POSICIÓN Y EVOLUCIÓN DE LOS FORMANTES DEL HABLA. ESTADO DEL ARTE

[1] Jesús Bobadilla, [2] Pedro Gómez y [1] Jesús Bernal

[1]
Departamento de Informática Aplicada
Escuela Universitaria de Informática
Ctra. de Valencia Km. 7, 28031 Madrid
Tel: +34.91.3367862, Fax: +34.91.3367527
e-mail: jbobi@eui.upm.es, jbernal@eui.upm.es

[2]
Departamento de Arquitectura y Tecnología de Sistemas Informáticos
Universidad Politécnica de Madrid, Campus de Montegancedo, s/n,
Boadilla del Monte, 28660 Madrid
Tel: +34.91.3367384, Fax: +34.91.3367412
e-mail: pedro@pino.datsi.fi.upm.es

RESUMEN

En este artículo se presenta una revisión de diversas publicaciones centradas en el campo de la fonética acústica y relacionadas con estudios sobre la posición y evolución de los formantes del habla. En total se exponen 52 publicaciones comentadas que pueden servir como referencia para diseñar nuevos experimentos de fonética acústica, evitar repeticiones en las investigaciones y asimilar los conceptos fundamentales relacionados con el área.

ABSTRACT

This article presents a publication review centered in the acoustic phonetics area, more specifically in the speech formants position and evolution studies field. As a whole, 52 commented publications are exposed. They can serve as a reference to design new acoustic phonetics experiments, to avoid research duplications and to learn the most important concepts of the area.

1. INTRODUCCIÓN

En este artículo se pretende dar un breve repaso a diferentes investigaciones desarrolladas en el campo de la fonética acústica, centrándose en aquellas publicaciones que inciden en el estudio de las características acústicas del habla como base para la realización de aplicaciones en el área del tratamiento automático de la voz.

Es importante destacar que a lo largo de los últimos 45 años, el número de publicaciones realizadas en este campo tan extenso es innumerable, y que la selección aquí mostrada incide únicamente en aquellos aspectos que se han considerado de mayor interés para la consecución de objetivos de índole práctica. Para una mayor profundización en el campo, en [TOH92] se recopilan por temas aportaciones en la misma línea que las aquí presentadas.

La selección realizada pretende así mismo asegurar que los lectores tomen como punto de partida los últimos avances logrados en el área, motivando nuevos enfoques y evitando la repetición de investigaciones puntuales.

La extracción de formantes es un objetivo muy importante, debido a la importancia que tiene esta información para conseguir una correcta interpretación de los espectros de voz [PET52], [KAT95]; sin embargo, una vez obtenidos los formantes es necesario conocer sus peculiaridades, que varían apreciablemente según quién sea el hablante, los distintos contextos que presentan los sonidos, la entonación empleada en las frases, etc. [TOK93].

2. DESARROLLO

El estudio con el que comenzaremos será el clásico trabajo de Peterson & Barney [PET52], realizado en los laboratorios de Bell Telephone, que ha marcado e influido tan fuertemente las investigaciones posteriores en el campo que incluso 38 años después se pueden encontrar artículos [WAT90] que no sólo lo referencian, sino que hacen de este estudio el núcleo de la publicación.

En este trabajo se grabaron 10 vocales en un contexto /hVd/ pronunciado por 33 hombres, 28 mujeres y 15 niños. Una vez procesados los datos, se obtuvieron medidas de los formantes F1 a F3 y del tono fundamental (F0). Se descubrió que existe una considerable variabilidad en las frecuencias de los formantes de los distintos hablantes y que se producen solapamientos entre ocurrencias de vocales adyacentes, a pesar de ello las vocales presentaron un alto porcentaje de aciertos en su identificación.

En la producción del habla se producen muchas variaciones debidas a la complejidad del proceso, la intervención de diferentes individuos, los cambios físicos y de estado de ánimo de un mismo hablante a lo largo del tiempo, influencias dialectales, etc.

En [PET52] se diseña un experimento que permite cuantificar y relacionar parámetros espectrales de las vocales inglesas en contextos /hVd/; así mismo, se comprueba cuáles son las vocales mejor reconocidas en pruebas de audición atendiendo a las posiciones de sus formantes.

El resultado más importante obtenido es un mapa de vocales inglesas en el plano Formante 1 / Formante 2; este tipo de mapas ha sido utilizado y mejorado repetidamente hasta nuestros días.

Este estudio tuvo una gran importancia en su época, puesto que sentó las bases de utilización de técnicas de calibración y medida en espectrógrafos, además de mostrar un método que incluye obtención

de datos, aleatorización de procedimientos de medida y audición, uso de técnicas estadísticas, etc.

La base de datos utilizada en [PET52] tiene bastantes limitaciones de diseño, en primer lugar, las medidas fueron tomadas en instantes puntuales de tiempo, por lo cual no existe información de la evolución de los parámetros espectrales. Esta laguna es especialmente importante debido al papel fundamental que ejercen en el reconocimiento de voz las propiedades dinámicas del habla tales como la evolución de los formantes y la duración espectral de cada sonido.

Otras limitaciones son [HIL95]:

1. No se tiene información de los dialectos de los hablantes que realizaron las grabaciones.
2. Los resultados no se agruparon separadamente para hombres, mujeres y niños.
3. No existe información de las edades de los niños empleados en el estudio y además el número de los mismos no era suficientemente grande como para realizar un adecuado estudio estadístico.
4. No se puede determinar el origen (por sexo, edades, etc.) de las muestras empleadas.

En [HIL95] se intenta minimizar estas limitaciones, se grabaron secuencias /hVd/ pronunciadas por un numeroso grupo de hombres, mujeres y niños; se tomaron medidas de la duración de las vocales, frecuencias fundamentales y primeros formantes. Al igual que en [PET52] las señales se presentaron a un conjunto de oyentes para su identificación.

