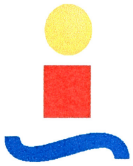

UNIVERSIDAD DE SEVILLA
ESCUELA SUPERIOR DE INGENIEROS
INGENIERÍA DE TELECOMUNICACIÓN

DEPARTAMENTO DE INGENIERÍA ELECTRÓNICA
ÁREA DE TEORÍA DE LA SEÑAL Y COMUNICACIONES

PROYECTO FIN DE CARRERA

MEJORA DE SEÑALES VOCALES VÍA FILTRADO MORFOLÓGICO PARA RECONOCIMIENTO AUTOMÁTICO

Autor: Vicente Javier Rivas Bastante
Tutora: Begoña Acha Piñero
Julio 2004

PROYECTO FIN DE CARRERA:
MEJORA DE SEÑALES VOCALES VÍA FILTRADO MORFOLÓGICO PARA RECONOCIMIENTO AUTOMÁTICO

Autor: Vicente Javier Rivas Bastante
Tutora: Begoña Acha Piñero

El tribunal nombrado para juzgar el Proyecto Fin de Carrera arriba citado, compuesto por:

Presidente:

Vocal:

Vocal secretario:

Acuerda otorgarle la calificación de:

Sevilla, a de de 2004.

Capítulo 1

INTRODUCCIÓN

Las Telecomunicaciones han recorrido un largo camino desde la era de la telegrafía y las primeras conversaciones telefónicas a principios del siglo XX hasta la actualidad dominada por la tecnología digital. En los últimos 20 años hemos sido testigos del desarrollo de sistemas de comunicación digitales tales como la telefonía móvil, las teleconferencias, etc... donde el procesamiento de la señal de voz y especialmente su mejora en entornos adversos juega un papel crucial.

La mejora de la señal voz consiste en procesar una señal corrompida por ruido para tratar de mejorar su calidad o su inteligibilidad, de manera que pueda ser mejor explotada por otros sistemas de procesamiento vocal (compresión, reconocimiento, autentificación...) o mejor reconocida por el oído humano. Este procesamiento es necesario en el sentido de que cada vez que una señal vocal es registrada, algún tipo de ruido aditivo es registrado junto con dicha señal. Por ejemplo, en el entorno de las comunicaciones móviles la señal vocal esta expuesta a todo tipo de ruidos de fondo, que pueden ir desde el ligero zumbido de un aparato de aire acondicionado en una oficina hasta el bullicio producido en una estación de autobuses en hora punta. En ciertas condiciones el nivel de ruido puede incluso exceder el nivel de la señal vocal haciendo caer la relación señal-ruido por debajo de los 0 dB, en cuyo caso la inteligibilidad es baja y la calidad de la señal trasmitida es pobre. Apoyando este hecho podemos ver cómo el nivel de exactitud de un sistema de reconocimiento vocal va decayendo conforme el ruido presente en la señal va aumentando.

