCR con generación del grafo por alineamiento de texto). Los resultados se presentan en la tabla 5-10.

Tasa de Error de Concepto combinando las dos mejores hipótesis de los dos reconocedores	
Método de combinación	**Tasa de Error**
Referencia	25,1%
FLCR	24,6%
CGCR	24,6%
FLCR (caso ideal)	20,6%
CGCR (caso ideal)	19,8%

Tabla 5-10: Tasa de Error de Concepto obtenida combinando las dos mejores hipótesis de diferentes reconocedores.

En este caso no hay diferencias entre los métodos FLCR y CGCR porque las secuencias de conceptos son más cortas (1,4 conceptos por frase) que las secuencias de palabras, lo que hace que ambos comportamientos sean muy parecidos. En este caso las Tasas de Error Ideales muestran aún un amplio margen de mejora.

Aunque las diferencias no son significativas, los resultados muestran que la primera de las opciones, que consiste en aumentar cuanto antes la tasa de reconocimiento, funciona mejor que intentar recuperar errores en las secuencias de conceptos utilizando medidas de confianza a este nivel.

5.3 Medidas de Confianza sobre el reconocedor de nombres deletreados

El sistema de reconocimiento de nombres deletreados sobre el que hemos trabajado es el desarrollado en la presente tesis doctoral. Este sistema tiene una estructura de reconocimiento basada en un esquema de hipótesis y verificación. La fase de hipótesis consiste en dos pasos: obtención de las N mejores cadenas de letras y la comparación posterior de estas cadenas con los nombres del diccionario para obtener los más parecidos a dichas secuencias. En la fase de verificación se utilizan estos nombres del diccionario para volver a realizar un proceso de reconocimiento completo y seleccionar finalmente el nombre reconocido (para más detalles ver capítulo 3). A la hora de obtener medidas de confianza trabajaremos con parámetros obtenidos tanto de la parte de hipótesis como de la de verificación.

En este sistema trabajaremos al nivel de frase completa, o secuencia de letras completa. Las medidas de confianza sobre este sistema intentarán informar sobre la certeza en el reconocimiento del nombre completo deletreado y no de cada una de las letras que lo componen. Además en este caso aplicaremos las medidas de confianza para detectar nombres que no están en el diccionario: OOV (Out of Vocabulary)

5.3.1 Base de datos

La base de datos utilizada para los experimentos realizados sobre este sistema es SpeechDat (Moreno, 1997), la misma que la utilizada para entrenar y evaluar el reconocedor de nombres deletreados. En esta base de datos se dispone de 3.000 ficheros de voz con nombres de ciudad, nombres de pila y cadenas de letras aleatorias, deletreadas por 1.000 locutores diferentes a través de la red telefónica fija. Como comentamos en el capítulo 3, para realizar los experimentos de tasa de reconocimiento sobre este sistema, seleccionamos aleatoriamente 600 ficheros para considerar dos conjuntos de validación y ajuste de parámetros intermedios (300 ficheros para cada conjunto) y 300 ficheros para evaluar la tasa final del reconocedor, dejando alrededor de 2.100 para entrenar los modelos acústicos. Esta distribución se repitió 6 veces mediante un algoritmo de Round Robin. De esta forma se evaluó el sistema con 1.800 ficheros diferentes. Esta lista de ficheros será la utilizada para evaluar las medidas de confianza obtenidas en este reconocedor. A la hora de realizar los experimentos de este capítulo, hemos dividido aleatoriamente el conjunto de 1.800 nombres deletreados en tres subconjuntos: 66% de los nombres para el entrenamiento de la Red Neuronal utilizada en la combinación de los parámetros (ver apartado 5.3.3), 17% para su validación y el 17% para evaluación. Esta división se ha repetido 6 veces realizando un proceso Round-Robin, de forma que cada vez, se van utilizando unos datos diferentes para entrenar, validar o evaluar la Red Neuronal, consiguiendo que se usen todos los datos disponibles una vez para evaluar. Los resultados presentados en este apartado son la media de los valores obtenidos en todos los experimentos. El intervalo de confianza para el Error de Clasificación (al 95%) es menor del ±1,4%. Este valor se ha obtenido con la fórmula presentada en el apartado 3.3.1.1, donde la variable **p** es el Error de Clasificación de referencia 90,3% y **n** es el número de ejemplos de evaluación; 1800. Los intervalos de confianza para los valores de Rechazo Correcto se presentan en las tablas de resultados.

La etiquetación de cada ejemplo como acierto o fallo se realiza de forma automática comparando el nombre reconocido con el nombre deletreado. Los resultados de reconocimiento utilizados son los obtenidos para el diccionario de 10.000 nombres (San-Segundo et al, 2002). En este caso la tasa de reconocimiento del sistema fue del 90,3% (Tasa de Error de 9,7%). Esta tasa de error es nuestro Error de Referencia para evaluar nuestras medidas de confianza. Esta evaluación se realizará considerando las mismas medidas descritas en el apartado 5.2.3.

5.3.2 Parámetros de medida de confianza

Los parámetros considerados se han obtenido de las dos partes de las que consta el reconocedor: hipótesis y verificación. En los trabajos de Macías-Guarasa (Macías-Guarasa et al, 2000b; Macías-Guarasa, 2001) podemos consultar un análisis de medidas

de confianza sobre un sistema similar, dividido también en dos fases: hipótesis y verificación, pero desarrollado para el reconocimiento de habla aislada y/o expresiones cortas. A continuación se describen los parámetros utilizados en nuestro sistema de reconocimiento (San-Segundo et al, 2001b). Para cada uno de los parámetros se ha definido un código de forma que sea fácil su identificación en la tabla de resultados.

5.3.2.1 Parámetros de la etapa de Hipótesis

Esta etapa está formada por dos pasos: obtención de las N mejores cadenas de letras y comparación de dichas cadenas con los nombres del diccionario. Al conjunto de parámetros obtenidos del primer paso lo referenciaremos con el código H-1 y son los siguientes:

- **Verosimilitud Acumulada por trama para la mejor cadena de letras (VA1-1):** es el logaritmo de la verosimilitud obtenida para la mejor secuencia de letras dividido por el número de tramas. En casos de error o de nombre fuera del diccionario, esta verosimilitud debe ser generalmente menor que los casos en lo que tengamos un acierto de reconocimiento.

- **Diferencia de Verosimilitud Acumulada (DVA-1):** es la diferencia del logaritmo de las verosimilitudes obtenidas para la primera y segunda cadena de letras, dividida por el número de tramas. En casos de mayor diferencia la confianza sobre lo reconocido será generalmente mayor pues revelará una mayor diferencia acústica entre los candidatos.

Los parámetros obtenidos del proceso de comparación de las secuencias de letras con los nombres del diccionario, son los siguientes (H-2):

- **Mejor Coste de Alineamiento (MCA-2):** distancia léxica mínima entre la mejor secuencia de letras y los nombres del diccionario, dividido por la longitud de la cadena de letras. Ante una distancia léxica menor se refleja una mayor certeza del reconocimiento.

- **Diferencia entre los dos mejores Costes de Alineamiento (DCA-2):** diferencia entre la distancia de la mejor cadena de letras con los dos nombres del diccionario más cercanos, dividido por la longitud de la cadena de letras. De forma análoga al parámetro DVA-1, una mayor diferencia revela generalmente una mayor confianza en el nombre reconocido.