Tras realizar diversos análisis discriminantes se clasificaron las señales usando varias combinaciones de las medidas acústicas. Un resultado fundamental del estudio es la necesidad de utilizar algo más que un simple instante de tiempo por muestra para conseguir una correcta clasificación vocálica.

Los trabajos sobre identificación de formantes forman un grueso bloque de investigación en un área que se considera fundamental para el avance en casi todos los campos del tratamiento automático de la voz, incluidos el reconocimiento y la síntesis del habla. La mayor parte del contenido de este apartado se centra en la recopilación de estudios basados en este concepto

Si bien es conocido que la altura de los formantes y de la frecuencia fundamental varía con la edad y el sexo, (la altura de los formantes producidos por los niños decrece con la edad y en las niñas es mayor que en los niños, aproximadamente un 10% superior), en [BUS95] se realiza un experimento que analiza dichas variaciones. En este estudio se pretende determinar si existen diferencias claras relacionadas con el sexo y la edad en los valores frecuenciales de los formantes producidos por niños y niñas de 5,7,9 y 11 años de edad.

Como resultados principales tenemos que los valores de F0 varían de forma inversamente proporcional al incremento de edad, sin embargo, no se aprecian diferencias significativas de este parámetro atendiendo al sexo. En el caso de la altura de F1, F2 y F3, los valores se hacen menores a medida que aumenta la edad, y las posiciones frecuenciales de las niñas son superiores a las presentadas por los niños.

En [KAT95] se obtuvo F0 a partir de las grabaciones de vocales en contextos /hVd/ producidos por 10 hombres, 10 mujeres y 30 niños con edades entre los 3 y los 7 años. Después se sintetizaron estas mismas vocales con F0 constante (sin variaciones a lo largo del tiempo) y con F0 a una altura frecuencial igual a la media obtenida en los hablantes. Tras un experimento de identificación vocálica, se obtuvo como conclusión que las variaciones de F0 no representan un factor importante en la identificación de vocales sobre sílabas aisladas.

El experimento que se realizó en [ASS95] complementa y amplía al diseñado en [KAT95]; para ello se siguen los mismos pasos, pero trabajando con los 4 primeros formantes (F1-F4). En este caso, la

identificación vocálica empeora significativamente cuando se utilizan formantes sintetizados planos (sin variaciones temporales) y promediados. Estos resultados enfatizan la importancia que en el reconocimiento de la voz tiene la información que proporciona la situación y evolución de los formantes.

Con el fin de situar las vocales castellanas en el plano F1-F2, se realiza en [FER93] un estudio de dispersión de las 5 vocales creadas mediante síntesis de voz, para ello se realiza un experimento de audición, en el que en primer lugar se generan sintéticamente diversas vocales y después se valida su percepción mediante un conjunto de oyentes.

Los resultados del experimento se centran en los valores frecuenciales de las vocales halladas (superiores a los expuestos en diversas publicaciones [MAR94], [ROM88]). Como conclusión se aprecia que el campo de dispersión de las vocales es mayor desde el punto de vista perceptivo que desde el punto de vista de producción.

En [MAR90], Martínez Celdrán realiza una herramienta (programa informático) de representación visual de vectores bidimensionales, lo que viene a mostrar la falta de software de propósito general circulando entre los investigadores hasta el momento.

El estudio de las características espectrales de la voz se complica mucho debido a la variabilidad que se produce en el habla cuando se pronuncian varios sonidos seguidos de forma que existan interacciones entre ellos, lo cual, lejos de resultar extraño, es la situación normal del habla continua.

Una fuente importante de variación en la realización espectral de las vocales está dada por el contexto de consonantes que puede existir. A la influencia de este fenómeno habitualmente se le denomina coarticulación. Se ha desarrollado una gran cantidad de sistemas que modelizan la coarticulación, pero ninguno de ellos parece poder explicar todas las observaciones encontradas [TOK93]. Cada modelo

tiene que resolver el problema de abarcar la variación de las realizaciones de las vocales aisladas respecto a los espectros producidos en segmentos de voz coarticulados.

Gran parte de la dificultad que se presenta viene dada por la variabilidad que introduce la utilización de diversos hablantes, habiéndose observado una marcada dependencia de la persona que pronuncia en los fenómenos de coarticulación, hasta el punto de que éste podría ser considerado como un parámetro de identificación del hablante [SU74]. En este estudio se descubre que la coarticulación en las nasales (especialmente en la 'm') varía significativamente con las distintas personas, y como consecuencia, este factor puede ser tomado en cuenta en el campo del reconocimiento del hablante. Por otra parte, la variación existente entre diversos individuos en el fenómeno de coarticulación, nos hace descartar la hipótesis de universalidad en las características de articulación fonética.

Un reciente estudio en el campo se documenta en [HEU96], donde se realiza un experimento para determinar la variabilidad en el hablante que se produce en la coarticulación de las vocales holandesas 'a','i','u'. Primero se extrajeron los formantes F1 a F3 de muestras de voz coarticuladas con diferentes combinaciones de consonantes, después se efectuaron diversas operaciones estadísticas basadas en análisis lineales discriminantes, y por fin, analizando los resultados, se obtuvo la conclusión de que si bien el efecto de coarticulación ofrece una medida razonable para la correcta identificación de hablantes, no es por si solo lo suficientemente seguro como para poder emplearlo aisladamente en este fin.

En [MAR95] se buscan propiedades acústicas invariantes que el sistema perceptivo humano puede captar en los sonidos del español. Estas propiedades se cuantificarán como parámetros, se asociarán a rasgos y finalmente se relacionarán con fonemas. Este mecanismo serviría de base para la implementación de un sistema de reconocimiento automático del habla basado en rasgos.

Un importante esfuerzo en la búsqueda de propiedades acústicas invariantes en el español ha sido realizado por el grupo de investigación del laboratorio de fonética de la Facultad de Filología de Barcelona. Concretamente, entre otros estudios se han publicado artículos relacionados con la invariación acústica de las oclusivas [RAL95], [VIL95], [GAL95], fricativas [ROM95] y nasales [SAL95].

