FUNDAMENTOS DEL RECONOCIMIENTO AUTOMÁTICO DE LA VOZ

"Historia de los sistemas de reconocimiento automático del habla"

Agustín Álvarez Marquina

Introducción (I)

- **Durante las últimas décadas, la investigación en el campo del reconocimiento automático del habla se ha venido desarrollando de una forma intensa, empujada por los avances en procesamiento de señal, algoritmos, arquitecturas y plataformas de cómputo.**

Introducción (II)

- Durante este periodo se han construido sistemas para una amplia gama de aplicaciones, que abarcan desde tareas de reconocimiento de pequeños conjuntos de palabras sobre líneas telefónicas, hasta máquinas de dictado para grandes vocabularios con capacidad para asimilar cualquier tipo de habla [RAB96].

Introducción (III)

- La historia de la investigación en el campo que nos ocupa [RAB93], [JUN96], [JUA98], se ha venido llevando a cabo durante un periodo que abarca la segunda mitad del siglo XX.

- Los primeros intentos por construir máquinas que realizaran tareas de reconocimiento se remontan a la década de los 50, cuando diversos investigadores trataban de explotar los principios fundamentales de la fonética acústica.

Primeros sistemas. Dispositivos electrónicos (I)

- En 1952, en los laboratorios Bell, K. Davis, R. Biddulph y S. Balashek crearon un sistema electrónico que permitía identificar para un solo hablante, pronunciaciones de los 10 dígitos realizadas de forma aislada [DAV52].

- El fundamento de esta máquina se basaba en medidas de las resonancias espectrales del tracto vocal para cada dígito. Las medidas se obtenían mediante el uso de bancos de filtros analógicos.

Primeros sistemas. Dispositivos electrónicos (II)

- En 1959, en la University College de Londres, P. Denes trataba de desarrollar un sistema para reconocer 4 vocales y 9 consonantes [RAB93].

- El aspecto más novedoso de su trabajo era el uso de información estadística, acerca de las secuencias válidas de fonemas en inglés.
  - El objetivo es mejorar el rendimiento de la tarea para palabras que contuvieran dos o más fonemas.
  - Este constituye el primer intento de incorporar conocimiento lingüístico en este tipo de sistemas.

Primeros sistemas. Dispositivos electrónicos (III)

- Hasta este momento, todos los sistemas son dispositivos electrónicos.
- Los primeros experimentos de reconocimiento desarrollados en ordenadores tienen lugar al final de los años 50 y comienzo de los 60, principalmente en el Lincoln Laboratory a cargo de J. Forgie y C. Forgie [JUA98].

Años 60. Sistemas con hardware específico

- Durante la década de los 60, aparecen los primeros desarrollos realizados en Japón aplicando todavía, piezas de hardware específico aplicadas al reconocimiento de:
  - Vocales (J. Suzuki y K. Nakata Radio Research Lab. de Tokio, 1961).
  - Fonemas (T. Sakai y S. Doshita, Universidad de Kioto, 1962)
  - Dígitos (K. Nagata, Y. Kato y S. Chiba laboratorios NEC, 1963) [RAB93].

Años 60. Avances relevantes (I)

- **Sin embargo es durante este periodo cuando se generaliza el uso de computadores en este campo.**
- **En estos años, se inician 3 proyectos que modifican el curso de la investigación y desarrollo en el área del reconocimiento de voz de manera notable.**

Años 60. Avances relevantes (II)

- **El primero de ellos lo realizan T. Martin, A. Nelson y H. Zadell en los RCA Laboratories (1964) [RAB93].**
  - Su objetivo fundamental era desarrollar soluciones realistas para los problemas asociados con la falta de uniformidad de las escalas de tiempo en los hechos de habla.
  - Como consecuencia de este trabajo, los autores diseñaron un conjunto de métodos elementales de normalización en el tiempo, que se basaban en la detección fiable de los puntos de principio y fin de discurso.

Años 60. Avances relevantes (III)

- De esta forma conseguían reducir la variabilidad en las tasas de reconocimiento.
- T. Martin en última instancia completó el método y fundó una de las primeras compañías, *Threshold Technology*, que construía, promocionaba y vendía productos de reconocimiento de voz.