- **Coste de Alineamiento Medio para los 50 mejores nombres del diccionario (CAM-2):** es el coste medio de alineamiento entre la mejor cadena de letras y los 50 mejores nombres del diccionario, divido por la longitud de la secuencia de letras.

- **Varianza de Costes de Alineamiento para los 50 mejores nombres del diccionario (VCA-2):** es la varianza de costes para los 50 mejores nombres dividida por la longitud de la mejor secuencia de letras. En este caso, varianzas

de coste mayores revelan una mayor diferencia entre los nombres seleccionados, lo que pone de manifiesto una mayor facilidad de reconocimiento, y por tanto, una mayor certeza.

Macías (Macías-Guarasa, 2001) hace un análisis más detallado de estos dos últimos parámetros en el que se varía el número de nombres candidatos sobre los que se realiza la media y la varianza. En nuestro caso hemos considerado 50 candidatos porque es el número de nombres que el reconocedor extrae del módulo de hipótesis y los pasa al módulo de verificación para su posterior reordenación.

5.3.2.2 Parámetros de la etapa de Verificación

Al conjunto de parámetros extraídos de este módulo los referenciaremos en las tablas de resultados con el código V-3, y son los siguientes:

- **Verosimilitud Acumulada por trama para el mejor Nombre reconocido (VAN-3):** es el logaritmo de la verosimilitud acumulada en el reconocimiento del mejor nombre, una vez reordenados los nombres en la etapa de verificación, y dividido por el número de tramas. Al igual que el parámetro VA1-1, mayores verosimilitudes reflejan una mayor confianza en lo reconocido.

- **Diferencia de Verosimilitud para los dos mejores nombres (DVA-3):** diferencia de verosimilitudes para los dos mejores nombres reconocidos en la etapa de verificación, divida por el número de tramas.

- **Verosimilitud media por trama de los 50 nombres (VM-3):** es la media del logaritmo de la verosimilitud para los 50 nombres reordenados en la etapa de verificación, dividida por el número de tramas.

- **Varianza de la Verosimilitud para los 50 nombres candidato (VV-3):** es la varianza del logaritmo de la verosimilitud obtenida para los 50 nombres reordenados, dividida por el número de tramas de voz.

  Los comportamientos esperados para estos dos últimos parámetros y su relación con las medidas de confianza son equivalentes a los previstos para los parámetros CAM-2 y VCA-2.

- **Diferencia de Verosimilitudes entre Módulos (DVM-3):** es la diferencia entre los logaritmos de la verosimilitud obtenida para la mejor secuencia de letras (en la fase de hipótesis, parámetro VA1-1), y la verosimilitud del mejor nombre reconocido en la fase de verificación (parámetro VAN-3), dividido por el número de tramas.

5.3.3 Combinación de parámetros.

Al igual que en el caso anterior, hemos considerado una Red Neuronal para combinar los parámetros y obtener una única medida de confianza. La red utilizada ha sido un

Perceptrón Multicapa sencillo. En este caso, debido a que no tenemos muchos datos de entrenamiento no hemos realizado ninguna codificación de los parámetros para no incrementar en exceso el número de pesos de la red. Estos parámetros los aplicaremos directamente a las entradas de la Red Neuronal. En este caso es necesario hacer un reescalado para adecuar el rango de cada parámetro al intervalo [0-1]. La capa oculta está formada por 10 neuronas y la capa de salida por una única neurona. En el entrenamiento de los pesos se etiqueta la salida con un 1 cuando tenemos un acierto de reconocimiento y con un 0 cuando es un error o es una palabra fuera del diccionario (OOV), según sea el comportamiento a analizar.

5.3.4 Detección de errores de reconocimiento

En este apartado analizaremos el poder de discriminación de los parámetros comentados en 5.3.2 para la detección de errores de reconocimiento. La tabla 5-11 presenta las tasas de Rechazo Correcto (ver definición en apartado 5.2.4) para tasas de Rechazo Incorrecto (ver apartado 5.2.3) del 2,5% y del 5.0%% y el Error de Clasificación (con sus dos contribuciones ECa y ECe) para el caso de RI del 5%. En esta tabla también se presenta el Error de Clasificación de Referencia y el mínimo obtenido en cada caso.

Detección de Errores de reconocimiento. Error de Referencia: 9.7%							
Parámetros		**Rechazo Correcto (%)**		**5,0% RI**			**Mínimo Error**
		2,5% RI	**5,0% RI**	**ECe**	**ECa**	**EC**	
Hipótesis	**H-1**	**7,1 (±3,8%)**	**12,9 (±5%)**	**8,4%**	**4,5%**	**12,9%**	**9,7%**
	MCA-2 DCA-2	20,0 (±5,9%)	26,5 (±6,5%)	7,1%	4,5%	11,6%	9,4%
	CAM-2 VCA-2	18,3 (±5,7%)	26,0 (±6,5%)	7,2%	4,5%	11,7%	9,4%
	H-2	**22,3 (±6,2%)**	**29,5 (±6,8%)**	**6,8%**	**4,5%**	**11,3%**	**9,2%**
Verificación	VAN-3 DVA-3	40,5 (±7,3%)	54,3 (±7,4%)	4,4%	4,5%	8,9%	8,0%
	VM-3 VV-3	27,1 (±6,6%)	38,2 (±7,2%)	6,0%	4,5%	10,5%	9,0%
	DVM-3	30,1 (±6,8%)	37,4 (±7,2%)	6,0%	4,5%	10,5%	9,0%
	V-3	**46,7 (±7,4%)**	**57,4 (±7,3%)**	**4,1%**	**4,5%**	**8,6%**	**7,6%**
	H-2 y V-3	**44,7(±7,4%)**	**57,9 (±7,3%)**	**4,1%**	**4,5%**	**8,6%**	**7,5%**

Tabla 5-11: Rechazo Correcto de errores para Rechazos Incorrectos del 2,5% y 5% considerando parámetros de forma aislada o grupos reducidos de parámetros. También se muestran los Errores de Clasificación para RI del 5,0%, Mínimo Error de Clasificación y el Error de Referencia.

Como se puede observar, los parámetros de la fase de verificación presentan un mayor poder de discriminación. Considerando todos los parámetros de la etapa de verificación, V-3, podemos detectar el 57,4% de los errores con un RI del 5%. Este valor es muy similar al obtenido para el caso en el que combinamos los parámetros H-2

y los V-3, luego los parámetros H-2 aportan poca información adicional a la ya considera con el grupo de parámetros del proceso de verificación. Los parámetros H-1 no han sido considerados para el cálculo de la medida de confianza final, puesto que ofrecen muy bajo poder de discriminación. Considerando que el Error de Referencia es 9,7%, con estas medidas reducimos 2,2 puntos el Error de Clasificación (22,7% relativo). En la figura 5-15 se muestra la evolución del Rechazo Correcto para diferentes valores de Rechazo Incorrecto.