En [NUÑ95] se estudian espectros de consonantes oclusivas castellanas a partir del comienzo de la explosión y con una duración de 10 á 26 ms. El objetivo perseguido es la determinación de las propiedades acústicas invariantes, buscándose una adaptación a plantillas espectrales con tendencias ascendentes o descendentes. Los resultados obtenidos no permiten una adecuada clasificación de las muestras a las plantillas establecidas.

La invariante acústica de las oclusivas sordas del castellano ha sido estudiada en [RAL95]. En este caso se establecieron una serie de experimentos basados en la capacidad que presentan diversos oyentes para identificar estímulos CV sintetizados como una serie de sílabas de corta duración. Los resultados sugieren que estos estímulos son suficientes para distinguir entre [p], [t] y [k]. Además, se reconoce la importancia de la evolución de los formantes para reconocer las vocales posteriores a las consonantes oclusivas.

[GAL95] presenta un enfoque interesante para situar la posición del locus en las oclusivas sordas como intersección de dos polinomios (de grado 2 y 3 respectivamente). Estos polinomios se generan a partir de las trayectorias del segundo formante en las vocales adyacentes a la consonante oclusiva (empleando secuencias VCV).

En [MEM78] se realiza uno de los estudios primarios destinados a identificar las posiciones de los formantes en vocales con contextos consonánticos. [KEW95] amplía éste y otros trabajos con el objetivo de determinar los efectos de los contextos consonánticos en la discriminación de la frecuencia de los formantes.

Debido a la gran variedad de combinaciones existentes entre vocales y consonantes, la investigación se centró únicamente sobre la /i/ sintetizada aisladamente y en contextos /CVC/ con las consonantes /b,d,g,z,m,l/. La elección de la vocal /i/ no es casual, sino que se trata de uno de los dos casos más complejos (/i/, /ae/) analizados en [MEM78].

Examinando los parámetros de estudio empleados, se destacan dos factores discriminantes: la longitud de la porción estable de la vocal y la separación de F1 y F2 en las transiciones de los formantes. Ambos factores se realzan en vocales con contextos consonánticos. Otra conclusión significativa se centra en la mayor influencia que ejercen las consonantes sobre la evolución del formante F2 respecto a F1.

En [PIC95] se realiza una completísima labor de recopilación de los estudios principales realizados hasta la fecha en el campo de la caracterización fonética de las consonantes intervocálicas. La importancia del tema viene dada no sólo por la frecuencia con la que se presenta este tipo de estructuras dentro de las sílabas, sino porque también es un factor clave su influencia en la percepción del habla cuando estas consonantes se encuentran entre sílabas. Resulta importante obtener un mayor conocimiento en el campo para poder clarificar las diferentes teorías existentes en la percepción de los sonidos.

En el habla continua se producen posiciones intervocálicas de las consonantes con mayor frecuencia que en comienzos o finales de frase. La mayor parte de los estudios de fonética acústica en el área se basan en unidades de tipo sílaba como patrones de partida, esto tiene la ventaja de reducir a un tamaño razonable el número de combinaciones posibles a tratar; por otra parte, el uso de unidades silábicas o similares presenta el problema de que los principios y conclusiones hallados en su estudio no siempre son ciertos y aplicables en el habla continua.

En la percepción de las consonantes intervocálicas intervienen diferentes factores acústicos como los siguientes: variaciones en las frecuencias y anchos de banda de los formantes, segmentos de sonoridad, barras de oclusión, fricación, aspiraciones, silencios, ceros, polos, etc. En algunos casos se requiere la simultaneidad de varios de ellos para la identificación fonética, mientras que otras veces es suficiente que se presente alguno de forma aislada, por ejemplo silencio+barra de oclusión, variación de las frecuencias a lo largo del tiempo, etc.

Diferentes estudios [MAS75], [REP78] muestran la mayor importancia del segmento consonante-vocal (CV) en las ocurrencias VCV analizadas, esto se documenta claramente en [REP78] en donde se aprecia una mayor dificultad para la discriminación VC que en los segmentos CV. Sólo en los casos en los que las pistas proporcionadas en CV no son suficientes, se resuelve la ambigüedad mediante la porción VC.

Las transiciones vocálicas ofrecen una información vital en la localización del punto de articulación consonántico, en [MAR94] se expone que las consonantes se perfilan por las transiciones de las vocales adyacentes y en el caso de las oclusivas, por la altura de la mayor intensidad en la barra de explosión.

Buena parte de los estudios realizados sobre la evolución de los formantes, determinan la posición aproximada de los diferentes locus, sin precisar las diferencias existentes entre las vocales posteriores y anteriores a las consonantes estudiadas. El objetivo de [MOR90] es el de precisar el papel que cumplen las transiciones de las vocales anterior y posterior a una consonante en la identificación del punto de articulación de dicha consonante, para ello, y mediante el uso de un espectrógrafo se generaron palabras de estudio conteniendo consonantes oclusivas sordas y se realizó un experimento auditivo destinado a determinar cuál de las transiciones vocálicas tenía más importancia para predecir el punto de articulación de la oclusiva intervocálica.

Los resultados de [MOR90] muestran con mucha claridad el papel primordial de las transiciones de los formantes de la vocal posterior a la consonante oclusiva estudiada, mientras que las transiciones de los formantes de la vocal anterior no aparecen como representativos para el fin perseguido: la identificación del punto de articulación de las consonantes oclusivas sordas.

En la identificación de consonantes oclusivas, existen varios factores que proporcionan pistas para hallar el lugar de articulación. Las primeras investigaciones se centraron en características individuales de la señal de voz, tales como las barras de oclusión y la transición de formantes [BLU79], [BLU80].