Años 60. Avances relevantes (IV)

- **Al mismo tiempo en la Unión Soviética, T. K. Vintsyuk, propone la utilización de métodos de programación dinámica para conseguir el alineamiento temporal de pares de realizaciones de habla [JUA98].**
  - "Speech discrimination by dinamic programming" Kibernetika 4(2), pp. 81-88 enero-febrero, 1968.
  - Esencia de los conceptos relativos a la técnica de distorsión dinámica temporal o DTW (*Dynamic Time Warping*).

Años 60. Avances relevantes (V)

- Implementación de versiones rudimentarias de algoritmos para realizar reconocimiento de palabras conectadas.
- Su difusión en el resto de Occidente no se produce hasta casi 15 años después.
- Para aquel entonces, otros autores ya han creado e implementado métodos más formales.

Años 60. Avances relevantes (VI)

- **El tercer trabajo lo realiza D. R. Reddy (Stanford University, 1966) [JUA98] en el campo del reconocimiento de habla continua mediante el seguimiento dinámico de fonemas.**
  - La aplicación de sus ideas concluye con el reconocedor de oraciones, dependiente del hablante, para un vocabulario de 561 palabras, que aparece recogida en la tesis doctoral de P. Vicens (1969) [JUN96].

Años 70. Críticas a la viabilidad del reconocimiento de voz

- **Tras 2 décadas de investigación aparecen opiniones muy críticas vertidas contra la utilidad y viabilidad de los métodos de reconocimiento automático de discurso, como las expresados en [PIE69]:**
  - J. R. Pierce, "Whither Speech Recognition?", *Journal of Acoustic Society of America*, Vol. 46, Nº 4, Parte 2, junio 1969, pp. 1049-1050.

Años 70. Principales campos de estudio (I)

- **Sin embargo, los años 70 representan un periodo muy activo para esta disciplina, distinguiéndose dos tipos de actividades principales [JUN96]:**
  1. El reconocimiento de palabras aisladas.
  2. Primeros intentos de construir reconocedores de habla continua y de grandes vocabularios.
     - Basados en el uso de conocimiento de alto nivel, fundamentalmente de tipo sintáctico.

Años 70 . Reconocimiento de palabras aisladas (I)

- **El reconocimiento de palabras aisladas comienza a ser viable y utilizable en la práctica, como consecuencia de los trabajos de:**
  - V. M. Velichko y N. G. Zagoruyko en la Unión Soviética (1970).
  - H. Sakoe y S. Chiba en Japón [SAK78].
  - F. Itakura en los Estados Unidos [ITA70].

Años 70. Reconocimiento de palabras aisladas (II)

- **Los primeros contribuyeron al avance del uso de procedimientos de encaje de patrones en el terreno del tratamiento de la voz.**
- **El grupo japonés estableció de manera formal los algoritmos, que fundamentados en la programación dinámica, podían aplicarse a la resolución de este tipo de problemas.**

Años 70. Reconocimiento de palabras aisladas (III)

- **Por último los trabajos de F. Itakura mostraban cómo los principios de las técnicas LPC (*Linear Predictive Coding*) podían extenderse al reconocimiento.**
  - Empleadas con éxito en la codificación y compresión de la voz.
  - Aplicación mediante el uso de medidas de distancia adecuadas sobre el conjunto de parámetros espectrales LPC.

Años 70. Primeros intentos de reconocimiento de habla continua (I)

- **El fin que se persigue es poder compensar de esta forma los errores cometidos durante la fase de decodificación fonética.**
- **Muchos de estos desarrollos se realizan dentro del marco ARPA Speech Understanding Research (1971-1976) o inspirados por él.**

Años 70. Primeros intentos de reconocimiento de habla continua (II)

- **Los objetivos iniciales cubrían tareas de reconocimiento de oraciones para vocabularios de unas 1000 palabras realizadas por un solo hablante de manera continua.**
- **Es en este momento cuando se advierte que el conocimiento sintáctico, semántico y contextual son fuentes de información.**
  - Permiten reducir el número de posibles alternativas que todo sistema automático de diálogo hombre-máquina debe considerar.