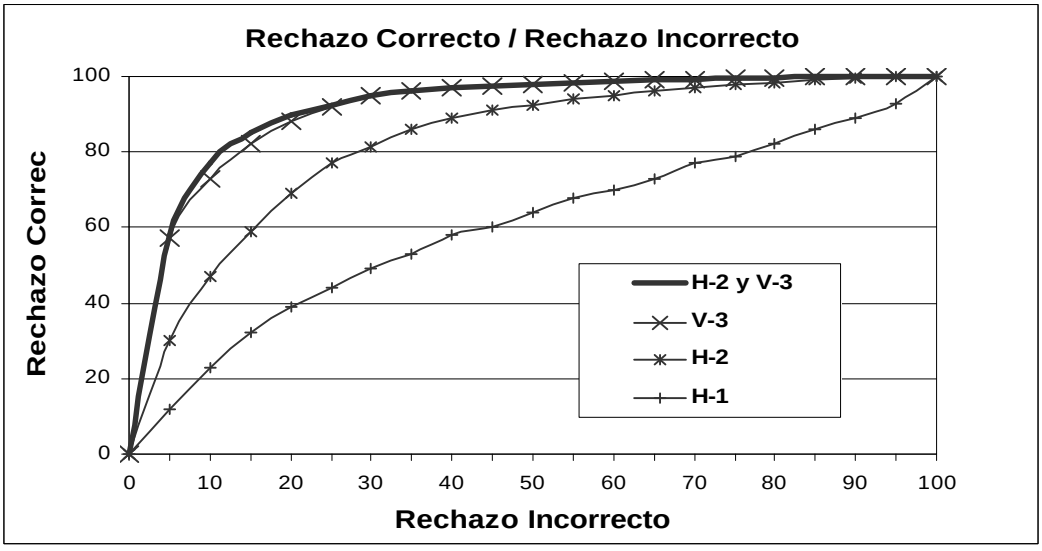

Figura 5-15: Rechazo Correcto vs. Rechazo Incorrecto (RI) para la detección de errores de reconocimiento considerando por separado parámetros de la etapa de hipótesis (H-1, H-2), de la etapa de verificación (V-3), y de ambas etapas (H-2 y V-3).

La principal conclusión que podemos deducir de los resultados presentados es que a medida que utilizamos parámetros de etapas de reconocimiento más avanzadas, el poder de discriminación aumenta. Este hecho tiene su justificación en que a medida que avanzamos en el proceso de reconocimiento, vamos aplicando información más potente (modelos acústicos más refinados o gramáticas más restrictivas). El comportamiento del reconocedor al usar esta información permite obtener medidas de confianza mejores. Por otro lado, las etapas finales son las responsables de decidir finalmente el resultado del reconocimiento, luego las medidas obtenidas en ellas, podrán reflejar mejor la confianza del resultado obtenido.

5.3.5 Detección de nombres fuera del diccionario de reconocimiento (OOV: Out of Vocabulary)

En este apartado analizamos el poder de clasificación de los parámetros comentados en 5.3.2 para la detección de nombres que no pertenecen al diccionario de reconocimiento. Para simular la pronunciación de nombres fuera del vocabulario eliminamos del diccionario nombres aleatoriamente de forma que, para el 21,5% de los ficheros, el nombre deletreado no se encontrase en el diccionario. En estos experimentos se intentará detectar estos casos. En la tabla 5-12 se resumen los resultados obtenidos, y en la figura 5-15 se muestra la evolución del RC con el RI.

Detección de nombres fuera del diccionario (Error de Referencia: 21,5%)							
Parámetros		Rechazo Correcto (%)		5,0% RI			Mínimo Error
		2,5% RI	5,0% RI	ECe	ECa	EC	
Hipótesis	**H-1**	**2,9 (±1,7%)**	**5,7 (±2,3%)**	**20,3%**	**3,9%**	**24,2%**	**21,5%**
	MCA-2 DCA-2	17,6 (±3,8%)	33,4 (±4,7%)	14,3%	3,9%	18,2%	17,7%
	CAM-2 VCA-2	3,0 (±1,7%)	5,3 (±2,2%)	20,4%	3,9%	24,1%	21,5%
	H-2	**17,5 (±3,8%)**	**34,5 (±4,7%)**	**14,1%**	**3,9%**	**18,0%**	**17,7%**
Verificación	VAN-3 DVA-3	9,3 (±2,9%)	15,5 (±3,6%)	18,2%	3,9%	21,1%	21,5%
	VM-3 VV-3	3,0 (±1,7%)	6,3 (±2,4%)	20,1%	3,9%	24,0%	21,5%
	DVM-3	53,0 (±5%)	66,3 (±4,7%)	7,3%	3,9%	11,2%	11,2%
	V-3	**53,5 (±5%)**	**67,9 (±4,7%)**	**7,0%**	**3,9%**	**10,9%**	**10,9%**
	H-2 y V-3	**56,2 (±5%)**	**68,3 (±4,6%)**	**7,0%**	**3,9%**	**10,9%**	**10,9%**

Tabla 5-12: Rechazo Correcto de nombres fuera del vocabulario para Rechazos Incorrectos del 2,5% y 5% considerando parámetros de forma aislada o grupos reducidos de parámetros. También se muestran los Errores de Clasificación para RI del 5,0%, Mínimo Error de Clasificación y el Error de Referencia.

La Diferencia de Verosimilitudes entre Módulos (DVM-3) es, sin lugar a dudas, el mejor parámetro para la detección de nombres fuera del vocabulario de reconocimiento. Utilizando únicamente este parámetro, podemos detectar más del 67% de nombres fuera del vocabulario de reconocimiento, rechazando incorrectamente sólo un 5% de nombres pertenecientes al diccionario. Considerando que el Error de Referencia es del 21,5%, con este único parámetro podemos reducir 10,6 puntos (50,7% relativo) el error mínimo de clasificación.

En el módulo de hipótesis realizamos un proceso de decodificación para obtener la secuencia de letras que mejor encaja acústicamente con la secuencia pronunciada por el usuario. En este módulo permitimos prácticamente el reconocimiento de cualquier secuencia de letras. En la fase de verificación se vuelve a repetir este proceso pero permitiendo únicamente secuencias de letras que correspondan con nombres del diccionario. La diferencia entre las verosimilitudes conseguidas de una u otra forma nos ofrece mucha información sobre si la secuencia de letras pronunciada corresponde o no con un nombre del diccionario. En casos de nombres no pertenecientes al diccionario, se obtienen valores altos de verosimilitud en la fase de hipótesis, y valores muy pequeños cuando se imponen las restricciones de los nombres del diccionario.

Al igual que en el apartado anterior, los mejores resultados se obtienen combinando los parámetros H-2 y V-3. En este caso detectamos un 68,3% de nombres fuera del diccionario para un rechazo incorrecto del 5%. Tampoco hemos considerado los parámetros H-1 debido a su bajo poder de discriminación. En la figura 5-12, podemos

observar como la gráfica de RC vs. RI para H-1 es prácticamente lineal luego el poder de discriminación de estos parámetros es nulo.