Investigaciones más recientes mantienen la importancia del análisis espectral enfocado en la región de cambio entre el fin de la barra de oclusión y el comienzo de los formantes de la siguiente vocal. Las propiedades espectrales de las barras de oclusión y la evolución completa de los formantes forman pistas secundarias, su función es incrementar la efectividad del reconocimiento y servir de base cuando existen ambigüedades como por ejemplo en ambientes ruidosos.

Las características dinámicas de la voz tales como la evolución de formantes forman la base del aprendizaje. Se piensa que los niños inicialmente orientan su aprendizaje a sílabas completas, reduciendo el tamaño de las unidades a lo largo del tiempo hasta llegar al nivel de pequeños segmentos de voz conteniendo evoluciones cortas, de tal forma que los niños centran el reconocimiento en evoluciones de formantes, mientras que los adultos se basan fundamentalmente en secciones cortas de voz que incluyen parte de la barra de oclusión.

En [OHD95] se examinan las diferencias existentes en el uso de diferentes características acústicas para hallar el lugar de articulación en consonantes oclusivas, centrándose en la importancia de las evoluciones de los formantes en el reconocimiento atendiendo a las edades de los oyentes. Los resultados conseguidos contradicen en parte las ideas expresadas en el párrafo anterior, en primer lugar, la evolución de los formantes no se presenta como una característica

espectral obligatoria en el reconocimiento realizado por niños, por otra parte, la información de corta duración es empleada no solo por los adultos, sino también por los más jóvenes.

Se confirma el emplazamiento de la información fundamental, centrada en un intervalo entre 15 y 25 ms. con 10 ms. empleados en el comienzo de la vocal posterior a la consonante.

La barra de explosión, a pesar de contribuir a que la señal de voz no posea características estacionarias, en las consonantes oclusivas [SMI94], nos ofrece de forma aislada información interesante que puede ser analizada.

Entre los estudios de fonética acústica existentes, algunos se basan en la determinación de la importancia que las barras de explosión tienen en el reconocimiento de las consonantes oclusivas. En [BON96] se investiga sobre la percepción que aportan las barras de explosión y la ayuda que juegan las vocales posteriores, para ello se realizan varios experimentos basados en verificar la habilidad de diversos oyentes para identificar secuencias con la barra de explosión aislada y en otros casos unidas a vocales posteriores.

La principal conclusión que se obtiene en este estudio es que las porciones de voz correspondientes a barras de oclusión aisladas (sin contener ningún segmento vocálico), aportan información muy fiable acerca del lugar de articulación de las consonantes oclusivas. El porcentaje de aciertos en la identificación se sitúa en el 87%, sin embargo, no se puede obtener una perfecta identificación del lugar de articulación sin recurrir al resto de factores (tamaño completo de la barra de explosión y entorno de formantes vocálicos) presentados simultáneamente.

Otros autores [BLU79] con resultados similares, concluyeron que las barras de explosión de forma aislada proporcionan pistas invariantes e independientes del contexto suficientes para la identificación de las consonantes oclusivas, sin embargo, la posición y

evolución de los formantes de la siguiente vocal son necesarios para el perfecto reconocimiento de las secuencias de voz.

En [RAN95] se propone un método de identificación de las consonantes explosivas sordas, este grupo presenta características de señal de naturaleza no estacionaria, con lo que la dificultad para su correcta clasificación aumenta. El método propuesto se centra en la identificación de características de la barra de explosión utilizando análisis espectral básico. Aunque los resultados (porcentajes de acierto) no son brillantes, el método propuesto proporciona una vía para afinar los porcentajes conseguidos.

Fuera de la información aportada por las barras de explosión, el objetivo del trabajo realizado en [MUJ90] es la medición de la duración de las transiciones en la secuencia "oclusiva sorda+vocal" del castellano. Aunque diversos trabajos muestran la importancia de la evolución de los formantes en la determinación del punto de articulación consonántico, ésta investigación se centra únicamente en la duración de las transiciones de la vocal posterior a las consonantes oclusivas sordas del castellano. La concreción con que se ha fijado el objetivo de este trabajo, ayuda a conseguir resultados más fiables que en otras investigaciones de carácter general.

El método utilizado en el estudio es el siguiente:
1.- Grabación de las 15 sílabas que se pueden obtener combinando las consonantes oclusivas sordas con las vocales castellanas colocadas en posición posterior.
2.-Selección del área de transición (cambios bruscos de frecuencia situados junto a la barra de explosión).
3.- Selección de la vocal completa (transición+área estable).
4.- Medición de tiempos.
5.- Experimentos de percepción auditiva de los grupos seleccionados.

Además de los valores concretos de tiempos obtenidos, como aportaciones del trabajo se obtiene la siguiente conclusión: la duración de las transiciones desde el punto de vista perceptivo por sí sola no es

suficiente para identificar el punto de articulación de las consonantes oclusivas sordas.

Un objetivo muy complejo y no tan estudiado como otros en el área del tratamiento automático del habla, se basa en la capacidad para extraer y aislar conversaciones que se producen simultáneamente. En el campo de la percepción de la voz, se debe tener en cuenta la capacidad de los oyentes para comprender el habla en presencia de diferentes sonidos, incluidas diferentes conversaciones simultáneas. Estudios recientes han mostrado las pistas que permiten a los hablantes separar los sonidos de dos conversaciones. Los dos principales factores encontrados son: las diferencias en la frecuencia fundamental (F0) y el patrón de continuidad de formantes, esto es, la tendencia que las resonancias del tracto vocal tienen a variar de forma lenta y continua sus frecuencias a lo largo del tiempo.

En [ASS95b] se realiza un estudio encaminado a determinar cómo los oyentes explotan la característica de continuidad de formantes a lo largo del tiempo, con el fin de separar los sonidos creados por dos hablantes diferentes, para ello, se elaboran varios experimentos en los que se sintetizan sonidos, se presentan a distintos hablantes y se establecen conclusiones con los resultados obtenidos.

