Fusión de sistemas de reconocimiento basados en características de alto y bajo nivel

Mireia Farrús[1], Jan Anguita[1], Javier Hernando[1], Ramon Cerdà[2]
[1] Universitat Politècnica de Catalunya, [2] Universitat de Barcelona

Resumen

Los sistemas automáticos de reconocimiento de locutor más utilizados recientemente están basados en características acústicas de bajo nivel, como las magnitudes espectrales, las frecuencias de los formantes, etc. Sin embargo, estos sistemas no tienen en cuenta algunas características de alto nivel que también pueden contribuir al reconocimiento, como el dialecto, el léxico o la entonación. En esta comunicación se presenta una técnica de fusión de ambos tipos de información con el objetivo de mejorar la tasa de reconocimiento.

1. Introducción

Es bien sabido que, en el área del reconocimiento automático de locutores, los sistemas basados en características acústicas de bajo nivel, tales como el tono, las magnitudes espectrales, las frecuencias de los formantes, los perfiles de energía, etc., son los más utilizados actualmente, al ser los sistemas que logran mejores resultados. Sin embargo, trabajos recientes demuestran que la información de alto nivel puede utilizarse en los sistemas de reconocimiento automático del locutor con el objetivo de mejorar la precisión e incrementar la robustez a las degradaciones del canal y los efectos del ruido, mucho más susceptibles en los sistemas basados en características de bajo nivel (Neuman 1996, Peskin 2003).

Así pues, la utilización de los sistemas de reconocimiento basados en características de alto nivel constituye un paso adelante al aportar información complementaria no analizada en los sistemas tradicionales. Sin embargo, los resultados obtenidos utilizando solamente rasgos de alto nivel son aún muy poco satisfactorios, de manera que deben complementarse necesariamente con los sistemas estadísticos de bajo nivel.

En esta comunicación se presenta una técnica de fusión de los dos tipos de sistemas de reconocimiento que logra mejorar la tasa de reconocimiento. A la vez, se comparan los dos sistemas utilizados y se analizan los errores de ambos con el fin de establecer el grado de complementariedad mutua. La comparación de los sistemas puede permitir el análisis futuro del grado de robustez de uno y otro frente, por ejemplo, a imitaciones llevadas a cabo por impostores.

2. Sistemas basados en información de alto nivel

2.1. Descripción

Las personas tienden a utilizar diferentes niveles de información para reconocer a los hablantes solamente con la voz: el timbre, una carcajada característica o un término léxico utilizado repetidamente pueden ser determinantes en el proceso de reconocimiento (Campbell 2003, Reynolds 2003). La información de alto nivel es la que se utiliza, por ejemplo, al escuchar una conversación a través de una pared o al intentar identificar el personaje que un actor cómico intenta imitar (Kajarekar 2003). Así pues, aunque estas características no proporcionan muy buenos resultados cuando se analizan aisladamente, sí que pueden aportar información complementaria y mejorar los resultados al fusionarse con los sistemas de bajo nivel, además de ser mucho más robustas al ruido y a la degradación del canal (Carey 1996).

Sin embargo, para trabajar con características de alto nivel se necesita una mayor cantidad de material de entrenamiento que en los sistemas de bajo nivel. Esto puede suponer un problema en cualquier aplicación que funcione a tiempo real. Por este motivo, la utilización de características de alto nivel es actualmente mucho más propicia en escenarios concretos, como el ámbito forense, en el cual se dispone de tiempo suficiente para tomar decisiones y recopilar toda la cantidad de datos que sea necesaria (García-Romero 2003).