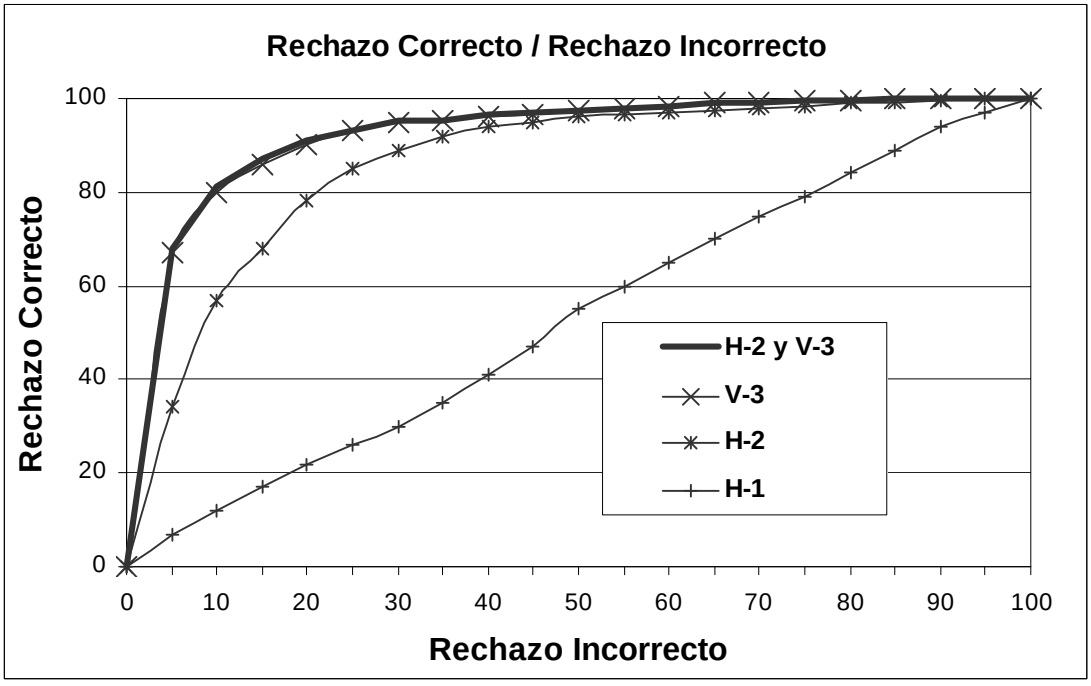

Figura 5-16: Rechazo Correcto (RC) vs. Rechazo Incorrecto (RI) para la detección de nombres fuera del diccionario, considerando por separado parámetros de la etapa de hipótesis (H-1, H-2), de la etapa de verificación (V-3), y de ambas etapas (H-2 y V-3).

5.3.6 Detección simultánea de errores y nombres fuera del diccionario de reconocimiento

A continuación presentamos los resultados para dos nuevos experimentos. En el primero utilizamos los conjuntos de parámetros H-2 y V-3 para detectar tanto errores de reconocimiento como nombres no pertenecientes al diccionario de reconocimiento.

Detección de errores y nombres fuera del diccionario (Error de Referencia: 29.2%)			
Parámetros	**Rechazo Correcto**		**Mínimo Error de Clasificación**
	2,5% RI	5,0% RI	
H-2 y V-3	54,8%	65,8%	13,1%

Tabla 5-13: Rechazo Correcto tanto de errores como de nombres fuera del vocabulario para Rechazos Incorrectos del 2,5% y 5% utilizando los parámetros H-2 y V-3.

Los resultados se muestran en la tabla 5-13. En este caso, la reducción relativa del error de clasificación es 55,1%, mayor que la obtenida en apartados anteriores.

En el segundo experimento intentamos discriminar entre aciertos, errores de reconocimiento y nombres fuera del diccionario. Para ello utilizamos también los parámetros de los grupos H-2 y V-3. En este caso hemos considerado tres neuronas de salida en el Perceptrón Multicapa utilizado. Durante el entrenamiento de los pesos, para los casos de acierto se puso a 1 la primera salida y el resto a 0, para el caso de errores se puso a 1 la segunda salida y el resto a 0, y para nombres fuera del diccionario, se activó la tercera neurona dejando a cero las dos primeras. Durante el proceso de evaluación se clasificará el ejemplo según la salida de mayor valor. En la tabla 5-14 se muestra la matriz de confusión resultante de este experimento.

Matriz de confusión (Error de Referencia: 29,2%)			
	Resultado de clasificación		
	Aciertos	**Errores**	**Nombres fuera del diccionario**
Aciertos	**94,9% (1,213)**	0,9% (13)	4,2% (52)
Errores	49,7% (68)	**18,0% (25)**	32,3% (44)
Nombres fuera del diccionario	24,0% (92)	3,6% (14)	**72,4% (279)**

Tabla 5-14: Matriz de confusión en la clasificación de un nombre reconocido como Acierto, Error o Nombre fuera del diccionario de reconocimiento.

Los mayores problemas de discriminación ocurren para los errores. En este caso la mayoría de los ejemplos se clasifican como aciertos o como nombres fuera del diccionario de reconocimiento. En el otro sentido, los ejemplos mejor clasificados son los aciertos para los que se consigue una Tasa de Error del 5,1%. El Error Total de Clasificación para este caso ha sido del 16,6%, lo que supone una reducción relativa del 43,2% sobre el error de referencia.

5.4 Medidas de confianza para el sistema de reconocimiento de fechas y horas

Por último, vamos a analizar las medidas de confianza sobre el reconocedor de fechas y horas desarrollado también en la presente tesis. Este sistema de reconocimiento pretende la decodificación de frases que expresan fechas u horas de forma continua. En el desarrollo de este reconocedor (capítulo 4) se analizó de forma separada la tarea de reconocimiento dependiendo del tipo de habla: leída o espontánea. En este apartado respetaremos esa división de forma que propondremos medidas de confianza dependiendo también del tipo de habla. En este sistema trabajaremos únicamente al nivel de palabra, no considerando los niveles de concepto y frase.

Sobre este reconocedor aplicaremos las medidas de confianza obtenidas para la recuperación de errores siguiendo los mecanismos descritos en el apartado 5.2.6. Estos

mecanismos se basan en la combinación de hipótesis obtenidas de diferentes reconocedores. En este caso consideraremos tres tipos de reconocedores: uno con modelos acústicos independientes del sexo, otro con modelos acústicos adaptados a la voz masculina y otro con modelos acústicos adaptados a la voz femenina.

5.4.1 Base de datos

La base de datos utilizada para los experimentos realizados sobre este sistema es SpeechDat (Moreno, 1997), la misma que la utilizada para entrenar y evaluar el reconocedor de fechas y horas. En esta base de datos se dispone de 5.000 ficheros de voz, 3.000 con fechas y 2.000 con horas, pronunciados por 1.000 locutores diferentes a través de la red telefónica fija. Como comentamos en el capítulo 4, para realizar los experimentos de tasa de reconocimiento sobre este sistema, seleccionamos aleatoriamente 3.000 ficheros para entrenar los modelos acústicos, 400 ficheros para validación y ajuste de parámetros intermedios, y 1.600 ficheros para evaluar: 800 con habla leída y 800 con habla espontánea.