En el estudio mencionado, los oyentes fueron capaces de identificar vocales emitidas en paralelo más eficientemente cuando una de ellas iba precedida o seguida por transiciones de formantes forzadas por consonantes líquidas o por glides, aunque solamente con pequeñas mejoras. Una explicación a estos resultados podría venir dada por recientes estudios que apuntan hacia la 'estrategia de cancelación'. De acuerdo con esta teoría, uno de los sonidos que compiten por el reconocimiento es eliminado o 'cancelado' de la mezcla de dos voces mediante la identificación de algún atributo que distinga la voz que interfiere, por ejemplo su estructura de armónicos o timbre. Los resultados de [ASS95b] podrían ser interpretados como un ejemplo del proceso de cancelación.

Un año más tarde, el mismo autor publica nuevos estudios relacionados [ASS96] que amplían las conclusiones obtenidas. En esta ocasión, se presentan dos modelos que tratan de predecir los efectos combinados de la transición de formantes y posición de la frecuencia fundamental con el objetivo de la identificación de vocales pronunciadas simultáneamente.

Los dos modelos predicen con suficiente claridad que las transiciones de los formantes (al comienzo o final de las vocales) generalmente no ayudan a los oyentes a identificar la vocal a la que se asocian estos formantes, por el contrario, se identifican con mayor seguridad las vocales aisladas. Las transiciones de los formantes son beneficiosas simplemente porque proporcionan al oyente una pista que ayuda al reconocimiento del habla en las regiones de transición entre vocales y consonantes.

Las siguientes referencias han sido escogidas para ofrecer una muestra del abanico de posibilidades de investigación que se puede tomar en los aspectos de la fonética acústica menos conocidos. Cabe resaltar la importancia que los estudios clásicos de identificación de formantes, tono de voz, etc. tienen en el desarrollo de todas las áreas del tratamiento automático de la voz.

En [KUS95] se realiza un experimento en el que se hacen grabaciones de 9 personas pertenecientes a 2 familias diferentes, al presentarse estas grabaciones para su reconocimiento se encontró un número significativo de aciertos en los casos en los que hablante y oyente coinciden en una misma familia. Estos resultados como cabría esperar se hacen más patentes a medida que la duración de las grabaciones aumenta hacia el tamaño de la oración, donde el factor de entonación prosódico actúa con una mayor intensidad.

Un sencillo experimento nos podría mostrar la capacidad humana para detectar variedades dialectales, y esta capacidad de reconocimiento se mantendría en unos parámetros razonables a pesar de las diferentes cualidades oratorias de los hablantes. En [TRE95] se diseña un estudio en el que se realizan grabaciones de frases y de

segmentos /hVd/. Los hablantes escogidos pertenecen a dos grupos: uno de raza blanca y el otro de personas de color. Los resultados nos muestran un porcentaje de aciertos muy significativo que mejora en las frases y disminuye en los segmentos escogidos.

[WAN95] ofrece un estudio en el que se pretende detectar emociones atendiendo a diversos parámetros físicos, para ello se escogió a ocho actores que grabaron frases en las que se expresan distintos sentimientos (ira, alegría, estado normal, nerviosismo, odio, etc.), los resultados conseguidos muestran la existencia de dos tipos de emociones: las 'no ambiguas' y las 'ambiguas', estas últimas (p.e. tristeza y depresión) se caracterizan por presentar valores paramétricos parecidos, por el contrario, en las expresiones en las que no existe ambigüedad, se consigue un buen porcentaje de clasificación.

En [COU95] se compara la capacidad para realizar identificación de vocales y discriminación de formantes de varios jóvenes sin problemas auditivos y de ancianos con disminuciones leves de percepción auditiva. Se presentaron 4 vocales a diferentes niveles sonoros (70 y 95 db.). En el caso de los ancianos la media de discriminación se situó en 69% y 80% respectivamente (con una fuerte varianza). Los jóvenes acertaron en una proporción muy cercana al 100%. Los resultados son muy similares para el primer formante, sin embargo, la tasa de aciertos en la discriminación de F2 se acercó significativamente entre los dos grupos, lo que sugiere que la identificación vocálica se consigue en parte gracias a la capacidad de los individuos para discriminar diferencias espectrales en las regiones en las que se sitúa el segundo formante.

Un correcto aprendizaje de la producción del habla requiere de la capacidad de escuchar ejemplos como mecanismo de aprendizaje/corrección [MON83]; para ello, existe la necesidad de llevar a cabo complejas sesiones de corrección de la pronunciación a las personas con deficiencias auditivas; en estos periodos de entrenamiento se utilizan logopedas altamente cualificados, cuyas habilidades sería conveniente simular en sistemas automáticos.

En [CHE95] se estudian las características acústicas que conforman una de las anormalidades más comunes que contribuyen a reducir la comprensión del habla producida por personas sordas: la nasalización en las vocales.

La corrección de la nasalización se complica debido a que es un fenómeno que suele ir unido a otro tipo de problemas en la pronunciación, por otra parte, la corrección humana provoca decisiones subjetivas que pueden confundir a los alumnos, finalmente, no es posible realizar una inspección visual del tracto nasal para corregir estos problemas desde un punto de vista articulatorio.

Con el fin de detectar las diferencias existentes entre hablantes con algún nivel de sordera y aquellos que oyen con normalidad, en primer lugar se realizaron grabaciones (entre miembros de los dos grupos) escogidas por su contenido nasal, después se obtuvieron los espectros de voz y se compararon entre sí. Del estudio de los espectros se deduce la aparición de un pico adicional en las vocales nasales, que en el caso de las personas sin sordera se sitúa alrededor de los 950 Hz. y en el grupo con minusvalía sobre los 930 Hz. Cuanto mayor sea la amplitud de este pico mejor será la nasalización conseguida.

La principal aportación del estudio se basa en la determinación del factor A1-P1 como medida de cuantificación de la nasalización (A1 $\Rightarrow$ amplitud del primer formante, P1 $\Rightarrow$ amplitud del pico extra).