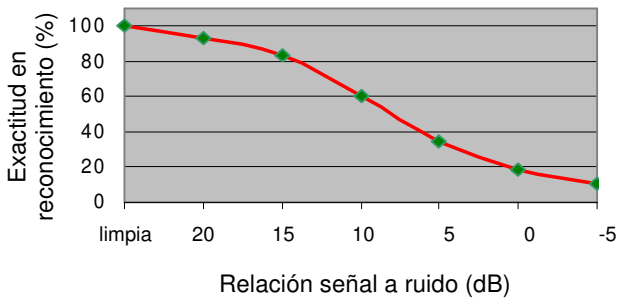

Figura 1.1 Dependencia de la exactitud del reconocimiento con la SNR

Los sistemas de reconocimiento vocal deberían tener un comportamiento robusto frente al ruido y dar buenos resultados en cualquier entorno, condición indispensable para el desarrollo de aplicaciones en al ámbito doméstico o la telefonía móvil por ejemplo. Para alcanzar este objetivo existen diferentes enfoques en la actualidad: intentar mejorar la señal de entrada, adaptar el reconocedor al entorno, etc. Unos de los procedimientos más ventajosos, desde el punto de vista de la independencia del proceso respecto al reconocedor utilizado, es preprocesar las señales para tratar de separar el ruido de fondo de la señal vocal. Esta tarea no es en absoluto sencilla ya que ambas señales comparten la misma banda (lo que elimina la posibilidad de utilizar un filtro paso de banda) y se superponen en el dominio temporal, modificando el valor de cada muestra de la señal vocal y haciendo prácticamente imposible distinguir una señal de otra en condiciones de ruido severas. Se ha investigado mucho en este campo, pero la mayoría de las técnicas utilizadas actualmente no son todavía eficientes cuando se trata de señales en condiciones cercanas a los 0dB

Substracción Espectral

Las técnicas más conocidas y sobre las que más se ha investigado son las llamadas técnicas de Substracción Espectral, cuyo primer referente es el trabajo de Boll [6]. El procedimiento utilizado para compensar el ruido presente en la señal parte de una estimación de la potencia de ruido durante los intervalos donde la señal vocal esta ausente (comienzo de la grabación o pausas entre palabras). Esta estimación es considerada como la potencia de ruido que esta corrompiendo la señal en el periodo inmediatamente posterior y por lo tanto es substraída de la densidad espectral de potencia de la señal original para obtener la señal de voz 'limpia'.

Desafortunadamente, este enfoque requiere una detección fiable de los intervalos de actividad vocal, lo cual se hace impracticable cuando las señales son muy débiles o cuando la relación señal-ruido es muy baja. Además, los periodos que presumiblemente no contienen actividad vocal pueden ser insuficientes para estimar la potencia de ruido de una manera más o menos precisa. Otro problema aún más grave consiste en la asunción de que la potencia de ruido es estacionaria y constante durante los periodos de no actividad vocal y los siguientes intervalos de voz. Esto no es en absoluto correcto ya que las variaciones instantáneas de la potencia de ruido muestran un patrón aleatorio, como puede comprobarse en la figura 1.2. La línea continua

representa los valores reales de la potencia de ruido en cada instante y la discontinua la media que suele utilizarse en la compensación [7]. De todas formas Boll es considerado como el precursor de los trabajos de investigación sobre substracción espectral y cualquier otro estudio posterior se basa en sus ideas.

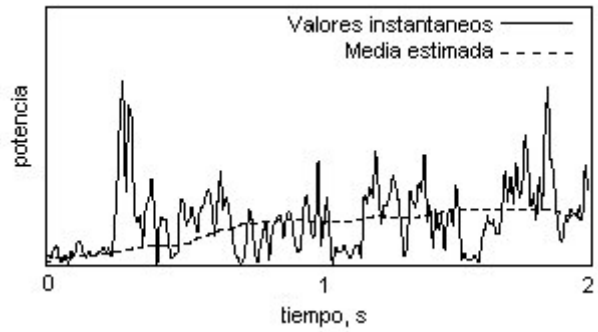

Figura 1.2 Diferencia entre los valores instantáneos de ruido y la media estimada

Para evitar los diversos problemas mencionados anteriormente, especialmente el de la necesidad de disponer de un detector fiable que discriminelos intervalos de actividad y de no actividad vocal, se han desarrollado múltiples nuevos algoritmos basados en la estimación continua de la potencia de ruido, como son los acercamientos de Martín [8], Arslan [10], Stahl [11], Hirsch-Ehrlicher [12] y Evans-Mason [7] entre otros. Con estos acercamientos se consiguen mejoras notables, sobre todo en caso de ruidos no estacionarios; pero siempre tenemos el problema de la diferencia entre los valores instantáneos del ruido y la estimación media.