A la hora de realizar los experimentos sobre medidas de confianza, hemos considerado los conjuntos de evaluación de forma independiente para cada tipo de habla. De esta forma, para habla leída tenemos 800 ficheros con un total de 6.368 palabras, y para habla espontánea tenemos otros 800 ficheros completando un total de 5.104 palabras. Estos ficheros los hemos dividido en tres conjuntos: 66% de los ficheros para el entrenamiento de la Red Neuronal, 17% para su validación y el 17% para evaluación. Esta división se ha repetido 6 veces realizando un proceso Round-Robin, de forma que cada vez, se van utilizando unos datos diferentes para entrenar, validar o evaluar la Red Neuronal, consiguiendo que se usen todos los datos disponibles para evaluar una vez la red. Los resultados presentados en este apartado son la media de los valores obtenidos en todos los experimentos. El intervalo de confianza para el Error de Clasificación, calculado al 95%, es menor del ±1,9% para habla leída y menor del ±2,2% para habla espontánea. Estos valores se han obtenido con la fórmula presentada en el apartado 3.3.1.1, donde la variable **p** es el Error de Clasificación de referencia (Tasas de Error de reconocimiento obtenidas para cada tipo de habla) y **n** es el número de ejemplos de evaluación; 6.368 y 5.104 respectivamente. Los intervalos de confianza para los valores de Rechazo Correcto se presentan en las tablas de resultados.

La etiquetación de cada palabra como acierto o fallo se realiza de forma automática alineando la hipótesis reconocida con la frase de referencia, obteniendo así las palabras correctas, sustituidas, insertadas y borradas. La Tasa de Error de Referencia considerará únicamente los casos de palabras sustituidas e insertadas y dependerá del tipo de habla. La evaluación de las medidas de confianza se realizará según las pautas descritas en el apartado 5.2.4.

5.4.2 Parámetros utilizados

Dado que vamos a trabajar únicamente al nivel de palabra, los parámetros considerados se obtendrán principalmente del proceso de decodificación y del modelo

de lenguaje considerado. En este caso concreto hemos utilizado los mismos parámetros propuestos en el apartado 5.2.2.1. con algunas variaciones que pasamos a comentar.

En cuanto a los parámetros obtenidos del proceso de decodificación, las variaciones han sido las siguientes:

- En relación con el parámetro Homogeneidad de la palabra en la lista de las N mejores hipótesis, en lugar de trabajar con 100 hipótesis trabajaremos únicamente con 10 hipótesis. En este caso, la resolución del parámetro es menor.

- Dado que en este sistema de reconocimiento no tenemos implementada ninguna técnica de Beam Search, el parámetro Perplejidad de fonemas, no lo podremos utilizar.

- Añadiremos dos nuevos parámetros:

  - **Número de nodos coincidentes (NNC):** número de nodos solapados en tiempo con el nodo correspondiente a la palabra analizada, calculados sobre el grafo de palabras obtenido durante la primera etapa de reconocimiento. Este parámetro se normaliza por el número de nodos (del grafo) por palabra. Este valor de normalización se obtiene sumando todos los nodos del grafo y dividiéndolo por el número de palabras de la frase reconocida.

  - **Posición de la palabra en la frase (PPF):** indicación sobre si la palabra es comienzo de frase, final o palabra intermedia. Este parámetro por si sólo no aporta ningún poder de discriminación pero puede ayudar a modelar comportamientos diferentes según la posición. Como veremos más adelante sí ayuda a mejorar los resultados pero su aportación no es significativa.

En cuanto a los parámetros provenientes del modelo de lenguaje utilizaremos los dos parámetros presentados anteriormente: tanto el comportamiento (back-off) como la probabilidad. En este caso, el modelo de lenguaje utilizado ha sido entrenado con una cantidad muy reducida de datos: 2.000 frases para el caso de habla leída y con 1.000 para habla espontánea, frente a las 30.000 frases utilizadas en el caso del sistema CU Communicator. Por esta razón, como veremos más adelante, los resultados obtenidos en este caso son peores a los presentados en el apartado 5.2.5.1.

5.4.3 Combinación de parámetros

Hemos considerado, al igual que en los casos anteriores, una Red Neuronal para combinar los parámetros y obtener una única medida de confianza. La red utilizada ha sido un Perceptrón Multicapa. En este caso, debido a que no tenemos muchos datos de entrenamiento (al igual que para el reconocedor de nombre deletreados) no hemos realizado ninguna codificación de los parámetros para no incrementar en exceso el número de pesos de la red. Estos parámetros los aplicaremos directamente a las entradas de la Red Neuronal, luego es necesario hacer un reescalado para adecuar el rango de cada parámetro al intervalo [0-1]. La capa oculta está formada por 16 neuronas y la capa

de salida por una única neurona. En el entrenamiento de los pesos se etiqueta la salida con un 1 cuando tenemos un acierto de reconocimiento y con un 0 cuando es un error.

5.4.4 Detección de errores en Habla Leída

En este primer apartado veremos los resultados obtenidos en la evaluación de las medidas de confianza para el rechazo de errores de reconocimiento en el caso de Habla Leída. En la tabla 5-15 se presentan las tasas de Rechazo Correcto (ver definición en apartado 5.2.4) para tasas de Rechazo Incorrecto (ver apartado 5.2.4) del 2,5% y del 5,0% para los casos siguientes:

- Considerando únicamente el parámetro Número de Nodos Coincidentes (NNC).
- Para todos los parámetros provenientes del proceso de decodificación (PD).
- Para los dos parámetros del modelo de lenguaje (LM).
- Considerando todos los parámetros unidos sin la Posición de la palabra en la frase (PD+ML) e incluyendo este parámetro (PD+ML+PPF).

No presentaremos los resultados obtenidos al utilizar únicamente la Posición de la Palabra en la Frase porque el poder de discriminación es muy reducido. En esta tabla también presentamos el Error de Clasificación de Referencia y el error mínimo.

Como se puede observar, el parámetro NNC tiene muy poco poder de discriminación, consiguiendo tasas de Rechazo Correcto muy bajas. Por otro lado la posición de la palabra en la frase (PPF) ayuda a mejorar los resultados de clasificación, aunque las diferencias en ningún caso son significativas. Al igual que ocurría en el sistema CU Communicator (tabla 5-2), los mejores parámetros se obtienen del modelo de lenguaje, y los mejores resultados se obtienen combinando todos los parámetros posibles.

Habla Leída (Error de Referencia: 5,9%)						
	Rechazo Correcto (%)		**5,0% RI**			**Mínimo Error de Clasificación**
	2,5% RI	**5,0% RI**	**ECa**	**Ece**	**EC**	
NNC	9,3 (±2,9%)	13,4 (±3,4%)	5,1%	4,7%	9,8%	5,9%
PD	15,7 (±3,7%)	23,4 (±4,3%)	4,5%	4,7%	9,2%	5,9%
ML	28,0 (±4,5%)	33,4 (±4,8%)	3,9%	4,7%	8,6%	5,5%
PD + ML	29,5 (±4,6%)	41,0 (±5%)	3,5%	4,7%	8,2%	5,3%
PD + ML + PPF	**31,3 (±4,7%)**	**42,3 (±5%)**	**3,4%**	**4,7%**	**8,1%**	**5,2%**

Tabla 5-15: Rechazo Correcto de errores para Rechazos Incorrectos de 2,5% y 5% en habla leída. También se muestran los Errores de Clasificación para RI del 5,0%, Mínimo Error de Clasificación y el Error de Referencia.

En este caso, aunque la reducción del Error de Clasificación es de un 11,7% relativa, las diferencias entre las tasas de clasificación son menores que los márgenes de confianza por lo que las diferencias no son significativas estadísticamente. En la gráfica 5-17 se muestra la evolución del Rechazo Correcto según el Rechazo Incorrecto, para los casos de considerar los parámetros del decodificador (PD), del modelo de lenguaje (ML), ambos combinados (PD+ML), y añadiendo la posición de la palabra en la frase (PD+ML+PPF).