La capacidad de aproximar las características de la señal de voz de un hablante a otro, presenta una serie de aplicaciones entre las que se encuentran: normalización del habla para su posterior reconocimiento, síntesis de voz adaptada a un patrón concreto, avances en identificación del hablante, etc.

En [SLI95] se presenta un trabajo de modificación del hablante basado en la variación de los polos obtenidos con técnicas LPC, desplazando estos polos, se consigue cambiar las posiciones de resonancia del tracto vocal (que determinan los formantes).

Las etapas realizadas son:
1.- Grabación de palabras de referencia obtenidas de varios hablantes.
2.- Segmentación y etiquetado manual de las palabras en porciones fonéticas significativas.
3.- Obtención de polos mediante LPC.
4.- Aproximación de los polos de un hablante hacia los de otro, modificando los ángulos y radios de los ceros mediante transformaciones lineales simples basadas en parámetros estadísticos (medias y varianzas).

Los resultados obtenidos evidencian el mal funcionamiento de este método ante situaciones de coarticulación, por lo que se deduce la necesidad de utilizar funciones más complejas (no lineales) en la etapa de aproximación de los polos de un hablante a otro.

El objetivo fundamental perseguido en [BOO96] es determinar como el reconocimiento de sonidos del habla empeora reduciendo sucesivamente los detalles espectrales. Este estudio pretende ayudar a comprender y cuantificar el impacto que la baja resolución y la introducción de ruido en el habla produce en la percepción de la voz para personas con problemas de audición.

El método seguido se basa en la aplicación de señales ruidosas de distintos anchos de banda sobre porciones de voz. Los límites de la comprensión aparecieron al aplicar filtros de 250 Hz y 8000 Hz. El reconocimiento por parte de los oyentes de palabras aisladas se presenta más susceptible a la introducción de ruido que la comprensión de fonemas aislados.

En [LEE96] se realiza un estudio sobre los efectos producidos por la introducción de una señal ruidosa en el habla y por la incorporación de un filtro que suaviza los picos y los valles espectrales. Una vez realizadas pruebas auditivas con diferentes señales degradadas se constata el hecho de que las personas sin problemas auditivos presentan una mayor tolerancia a este tipo de efectos.

En [ROS96] se investiga en el área del modelado de la voz con y sin presencia de un sistema automático que presente parámetros en tiempo real. En este caso, aunque el aprendizaje se realiza sobre nociones de canto, los resultados se podrían extrapolar a la enseñanza asistida del habla. Las conclusiones del estudio apuntan hacia una influencia directa y positiva en el aprendizaje al utilizarse los medios automáticos de visualización de información.

En [ZAH93] se realiza un estudio muy interesante que nos alerta sobre el peligro de despreciar información espectral importante al restringirnos únicamente a las características habituales como los formantes y el tono de voz.

La búsqueda de características invariantes de los sonidos permanece como uno de los principales y más complicados problemas en el campo del reconocimiento del habla. En el caso de las vocales, desde la publicación del artículo de Peterson y Barney [PET52] se han tomado los tres primeros formantes como la fuente fundamental de información espectral.

Cuando nos enfrentamos al reconocimiento del habla continua, debemos recurrir a un conjunto de información adicional como la evolución de los formantes a lo largo del tiempo [LIN67], cuya detección requiere técnicas más complicadas que un análisis estático de polos [SMI94], [BRO89]. Actualmente se utilizan conjuntamente datos espectrales estáticos y dinámicos (representando evoluciones a lo largo del tiempo) [GOT80], [STR89].

Restringirse a una representación espectral de formantes plantea una serie de consideraciones, tales como la crítica a la reducción de información que se realiza y a la inexacta localización que en algunos casos se produce en la posición de los formantes [BLA82].

En el trabajo realizado por Zahorian y Jagharghi [ZAH93], se realiza un experimento de clasificación automática de vocales, usado para comparar las cualidades del método clásico de extracción de

formantes como característica espectral frente al propuesto en el estudio, que se basa en la utilización de una envolvente suavizada del espectro completo. Los resultados básicos obtenidos apuntan hacia un mejor comportamiento (en este caso mejor clasificación) de la envolvente suavizada del espectro frente a los formantes aislados, sin olvidar que en el método clásico se maneja menos información, con las ventajas que esto conlleva.

En cuanto a las pruebas auditivas realizadas, el método propuesto se presenta más adecuado, lo cual no es sorprendente, puesto que se alberga más información en esta representación espectral que en el caso de tomar solamente formantes aislados. Una característica importante, es la mayor facilidad que existe para determinar trayectorias de picos (cimas) a lo largo del tiempo basándose en la envolvente espectral, frente a los algoritmos más complejos necesarios en el caso de los formantes [SCH95] [PLA95].

Con estas referencias se finaliza el capítulo dedicado al estudio del arte, de cuya lectura se puede deducir que aunque el campo tratado es muy amplio, toda evolución en los conocimientos, métodos y herramientas destinados a la caracterización de los formantes del habla, repercutirá positivamente en los desarrollos que se realicen en casi todas las áreas del tratamiento automático de la voz.

3. CONCLUSIONES

Existe una gran cantidad y variedad de estudios realizados en el campo de la fonética acústica, buena parte de los cuales se basan en el análisis de los primeros formantes de la voz. Gracias a estos estudios conocemos parte de las características espectrales más importantes de los sonidos que intervienen en la comunicación oral.

Un adecuado conocimiento de la posición y evolución de los formantes del habla, facilita considerablemente la comprensión de los espectros de voz.

Los estudios que se realizan sobre fonética acústica, ayudan de una manera muy importante al diseño y realización de aplicaciones de tratamiento de la voz, especialmente en los campos de reconocimiento y síntesis del habla.

La resolución de los diversos objetivos que plantea el tratamiento automático de la voz, obliga a la creación de equipos de investigación de naturaleza multidisciplinar. El campo cubierto por la fonética acústica entra de lleno en las áreas que deben ser cubiertas por estos equipos.