Años 70. Primeros intentos de reconocimiento de habla continua (III)

- **El sistema Hearsay I, construido por la CMU (Carnegie Mellon University) en 1973 era capaz de emplear información de tipo semántico para reducir el número de posibles alternativas que el reconocedor debía evaluar [JUA98].**
  - Consecuencia del impulso investigador de D. R. Reddy, que a finales de los 60 pasó a está universidad.

Años 70. Primeros intentos de reconocimiento de habla continua (IV)

- Ejemplo: tarea *Voice Chess*, consistente en reconocer realizaciones de habla referidas a movimientos de una partida de ajedrez.
  - El número de oraciones alternativas que podían producirse, se limitaba a todos los sinónimos de las posibles jugadas válidas.

Años 70. Primeros intentos de reconocimiento de habla continua (V)

- **Muchas de las aportaciones de estos proyectos vienen más por la parte de la estructura software de los sistemas basados en el conocimiento o K.B.S. (*Knowledge-Based Systems*), que por los avances intrínsecos en el reconocimiento de voz [JUN96].**

Años 70. Primeros intentos con grandes vocabularios (I)

- Otro hito durante esta década es el comienzo de los trabajos del grupo investigador de I.B.M., dedicado al dictado automático por voz para grandes vocabularios [JEL75].

- Finalmente, en los AT&T Bell Labs (ahora Bell Labs, Lucent Technologies y AT&T Labs-Research), los investigadores comenzaron una serie de experimentos orientados a conseguir reconocedores realmente independientes del locutor para su uso en aplicaciones telefónicas [JUA98].

Años 70. Primeros intentos con grandes vocabularios (II)

- Al final de este periodo, la implementación de sistemas reconocimiento de la voz o ASR (Automatic Speech Recognition), se ve favorecida por la disponibilidad de tarjetas microprocesador.
  - Se hace posible la aparición de los primeros reconocedores a un precio bajo.

Años 80. Generalización a sistemas de habla continua

- **Si en la década de los 70 los sistemas de reconocimiento de vocablos aislados alcanzan una cierta madurez, los años 80 se caracterizan por la generalización en la construcción de sistemas de reconocimiento.**
  - Ahora serán capaces de tratar con cadenas de palabras pronunciadas de una manera fluida.

Años 80. Avances importantes (I)

- **La extensión de las técnicas de programación dinámica al reconocimiento de palabras conectadas.**
- **En concreto:**
  - Método en dos niveles de H. Sakoe (1979).
  - Método de pasada única de J. Bridle y M. Brown (1979) [JUN96].

Años 80. Avances importantes (II)

- El giro metodológico que se produce como consecuencia de pasar de métodos basados en comparación de plantillas a los métodos basados en modelado estadístico.
- Debido a la extensión en el uso de los modelos ocultos de Markov o HMM (*Hidden Markov Models*) [RAB86] [RAB89].

Años 80. Avances importantes (III)

- Estos métodos habían sido desarrollados en la pasada década para tratar con problemas de habla continua [BAK75], pero su aceptación generalizada no sucedió hasta unos 10 años después.
- A partir de entonces se han desarrollado numerosas mejoras y actualmente constituyen los mejores modelos disponibles para capturar y modelar la variabilidad presente en el habla.

Años 80. Avances importantes (IV)

- La reintroducción de las redes neuronales [LIP87a], [LIP87b]. Los primeros modelos neuronales como por ejemplo el perceptrón, inicialmente propuesto en los años 50, volvieron a aparecer a finales de esta década gracias al desarrollo de algoritmos de aprendizaje mucho más eficaces.

Años 80. Avances importantes (V)

- La aparición de aproximaciones al problema de la decodificación acústico-fonética para habla continua, fundamentadas en el conocimiento directo de este proceso.

- La tecnología de sistemas expertos había sido postulada como base para diseñar unidades de decodificación fonética que se sirvieran de la experiencia de fonetistas en tareas de interpretación de espectrogramas [CAR87].