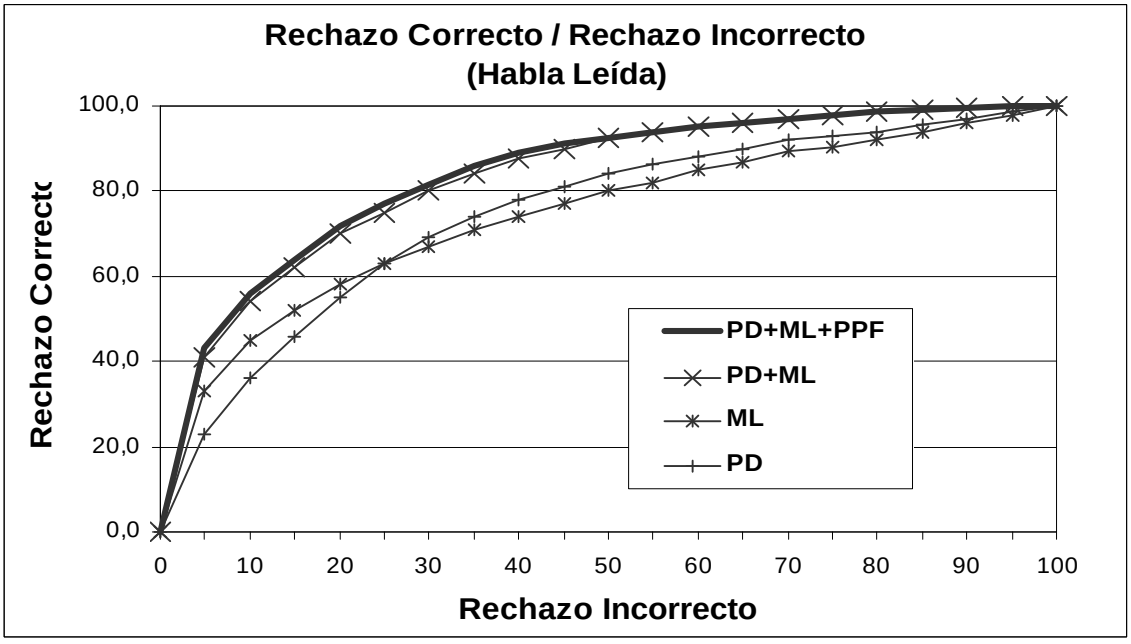

Figura 5-17: Rechazo Correcto (RC) vs. Rechazo Incorrecto (RI) para la detección de errores en el reconocedor de fechas y horas para habla leída.

Al igual que ocurría para el análisis realizado sobre el sistema CU Communicator al nivel de palabra, los parámetros del modelo de lenguaje se comportan mejor que los parámetros del proceso de decodificación para valores de RI pequeños, que es donde se encuentra el punto de trabajo deseado. A medida que aumentamos el Rechazo Incorrecto comienzan a aparecer errores más sutiles que pueden dar lugar a patrones de palabras válidos en el modelo de lenguaje considerado, lo que hace más difícil su detección con estos parámetros. En estos casos, el proceso de decodificación aporta una información más útil para su detección consiguiendo rechazos correctos mayores. Otro aspecto a comentar es que los parámetros ML y PD proporcionan información complementaria de forma que al combinarlos, la gráfica resultante mejora las dos anteriores para todos los valores de RI. Si bien estas diferencias no son significativas como lo eran en el caso del sistema CU Communicator por no disponer de suficientes datos de evaluación.

5.4.5 Detección de errores en Habla Espontánea

Los resultados análogos obtenidos para el caso de habla espontánea se presentan en la tabla 5-16 y en la figura 5-18.

Habla Leída (Error de Referencia: 16,4%)						
	Rechazo Correcto (%)		5,0% RI			Mínimo Error de Clasificación
	2,5% RI	5,0% RI	ECe	Eca	EC	
NNC	9,3 (±2,0%)	13,4 (±2,3%)	14,2%	4,2%	18,4%	16,4%
PD	13,4 (±2,3%)	20,1 (±2,7%)	13,1%	4,2%	17,3%	16,3%
ML	17,0 (±2,5%)	28,5 (±3,1%)	11,7%	4,2%	15,9%	15,8%
PD + ML	20,5 (±2,7%)	33,2 (±3,2%)	10,9%	4,2%	15,1%	14,8%
PD + ML + PPF	**23,8 (±2,9%)**	**34,3 (±3,2%)**	**10,8%**	**4,2%**	**15,0%**	**14,5%**

Tabla 5-16: Rechazo Correcto de errores en habla espontánea para Rechazos Incorrectos del 2,5% y del 5,0% considerando: únicamente el parámetro Número de Nodos Coincidentes (NNC), todos los parámetros provenientes del proceso de decodificación (PD), los dos parámetros del modelo de lenguaje (LM), todos los parámetros sin la Posición de la palabra en la frase (PD+ML) e incluyendo este parámetro (PD+ML+PPF). También se muestran los Errores de Clasificación para RI del 5,0%, Mínimo Error de Clasificación y el Error de Referencia.

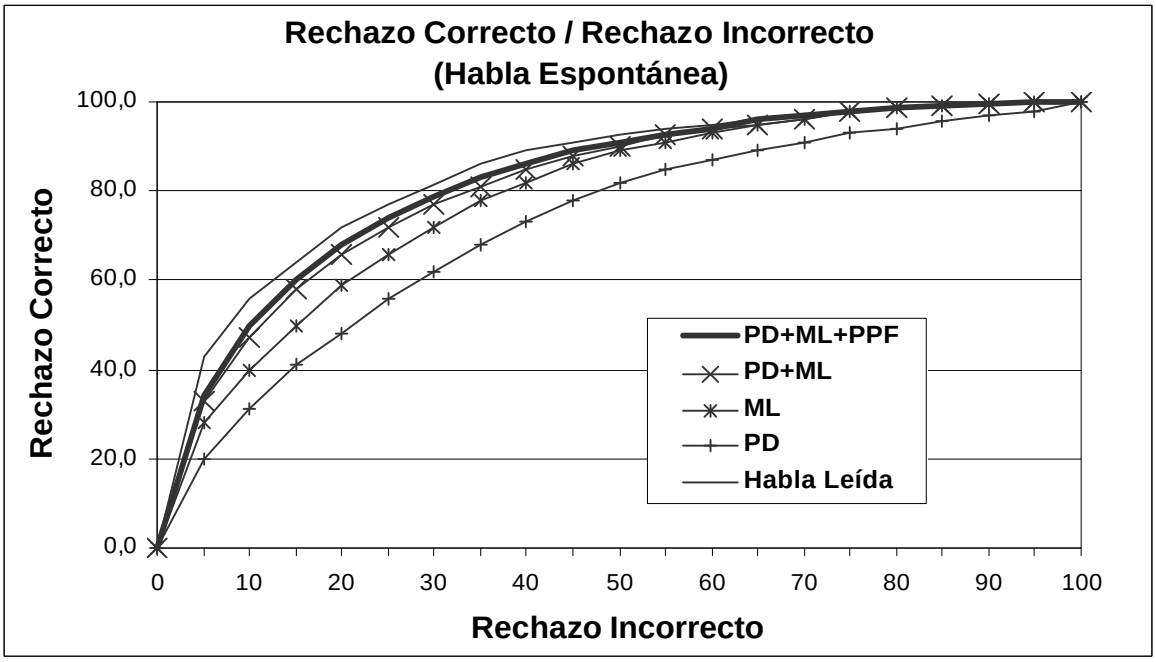

Figura 5-18: Rechazo Correcto (RC) vs. Rechazo Incorrecto (RI) para la detección de errores en el reconocedor de fechas y horas en habla espontánea.