Filtrado morfológico

Otro enfoque relativamente nuevo es el filtrado morfológico. Esta técnica es utilizada originalmente para procesar imágenes, dando buenos resultados en tareas como la detección de bordes, mejora de estructuras, identificación de objetos, suavizado de formas, segmentación, etc. La aplicación de esta técnica al campo del procesamiento de voz esta todavía en fase experimental, pero recientes trabajos han puesto de manifiesto su validez como vía para mejorar la calidad o inteligibilidad de una señal vocal inmersa en ruido. [1], [2],[3] y [4].

En términos generales, morfología implica el estudio de estructuras o formas. En el contexto del procesamiento de voz la señal vocal y la señal de ruido tienen 'formas' diferentes, lo que motiva la aplicación de esta técnica para intentar separarlas. La imagen sobre la que trabaja el filtrado morfológico en el contexto del procesamiento de voz y en la que se pueden distinguir estas distintas 'formas' es el espectrograma, que se trata de la más común de las representaciones tiempo-frecuencia (las mismas sobre las que se basa la substracción espectral). Aunque no existen muchas referencias en este campo, las que hay son bastante prometedoras. [4] muestra una mejora en media del 10% en exactitud del reconocimiento vocal sobre la señal ruidosa original en el rango que va de los 10dB a los 5dB. Hory en [1] propone un algoritmo basado en las estadísticas del espectrograma con capacidad de discriminar en un espectrograma las zonas que contienen solo ruido WGN y las que contienen además señal determinista. En [3] se lleva a cabo una implementación independiente del algoritmo corroborando los resultados obtenidos por Hory y como extensión se aplica el algoritmo por primera vez a señales vocales como pre-procesamiento para mejorar el comportamiento del reconocimiento automático de voz. Resultados preliminares sugieren que este enfoque (el filtrado morfológico) podría ser robusto incluso bajo condiciones severas de ruido con relaciones señal a ruido por debajo de los 0dB, usando sólo las formas espectrales como información que se manda al reconocedor; es decir, las zonas del espectrograma en las que se considera dominante la señal vocal se ajustan a un nivel de amplitud constante. El comportamiento obtenido es sorprendentemente bueno, llegando a mejoras del 40% sobre la señal sin tratar, lo que anima a seguir investigando en este campo.

Descripción del proyecto fin de carrera

Este proyecto profundiza en el estudio de la aplicación del algoritmo de Hory al campo vocal, optimizando el proceso para llevar a cabo esta tarea. Tomando como punto de partida [1] y [3] comenzamos con un estudio en detalle del algoritmo detectando insuficiencias y formas de mejora, para ir llevando a cabo una serie de ajustes y modificaciones del algoritmo que resultan en una completa optimización del proceso de filtrado morfológico para preprocesar señales vocales en el contexto del reconocimiento automático.

Para tener una visión objetiva de los resultados obtenidos con el filtrado morfológico se lleva a cabo un estudio de las prestaciones de la substracción espectral usando estimación de ruido basada en quantiles, en muestras de voz contaminadas con ruido gaussiano. El estudio comparativo muestra que tras la optimización los resultados son semejantes o ligeramente mejores que los obtenidos con la substracción espectral que es una técnica fuertemente establecida en este contexto. Como extensión y partiendo de los resultados obtenidos con las dos técnicas de forma separada, se realiza un estudio de las posibles formas de combinarlas para aprovechar las cualidades de ambos procesos.

Siendo conscientes de la necesidad de algoritmos con baja carga computacional en el entorno del tratamiento digital de voz por la necesidad de trabajar en muchos casos en tiempo real, se proponen varias modificaciones del algoritmo que reducen considerablemente los requerimientos computacionales sin perder competitividad en los resultados de reconocimiento. Estas modificaciones se concentran en reducir tanto las iteraciones del proceso mediante estimaciones de las condiciones finales del algoritmo, como en eliminar los costosos procesos de propagación de semillas en el algoritmo de región creciente. Los algoritmos propuestos consiguen efectivamente el objetivo deseado, alcanzando porcentajes de reducción del tiempo de ejecución del 95% respecto al algoritmo original y con tan sólo degradaciones del comportamiento en reconocimiento de voz del 3% por lo que los resultados siguen siendo altamente competitivos.