Años 80. Avances importantes (VI)

- La grabación de bases de datos de voz como por ejemplo TIMIT (1986), que contribuyen a los avances en la disciplina y que permiten comparar los resultados entre diferentes grupos de trabajo.

Años 80. Avances importantes (VII)

- Durante este mismo periodo, el programa DARPA (Defence Advance Research Agency), impulsó en Estados Unidos el desarrollo de mejores sistemas de reconocimiento para habla continua y vocabularios de tamaño medio y grande con independencia del locutor.

- Muchas de las contribuciones durante este periodo y el principio de los años 90, provienen de los esfuerzos de la CMU a través de su sistema SPHINX [LEE89a], [LEE89b], [LEE89c].

Años 90. Mejora y diversificación de las aplicaciones (I)

- **La década de los 90 supone en cierta manera la continuidad en los objetivos ya propuestos, ampliando eso sí, el tamaño de los vocabularios a la vez que se diversifican los campos de aplicación.**
  - Los servicios sobre la línea telefónica son unos de los que más atención acaparan en la actualidad [CHI97], [GAM97], [JUN97].

Años 90. Mejora y diversificación de las aplicaciones (II)

- **Interés por disponer de sistemas capaces de enfrentarse a situaciones cada vez más reales.**
- **En estos últimos años ha crecido el interés por el estudio de los procesos de reconocimiento en condiciones de ruido y adversas en general [ACE90], [MOR94], [KAS95], [JUN96].**

Resumen. Líneas futuras (I)

- **Algunas de las conclusiones que pueden derivarse de la experiencia acumulada a lo largo de casi 5 décadas son presentadas en [JUN96]:**
  - Los sistemas del presente y presumiblemente los que puedan venir en el futuro se basarán al menos en parte, en modelos y técnicas que aparecieron relativamente pronto en la historia del reconocimiento automático del habla.
  - La transformación de un prototipo de laboratorio de excelentes prestaciones en un sistema comercial fiable es un proceso arduo y no dominado en su totalidad.

Resumen. Líneas futuras (II)

- Las prestaciones del mejor sistema construido en la actualidad, en lo relativo a las tasas de reconocimiento, están por debajo en un orden de magnitud respecto a las que serían atribuibles al ser humano.
- La solución global al problema no se encontrará de manera inmediata por el trabajo de un investigador ingenioso, sino como consecuencia de un trabajo continuado, que incorpore conocimientos multidisciplinares, incluyendo trabajos de investigación básica en los campos de producción y percepción del habla.

Bibliografía (I)

[ACE90] A. Acero and R. M. Stern, "Environmental Robustness in Automatic Speech Recognition", *Proc. of ICASSP'90*, Alburquerque, Estados Unidos, 3-6 abril 1990, pp. 849-852.

[BAK75] J. K. Baker, "The DRAGON System- An Overview", *IEEE Transactions on Acoustic, Speech and Signal Processing*, Vol. ASSP-23, Nº. 1, febrero 1975, pp. 24-29.

[CAR87] N. Carbonell, D. Fohr and J. P. Haton, "APHODEX, An Acoustic-Phonetic Decoding Expert System", *International Journal of Pattern Recognition and Artificial Intelligence*, Vol. 1, Nº. 2, 1987, pp. 31-46.

[CHI97] J. T. Chien and H. C. Wang, "Telephone speech recognition based on Bayesian adaptation of hidden Markov models", *Speech Communication*, Vol. 22, 1997, pp. 369-384.

[DAV52] K. H. Davis, R. Biddulph and S. Balashek, "Automatic Recognition of Spoken Digits", *Journal of Acoustic Society of America*, Vol. 24, Nº 6, noviembre 1952, pp. 637-642.

[GAM97] S. Gamm, R. Haeb-Umbach and D. Langmann, "The development of a command-based speech interface for a telephone answering machine", *Speech Communication*, Vol. 23, 1997, pp. 161-171.