En cuanto a tasas de rechazo, tenemos resultados peores respecto al caso de habla leída (ver figura 5-18), aunque las diferencias no sean significativas por no disponer de suficientes datos de evaluación. En el capítulo 4, apartado 4.1.1 se describe la distribución de ficheros utilizados para entrenar los modelos acústicos utilizados en el reconocimiento tanto para el caso de habla leída como de habla espontánea. Como se

pudo ver, el número de ficheros disponibles de habla leída es el doble que de habla espontánea. Por otro lado, el número de frases disponibles para entrenar el modelo de lenguaje en el caso de habla leída, también ha sido el doble que el utilizado en el caso de habla espontánea. Debido a esto, tanto el modelado acústico como el modelado lingüístico introducido en el reconocedor de fechas y horas, es mejor para el tipo de habla leída, aunque los resultados obtenidos no sean significativamente diferentes por carecer de suficientes datos de evaluación.

Por otro lado el habla espontánea presenta una mayor variación en los efectos acústicos: mayor variación en la velocidad de locución, relajación de pronunciaciones, efectos de coarticulación más pronunciados y construcciones gramaticales más relajadas. Estas variaciones dan lugar a una mayor variedad de patrones de comportamiento que dificultan en mayor medida la tarea de clasificación. A diferencia del habla leída, los parámetros del modelo de lenguaje se comportan mejor que los obtenidos del proceso de decodificación, para todas la tasas de rechazo incorrecto. La utilización de un modelo de lenguaje entrenado con expresiones pronunciadas de forma espontánea, y el desbalanceo del tipo de fichero al entrenar los modelos acústicos hace que en este caso las diferencias sean mayores.

Siguiendo con el razonamiento comentado en los párrafos anteriores, es fácil justificar que los resultados obtenidos para el caso del reconocedor de fechas y horas sean significativamente peores que los obtenidos al nivel de palabra para el sistema CU Communicator (apartado 5.2.5.1) que utiliza el sistema de reconocimiento SPHINX. Una mayor calidad del modelado acústico y lingüístico en el reconocedor, se traduce en un mayor poder de discriminación de los parámetros obtenidos a partir de estas fuentes de información.

5.4.6 Aplicación de medidas de confianza para la recuperación de errores de reconocimiento

Siguiendo las recomendaciones propuestas en el apartado 5.2.6 en las que se proponen las medidas de confianza para la recuperación de errores de reconocimiento, hemos realizado varios experimentos combinando hipótesis de tres reconocedores diferentes: el primero utiliza modelos acústicos independientes del sexo, el segundo utiliza modelos acústicos adaptados a la voz masculina y el tercero utiliza modelos acústicos adaptados a la voz femenina. La obtención de los modelos adaptados a uno u otro sexo se han conseguido dando a cada fichero considerado en el entrenamiento un peso diferente en el proceso de estimación de los parámetros acústicos, según el sexo del locutor.

Para poder aplicar los algoritmos de combinación de hipótesis descritos en 5.2.6, el primer paso ha sido obtener los valores de confianza para cada reconocedor de forma independiente. El proceso es el mismo que el descrito en los apartados anteriores pero variando los modelos acústicos utilizados tanto en el reconocimiento, como en la estimación de las medidas de confianza.

El primer experimento fue analizar la mejora en tasa de reconocimiento conseguida mediante la combinación de las hipótesis obtenidas de los tres reconocedores. Esta combinación se ha realizado haciendo uso de los dos métodos descritos en 5.2.6.1: FLCR (Flat List Confidence Reescoring) y el WGCR (Word Graph Confidence Rescoring). Para el caso del WGCR utilizaremos el alineamiento basado en las marcas temporales para cada una de las palabras. En la tabla 5-17 se presentan los resultados tanto para habla leída como para habla espontánea. También se presentan los resultados obtenidos cuando se considera como heurístico el Incremento de Verosimilitud acumulada por trama a lo largo de la palabra, en lugar de su confianza.

Tasas de Error de Reconocimiento obtenidas combinando hipótesis de los tres reconocedores		
Habla Leída	**FLCR**	**WGCR**
Referencia	9,5%	9,5%
Incremento de Verosimilitud	9,0%	8,8%
Confianza	8,1%	7,8%
Habla Espontánea	**FLCR**	**WGCR**
Referencia	23,5%	23,5%
Incremento de Verosimilitud	22,8%	22,5%
Confianza	20,8%	20,0%

Tabla 5-17: Reducción de la tasa de error de reconocimiento mediante la combinación de hipótesis de diferentes reconocedores. Se presentan los resultados para los dos métodos FLCR y WGCR, considerando como heurístico la Verosimilitud Acumulada y la Confianza. Se detallan los resultados para los dos tipos de habla considerados: leída y espontánea.

Como se puede observar, utilizando las medidas de confianza para combinar varias hipótesis, podemos reducir la tasa de error 1,7 puntos (17,9% relativo) en habla leída, y 3,5 puntos en habla espontánea (14,9% relativo). Utilizando los Incrementos de Verosimilitud para cada palabra, conseguimos también reducir la tasa de error pero en menor medida. Otra vez se pone de manifiesto la mayor reducción conseguida con el método WGCR. En este caso, las diferencias obtenidas no son estadísticamente significativas como obtuvimos para el sistema CU Communicator.

Un detalle importante que conviene matizar es el heurístico concreto utilizado para seleccionar la mejor hipótesis tanto en el FLCR como en el WGCR. Este heurístico depende de la información utilizada: la confianza o el incremento de verosimilitud para cada palabra. Si consideramos la confianza, seleccionaremos la hipótesis que nos dé una mayor *Confianza Media por Palabra*: media de la confianza obtenida para las palabras que forman la hipótesis final. Por otro lado, si utilizamos el incremento de verosimilitud debemos seleccionar la hipótesis que nos ofrezca un mayor *Incremento de Verosimilitud Total*, obtenido sumando los incrementos para todas las palabras de la hipótesis final.

El siguiente experimento realizado ha consistido en la combinación de las N mejores hipótesis de cada reconocedor mediante el algoritmo WGCR, que ofreció mejores resultados en los experimentos anteriores. La evolución del error de reconocimiento con el número de hipótesis consideradas se presenta en la tabla 5-18. Como se puede apreciar, al aumentar el número de hipótesis la Tasa de Error de Reconocimiento también aumenta. Este hecho vuelve a reafirmar las justificaciones argumentadas a tenor de los resultados presentados en las tablas 5-6 y 5-9: por un lado, las medidas de confianza sobre la primera hipótesis no son aplicables a los errores de la hipótesis enésima porque a medida que el número de hipótesis aumente la afirmación de que son salidas posibles del reconocedor es menos cierta. Por otro lado, el modelo de error que supone del cálculo de medidas de confianza, obtenido de un subconjunto tan reducido de parámetros, aporta poco conocimiento para reducir la tasa de error reordenando las hipótesis propuestas por el propio reconocedor.