2.2. Características prosódicas

En esta comunicación se ha utilizado un sistema de reconocimiento de alto nivel basado en la extracción de 14 características diferentes, referentes a la duración de los segmentos, la frecuencia fundamental y las pausas. A continuación se listan detalladamente las características utilizadas:

a) 3 características referentes a la duración de los segmentos:

- log (#frames/palabra) promediado sobre todas las palabras
- duración media de los segmentos sonoros de las palabras
- duración media de los segmentos sordos de las palabras

b) 6 características relacionadas con la frecuencia fundamental:

- log (F0 media) promediado sobre todas las palabras
- log (F0 mínima) promediado sobre todas las palabras
- log (F0 máxima promediado sobre todas las palabras
- log (rango de F0) promediado sobre todas la palabras
- "pseudo pendiente" de F0: (última F0 - primera F0) / (longitud de la palabra)
- pendiente medio (por palabra) sobre todos los segmentos de la estilización lineal (PWL) de la F0

c) 5 características relacionadas con las pausas:

- frecuencia relativa de las pausas "cortas" (70-150 ms)
- frecuencia relativa de las pausas "medias" (151-1000 ms)
- log (duración de las pausas medias) promediado sobre las pausas "medias"
- frecuencia relativa de las pausas "largas" (>1000 ms)
- log (duración de las pausas largas) promediado sobre las pausas "largas"

3. Fusión de los niveles de información

La fusión de dos o más niveles de información puede ser vista desde dos perspectivas diferentes: la *fusión basada en reglas simples* y la *fusión basada en el aprendizaje* (García-Romero 2003, Fiérrez-Aguilar 2003). La fusión basada en reglas simples puede ser tratada como un problema de combinación de clasificadores. De todos los combinadores posibles considerados - reglas de la suma, producto, max, min, mediana, voto mayoritario, etc. - la regla de la suma es el combinador que proporciona mejores resultados.

La fusión basada en el aprendizaje, en cambio, puede ser tratada como un problema de clasificación de patrones si las puntuaciones obtenidas con los clasificadores individuales son consideradas como patrones de entrada que se etiquetarán como 'aceptado' o 'rechazado'. Con esta consideración, cualquier sistema basado en aprendizaje puede ser considerado como una estrategia de fusión. Entre las diferentes técnicas de clasificación de patrones utilizadas para la fusión se encuentran la *logistic regression, maximum a posteriori, k-nearest neighbors classifiers* (Higgins 1993), *multilayer perceptrons, binary decision trees, maximum likelihood, quadratic classifiers* y *linear classifiers* - ordenadas, según Verlinde *et al.* (2000), en sentido decreciente en cuanto a su funcionamiento.

La fusión de dos o más niveles de información puede realizarse en tres etapas diferentes del proceso de reconocimiento: en la extracción de características, en la determinación del nivel de decisión, o bien después de haber obtenido las diferentes puntuaciones *(scores)* para cada nivel de información. Esta última fusión es la preferible, ya que la combinación de *scores* y el acceso a ellos es mucho más fácil que en las otras dos (Jain 2005).

3.1. Normalización de los *scores*

Es muy habitual que los *scores* obtenidos para cada uno de los sistemas individuales no sean homogéneos, es decir, que estén definidos en diferentes escalas numéricas y que sigan diferentes distribuciones estadísticas. Sin embargo, para realizar la fusión de los distintos *scores*, estos deben ser totalmente homogéneos entre sí. La normalización de los *scores* es un paso esencial para transformar los *scores* de los sistemas individuales a un dominio común, antes de combinarse entre sí (Jain 2005).

Los dos tipos de normalización más utilizados son la normalización *min-max* y la normalización *z-score.* La normalización *min-max* es la técnica más simple: los valores mínimos y máximos de los *scores* se desplazan a los valores 0 y 1, respectivamente., y todos los *scores* se transforman en el rango [0,1], de manera que la distribución original se mantiene (excepto para el factor de escala). La normalización *z-score* es la técnica más utilizada. Consiste en transformar la distribución original de los *scores* en una distribución de media cero y variancia unitaria.

3.2. Métodos de fusión basada en reglas

Los métodos de fusión basada en reglas más frecuentes son: la regla de la suma, del máximo, del mínimo y del producto. La regla de la suma es la más habitual, y consiste en sumar los *scores* ponderados de cada uno de los niveles individuales (Ross 2001). Cuando la ponderación es la misma para cada nivel, se habla de suma simple.

En la regla del máximo el *score* final corresponde al *score* más alto de todos los obtenidos por los niveles individuales. Análogamente, en la regla del mínimo el *score* final corresponde al *score* de valor más bajo de todos los obtenidos por los niveles individuales. Finalmente, en la regla del producto el *score* final es el resultado de multiplicar los *scores* de cada uno de los niveles individuales.