Debido a la variabilidad que existe entre las diferentes realizaciones del habla, resulta imposible establecer unas reglas deterministas de caracterización espectral de sonidos. Sin embargo, conocemos las pautas que definen la posición y evolución de los formantes.

Los fenómenos de coarticulación dificultan el análisis de la evolución de los formantes, aunque también nos muestran parte de los principios básicos sobre los que se asienta la comprensión del lenguaje oral.

El campo de la fonética acústica requiere un enfoque experimental sobre el que se asienten las teorías que los investigadores desarrollan. Por ello, resulta muy importante desarrollar métodos y herramientas que faciliten la creación de aplicaciones que muestren con claridad las distintas características espectrales del habla humana.

4. BIBLIOGRAFÍA

[ASS95a] P.F. Assmann, W.F. Katz, K.M. Jenouri, P.W Hamilton, "Identification of natural and synthetised vowels produced by children and adults: Effects of formant frequency variation", *130th Meeting: Acoustic Society of America, Speech Communication:Studies of Voice*, 4pSC3, 1995

[ASS95b] P.F. Assmann, "The role of formant transitions in the perception of concurrent vowels", *Journal of the Acoustic Society of America*, Vol. 97 (1), Enero 1995, pp. 575-584

[ASS96] P.F. Assmann, "Modeling the perception of concurrent vowels: Role of formant transitions", *Journal of the Acoustic Society of America*, Vol. 100 (2), Agosto 1996, pp. 1141-1152

[BLA82] R.A. Bladon, "Arguments against formants in the auditory representation of speech", *The Representation of Speech in the Peripheral Auditory System*, 1982, pp. 95-102

[BLU79] S.E. Blumstein, K.N. Stevens, "Acoustic invariance in speech production: Evidence from measurements of the spectral characteristics of stop consonants", *Journal of the Acoustic Society of America*, Vol. 66, 1979, pp. 1001-1017

[BLU80] S.E. Blumstein, K.N. Stevens, "Perceptual invariance and onset spectra for stop consonants in different vowel environments", *Journal of the Acoustic Society of America*, Vol. 67 , 1980, pp. 648-662

[BON96] A. Bonneau, L. Djezzar, Y. Laprie, "Perception of the place of articulation of French stop bursts", *Journal of the Acoustic Society of America*, Vol. 100 (1), Julio 1996, pp. 555-564

[BOO96] A. Boothroyd, B. Mulhearn, J. Gong, J. Ostroff, "Effects of spectral smearing on phoneme and word recognition", *Journal of the Acoustic Society of America,* Vol.100 (3), Septiembre 1996, pp. 1807-1818

[BUS95] P.A. Busby, G.L Plant, "Formant frequency values of vowels produced by preadolescent boys and girls", *Journal of the Acoustic Society of America*, Vol. 97(4), 1995, pp. 2603-2607

[COU95] M.P. Coughlin, D. Kewley-Pot, L.E. Humes, "The relation between identification and discrimination of vowels by young normal-hearing and elderly hearing-impaired listeners", *130th Meeting: Acoustic Society of America, Speech Communication:Studies of Voice,* 4pSC1, 1995

[CHE95] M.Y. Chen, "Acoustic parameters of nasalized vowels in hearing-impaired and normal-hearing speakers", *Journal of the Acoustic Society of America*, Vol. 98 (5), Noviembre 1995, pp. 2443-2453

[FER93] A.M. Fernández, "Estudio del campo de dispersión de las vocales castellanas", *Estudios de Fonética Experimental*, Vol. 5, 1993, pp. 129-162

[GAL95] I. Galera, E. Martínez, S. Roig, "Determinación del locus de las oclusivas sordas mediante ecuaciones polinómicas: primera aproximación", Estudios de Fonética Experimental, Vol. 7, 1995, pp. 111-132

[GOT80] T.L. Gottfried, W. Strange, "Identification of coarticulated vowels", *Journal of the Acoustic Society of America*, Vol. 68, 1980, pp. 1626-1635

[HEU96] H.V. Heuvel, B.Cranen, T. Rietveld, "Speaker variability in the coarticulation of /a,i,u/", *Speech Communication*, Vol. 18, 1996, pp. 113-130

[HIL95] J. Hillenbrand, L.A. Getty, M.J. Clark, K. Wheeler, "Acoustic characteristics of American English vowels", *Journal of the Acoustic Society of America*, Vol. 97 (5), Mayo 1995, pp. 3099-3111

[KAT95] W.F. Katz, P.F. Assmann, K.M. Jenouri, "Identification of natural and synthetized vowels produced by children and adults: Effects of fundamental frequency variation", *130th Meeting: Acoustic Society of America, Speech Communication:Studies of Voice*, 4pSC2, 1995

[KEW95] D. Kewley-Port, "Thresholds for formant-frequency discrimination of vowels in consonantal context", *The Journal of the Acoustic Society of America*, Vol. 97 (5), Mayo 1995, pp. 3139-3146

[KUS95] R.E. Kushner, "Analysis and perception of voice similarities among family members", *130th Meeting: Acoustic Society of America, Speech Communication: Studies of Voice*, 3pSC1, 1995

[LEE96] M.R. Leek, V. Summers, "Reduced frequency selectivity and the preservation of spectral contrast in noise", *Journal of the Acoustic Society of America*, Vol. 100 (3), Septiembre 1996, pp. 1796-1806

[LIN67] B. Lindblom, M. Studdert-Kennedy, "On the role of formant transitions in vowel recognition", *Journal of the Acoustic Society of America*, Vol. 42, 1967, pp. 830-843

[MAR90] E. Martínez, "Una utilidad fonética: la carta de formantes por ordenador", *Estudios de Fonética Experimental*, Vol. 4, 1990, pp. 179-193