Tasas de Error de Reconocimiento obtenidas combinando las N mejores hipótesis de los tres reconocedores para habla espontánea	
Número de hipótesis	**Tasa de Error**
1	20,0%
2	20,5%
4	21,0%
8	22,1%

Tabla 5-18: Evolución del error de reconocimiento cuando se combinan las N mejores hipótesis de cada uno de los reconocedores. Evolución para habla espontánea (Martín, 2001).

En el sistema CU Communicator conseguimos reducir la tasa de error cuando combinábamos las dos primeras hipótesis de cada reconocedor (tabla 5-9) pero inmediatamente después el error aumentaba. En este caso no conseguimos mejorar la tasa para ningún valor de N porque el número de reconocedores combinados es distinto: tres en lugar de dos. Cuando en el sistema CU Communicator considerábamos sólo dos reconocedores, al considerar dos hipótesis por reconocedor pasamos a considerar cuatro hipótesis en total, en lugar de dos. Debido a esto, el efecto de aumentar el número de hipótesis fue más importante que el hecho de considerar hipótesis subóptimas. En este caso como partimos de tres reconocedores, al considerar dos hipótesis por reconocedor pasamos a utilizar seis frases en lugar de tres. En este caso, con tres hipótesis ya es un buen punto de partida, difícil de batir, y el efecto de introducir opciones subóptimas es predominante frente a utilizar más hipótesis.

Un detalle que quedó pendiente en el apartado 5.2.6.3 fue el hecho de que las medidas de confianza habían sido entrenadas para predecir errores en la mejor de las hipótesis y no en hipótesis subóptimas, donde la tipología de error es diferente. Estos errores muchas veces obedecen a la necesidad del reconocedor de tener que generar diferentes alternativas y no a problemas de reconocimiento propiamente dichos. En la tabla 5-19, presentamos los resultados obtenidos cuando se entrenan medidas de

confianza independientes para cada una de la hipótesis. A medida que aumentamos el número de hipótesis consideradas, se produce también un aumento del error similar al anterior (las diferencias no son significativas).

Tasas de Error de Reconocimiento obtenidas combinando las N mejores hipótesis de los tres reconocedores para habla espontánea	
Número de hipótesis	**Tasa de Error**
1	20,0%
2	20,8%
4	21,4%
8	23,0%

Tabla 5-19: Evolución del error de reconocimiento cuando se combinan las N mejores hipótesis de cada uno de los reconocedores utilizando medidas de confianza dependientes de la hipótesis considerada. Evolución para habla espontánea (Martín, 2001).

Por último, siguiendo la recomendación comentada en el apartado 5.2.6.2, hemos probado la repercusión sobre la tasa al relajar la condición de alineamiento entre las palabras de las diferentes hipótesis. En lugar de imponer que el comienzo o final de las palabras a unir fuesen exactamente iguales (para definir nodos de unión entre hipótesis), ofrecemos la posibilidad de que haya cierto margen de variación entre estos límites.

Como podemos ver en la tabla 5-20, al aumentar el margen de variación, la tasa de error aumenta. Aunque las diferencias no son estadísticamente significativas, los resultados ponen de manifiesto una tendencia de empeoramiento de los resultados a medida que permitimos mayor flexibilidad en la unión de palabras para generar el grafo.

Tasas de Error de Reconocimiento cuando relajamos el criterio de alineamiento temporal en el método WGCR		
Margen de tramas	**Habla Leída**	**Habla Espontánea**
0	7,8%	20,0%
1	7,8%	20,1%
2	8,0%	20,4%
3	8,2%	20,8%

Tabla 5-20: Evolución del error de reconocimiento cuando relajamos el criterio de alineamiento temporal en el WGCR. Los límites y el margen de variación se definen en número de tramas. En esta tabla se presentan los resultados para habla leída y espontánea (Martín, 2001).

5.5 Conclusiones

En este apartado se resumen las principales conclusiones obtenidas de este capítulo. En cuanto a los experimentos realizados sobre el sistema CU Communicator cabe resaltar los siguientes puntos:

- Considerando un Rechazo Incorrecto (RI) del 5%, hemos conseguido rechazar el 53,2% de palabras erróneas (al nivel de palabra), el 50% de conceptos incorrectos (al nivel de concepto), y el 76,1% de frases no comprendidas por el sistema (al nivel de frase).

- En cuanto al nivel de palabra, los parámetros obtenidos del modelo de lenguaje funcionan mejor para tasas de RI bajas. Combinando los parámetros del proceso de decodificación y del modelo de lenguaje se consiguen resultados bastante mejores que utilizando cada grupo de parámetros de forma independiente, lo que pone de manifiesto la complementariedad de ambas fuentes de información. Estas mismas conclusiones se obtuvieron para el caso del reconocedor de fechas y horas.

- En los niveles de concepto y frase, cabe comentar que las medidas obtenidas al nivel de palabra y de concepto respectivamente, son muy útiles para predecir la confianza en niveles superiores. Al nivel de frase, los parámetros provenientes del analizador semántico tienen una capacidad de discriminación muy importante.

- En el apartado 5.2.6, se proponen las medidas de confianza como heurístico para combinar varias hipótesis de reconocimiento de uno o varios decodificadores. Esta combinación se realiza siguiendo los dos métodos descritos: FLCR y el WGCR. Para el caso del CU Communicator conseguimos reducir la tasa de error un 16% relativo, y en el caso de fechas y horas la reducción es de un 18% para habla leída y un 15% para habla espontánea. La reducción del error se consigue cuando se combinan hipótesis de varios reconocedores, y no cuando combinamos exclusivamente hipótesis del mismo reconocedor.

Sobre el reconocedor de nombres deletreados las conclusiones son las siguientes:

- Los parámetros propuestos permiten detectar, para un RI del 5%, un 58% de errores de reconocimiento y un 68% de nombres deletreados no pertenecientes al diccionario. En este punto conviene resaltar el gran poder de discriminación ofrecido por el parámetro “Diferencia de Verosimilitudes entre Módulos (DVM-3)” para la detección de nombres fuera del diccionario de reconocimiento.

- En los sistemas de reconocimiento formados por varias etapas, a medida que se van utilizando parámetros de confianza de etapas más avanzadas, se consigue un mejor poder de discriminación. En este caso, se utilizan fuentes de información más potentes, y además, las decisiones de reconocimiento tomadas en estos módulos, repercuten en mayor medida sobre la tasa final de reconocimiento.

- La discriminación entre errores y palabras fuera del diccionario de reconocimiento es una tarea muy complicada y los resultados dependen fuertemente del número de casos de ejemplo disponibles para entrenar la Red Neuronal.

Para el caso del reconocedor de nombres de fechas y horas se han realizado experimentos para cada tipo de habla: leída y espontánea. La principal conclusión adicional a la comentada anteriormente es:

- Los resultados obtenidos en este caso son significativamente peores que los obtenidos para el sistema CU Communicator al nivel de palabra, lo que pone de manifiesto que un reconocedor con mejores modelos acústicos y/o lingüísticos permite obtener mejores parámetros para la obtención de medidas de confianza.