Bibliografía (II)

[ITA70] F. Itakura and S. Saito, "A Statical Method for Estimation of Speech Spectral Density and Formant Frequencies", *Electron. Communication*, Vol. 53, pp. 36-43

[JEL75] F. Jelinek, L. R. Bahl and R. L. Mercer, "Design of a Linguistic Statistical Decoder for the Recognition of the Continuous Speech", *IEEE Transactions on Information Theory*, Vol. IT-21, Nº. 3, mayo 1975, pp. 250- 256.

[JUA98] B. H. Juang, "The Past, Present, and Future of Speech Processing", *IEEE Signal Processing Magazine*, mayo 1998, pp. 24-48.

[JUN96] J. C. Junqua and J. P. Haton, *Robustness in Automatic Speech Recognition*, Kluwer Academic Publishers, 1996.

[JUN97] J. C. Junqua, "SmarTspelT™: A Multipass Recognition System for Name Retrieval over the Telephone", *IEEE Transactions on Speech and Audio Processing*, Vol. 5, Nº 2, marzo 1997, pp. 173-182.

[KAS95] K. Kasper et al., "A Fully Recurrent Neural Network for Recognition of Noisy Telephone Speech", *Proc. of ICASSP'95*, Detroit, Estados Unidos, 9-12 mayo 1995, Detroit, Estados Unidos, 9-12 mayo 1995, pp. 3331-3334.

[LEE89a] K. F. Lee et al., "The SPHINX Speech Recognition System", *Proc. of ICASSP'89*, Glasgow, Reino Unido, 23-26 mayo 1989, pp. 445-448.

Bibliografía (III)

[LEE89b] K. F. Lee, *Automatic Speech Recognition: The Development of the SPHINX System*, Kluwer Academic Publishers, Boston, Massachussetts, 1989.

[LEE89c] K. F. Lee and H. W. Hon, "Speaker-Independent Phone Recognition using Hidden Markov Models", *IEEE Transactions on Acoustic, Speech, and Signal Processing*, Vol. ASSP-37, Nº 11, febrero 1996, pp. 230-239.

[LIP87a] R. P. Lippmann, "An Introduction to Computing with Neural Nets", *IEEE ASSP Magazine*, pp. 4-23, abril 1987.

[LIP87b] R. P. Lippmann, E. A. Martin and D. B. Paul, "Multi-style Trainning for Robust isolated-word speech Recognition", *Proc. of ICASSP'87*, Dallas, 6-9 abril 1987, pp. 705-708.

[MOR94] P. J. Moreno and R. M. Stern, "Sources of Degradation of Speech Recognition in the Telephone Network" *Proc. of ICASSP'94*, Adelaida, Australia, 19-22 abril 1994, Vol. I, pp. 109-112.

[PIE69] J. R. Pierce, "Whither Speech Recognition?", *Journal of Acoustic Society of America*, Vol. 46, Nº 4, Parte 2, junio 1969, pp. 1049-1050.

Bibliografía (IV)

[RAB86] L. R. Rabiner and B. H. Juang, "An Introduction to Hidden Markov Models", *IEEE ASSP Magazine*, Vol. 3, Nº 1, enero 1996, pp. 4-16

[RAB89] L. R. Rabiner, "A Tutorial on Hidden Markov Models and Selected Applications in Speech Recognition", *Proc. of the IEEE*, Vol. 77, Nº 2, febrero 1989, pp. 257-286

[RAB93] L. R. Rabiner and B. H. Juang, *Fundamentals of Speech Recognition*, Prentice-Hall, Englewood Cliffs, N. J., 1993.

[RAB96] L. R. Rabiner, B. H. Juang and C. H. Lee, "An Overview of Automatic Speech Recognition", *Automatic Speech and Speaker Recognition: Advanced Topics*, C. H. Lee, F. K. Soong and K. K. Paliwal editores, Kluwer Academic Publisher, 1996, pp. 1-30.

[SAK78] H. Sakoe and S. Chiba, "Dynamic Programming Algorithm Optimization for Spoken Word Recognition", *IEEE Transactions on Acoustic, Speech, and Signal Processing*, Vol. ASSP-26, Nº 1, febrero 1978, pp. 43-49.