[MAR94] E. Martínez Celdrán, *Fonética*, Martínez Celdrán E., Teide, 1994

[MAS75] D.W. Massaro, "Preperceptual images, processing time, and perceptual units in speech perception", *Understanding language (Academic Press New York)*, 1975, pp. 125-150

[MEM78] P. Mermelstein, "Difference limens for formant frequencies on steady-state and consonant-bound formants", *Journal of the Acoustic Society of America*, Vol. 63, 1978, pp. 572-580

[MON83] R.B. Monsen, "General effects of deafness on phonation and articulation, Speech of the Hearing Impaired", *(University Park, Baltimore)*, pp. 23-24

[MOR90] MªA. Moreno, "Transiciones vocálicas y punto de articulación consonántico", *Estudios de Fonética Experimental*, Vol. 4, 1990, pp. 50-102

[MUJ90] E. Mújica, Mª M. Santos, J. Herraiz, "Duración de las transiciones en las oclusivas sordas del castellano", *Estudios de Fonética Experimental*, Vol. 4, 1990, pp. 103-122

[NUÑ95] B. Nuñez-Romero, "La invariación acústica en el punto de articulación de las oclusivas del español", Estudios de Fonética Experimental, Vol. 7, 1995, pp. 25-44

[OHD95] R.N. Ohde, K.L. Haley, H.K. Vorperian, "A developmental study of the perception of onset spectra for stop consonants in different vowel environments", *Journal of the Acoustic Society of America*, Vol. 97 (6), Junio 1995, pp. 3800-3812

[PET52] G.E. Peterson, H.L. Barney, "Control methods used in a study of the vowels", *Journal of the Acoustic Society of America*, Vol. 24, 1952, pp. 175-184

[PIC95] J.M. Pickett, H. Bunell, S. Revoile, "Phonetics of intervocalic consonant perception: retrospect and prospect", *Phonetica*, Vol. 52, 1995, pp. 1-40

[PLA95] F. Plante, W.A. Ainsworth, "Formant tracking using reassigned spectrum", *EUROSPEECH 95*, 1995, pp. 741-744

[RAL95] L. Rallo, A.M. Fernández, "La invariación acústica en las oclusivas sordas del castellano. Estudio perceptivo", Estudios de Fonética Experimental, Vol. 7, 1995, pp. 45-74

[RAN95] M. Rangoussi, A. Delopoulos, "Recognition of unvoiced stops from their time-frequency representation", *International Conference on Acoustics, Speech and Signal Processing,* Vol. 1, 1995, pp. 792-795

[REP78] B.H. Repp, "Perceptual integration and differentiation of spectral cues for intervocalic stop consonants", *Perception Psychophysics*, Vol. 24, 1978, pp. 471-485

[ROM88] J. Romero, "Campos de dispersión auditivos de las vocales del castellano. Percepción de las vocales", *Estudios de Fonética Experimental,* Vol. 3, 1988, pp. 86-95

[ROM95] J. Romero, A.Mª Fernández, "La invariación acústica en las fricativas del castellano. Estudio perceptivo", Estudios de Fonética Experimental, Vol. 7, 1995, pp. 133-160

[ROS96] D. Rossiter, D.M. Howard, M. DeCosta, "Voice development under training with and without the influence of real-time visually presented biofeedback", *Journal of the Acoustic Society of America,* Vol. 99 (5), Mayo 1996, pp. 3253-3256

[SAL95] L. Sala, A.Mª Fernández, "La invariación acústica en las nasales del castellano. Estudio perceptivo", Estudios de Fonética Experimental, Vol. 7, 1995, pp. 161-178

[SCH95] P. Schmid., E. Barnard, Robust , "N-best formant tracking", *EUROSPEECH95*, 1995, pp. 737-740

[SLI95] J. Slifka, T.R. Anderson, "Speaker modification with LPC pole analysis", *International Conference on Acoustics, Speech and Signal Processing*, Vol. 1, 1995, pp. 664-667

[SMI94] R. Smits, "Accuracy of quasistationary analysis of highly dynamic speech signals", *Journal of the Acoustic Society of America*, Vol. 96(6), 1994, pp. 3401-3415

[STR89] W. Strange, "Dynamic specification of coarticulated vowels spoken in sentence context", *Journal of the Acoustic Society of America*, Vol. 85, 1989, pp. 2135-2153

[SU74] L.S. Su, K.P. Li, K.S. Fu, "Identification of speakers by use of nasal coarticulation", *Journal of the Acoustic Society of America,* Vol. 56, 1974, pp. 1867-1882

[TOH92] Y. Tohkura, E. Vatikiotis-Bateson, Y. Sagisaka, *Speech perception, production and linguistic structure*, IOS Press, 1992

[TOK93] S. Tokuma, "Some arguments on vowel formant shift", *Speech, Hearing and Language: Work in Progress,* UCL, Vol. 7, 1993, pp. 233-254

[TRE95] S.A. Trent, "Voice quality: Listener identification of African-American versus Caucasian speakers", *130th Meeting: Acoustic Society of America, Speech Communication:Studies of Voice*, 3pSC2, 1995

[VIL95] X. Villalba, "Los invariantes acústicos y el punto de articulación de las oclusivas en español: Una revisión de Lahiri, Gewirth y Blumstein(1984)", Estudios de Fonética Experimental, Vol. 7, 1995, pp. 75-84

[WAN95] R. Wang, W.J. Strong, "Acoustic study of acted emotions in speech", *130th Meeting: Acoustic Society of America, Speech Communication:Studies of Voice*, 3pSC4, 1995

[WAT90] R. L. Watrous, "Current status of Peterson-Barney vowel formant data", *Journal of the Acoustic Society of America*, Vol. 89 (5), Mayo 1991, pp. 2459-2460

[ZAH93] S.A. Zahorian, A.J. Jagharghi, "Spectral-shape features versus formants as acoustic correlates for vowels", *Journal of the Acoustic Society of America*, Vol. 94 (4), 1993, pp. 1966-1982