4. Evaluación del sistema

4.1. Base de datos

Para la incorporación de información de alto nivel y la posterior fusión con las características de bajo nivel se ha utilizado la base de datos *Switchboard-I* (Campbell 1999, Godfrey 1992). La *Switchboard-I* es uno de los corpus orales más grandes de habla conversacional actualmente disponibles. Es un corpus de 2430 conversaciones telefónicas, de una duración media de seis minutos, realizadas por 543 locutores diferentes (302 hombres y 241 mujeres) de todas las regiones de los Estados Unidos. En total, son más de 240 horas de habla registrada y unos tres millones de palabras de texto, repartidas en 70 temas de conversación diferentes, de los cuales 50 aparecen frecuentemente y de manera que nunca dos locutores conversan juntos más de una vez y ningún locutor habla más de una vez sobre el mismo tema.

4.2. Sistema prosódico de alto nivel

Cada conversación de la *Switchboard-I* contiene dos canales correspondientes a cada uno de los dos hablantes que participan en la conversación. La idea principal es, para cada canal de conversación, obtener un vector de 14 dimensiones constituido por las 14 características prosódicas detalladas en el apartado 2.2.

Para el entrenamiento se han utilizado 8 conversaciones por locutor, siguiendo las directrices de evaluación de la *Extended Data task* de la NIST2001. En dicha evaluación, la base de datos se divide en 6 partes que permiten combinarse para realizar el test de evaluación. Las tres primeras partes se han utilizado para el entrenamiento de los locutores y las tres útlimas para proporcionar un modelo universal de *background.*

Para cada característica individual se calcula la media y la desviación estándar. Para la verificación se utiliza una conversación como locutor de test de la cual se extrae el correspondiente vector de características. Seguidamente, se calcula la distancia entre este vector de características y los *k* vectores correspondientes al locutor que reclama su identidad (*claimed speaker*), utilizando el método k-NN (*k-Nearest Neighbor*) y la divergencia de *Kullback-Leibler* simetrizada para el cálculo de la distancia, que pondera las características en función de su desviación estándar. Esta distancia se compara con la distancia entre el locutor test y el modelo de *background,* con lo que se obtiene el grado de similitud (*score*) entre el locutor que reclama una identidad determinada y su identidad real. Finalmente, a partir de todos los *scores* obtenidos se calcula el *Equal*

Error Rate (EER) del sistema, que da cuenta de su grado de funcionamiento. En la tabla 1 se presentan los EER correspondientes a cada una de las características prosódicas utilizadas, asignando k=3 en el método *k-Nearest Neighbor*. Los resultados son comparables a los obtenidos en Peskin (2003).

característica	EER (%)
log (#frames/palabra)	30.3
duración de los segmentos sonoros	31.5
duración de los segmentos sordos	31.5
log (F0 media)	19.2
log (F0 mínima)	21.3
log (F0 máxima)	21.5
log (rango de F0)	26.6
"pseudo pendiente" de F0	38.3
pendiente medio de F0	28.7
frecuencia relativa de las pausas cortas	46.7
frecuencia relativa de las pausas medias	44.9
log (duración de las pausas medias)	40.0
frecuencia relativa de las pausas largas	42.0
log (duración de las pausas largas)	40.1

Tabla 1. EER de las características prosódicas individuales.

Para fusionar las diferentes características prosódicas se ha utilizado la técnica de la suma simple, combinando los dos tipos de normalización más habituales: la nomalización *z-score* y la normalización *min-max.* Los resultados, en función de las distintas características agrupadas, se presentan en la tabla 2:

conjunto de características	EER (%)	
	z-score	min-max
3 duraciones de segmentos	31.5	31.5
6 características de F0	19.6	17.8
3 duraciones segmentos + 6 características F0	15.6	16.9
5 duraciones de pausa y frecuencias	40.0	36.9
las 14 características	31.0	24.0

Tabla 2. EER de las diferentes combinaciones de características prosódicas, en función del tipo de normalización utilizada.

4.3 Sistema acústico de bajo nivel

El sistema acústico de bajo nivel está formado por modelos de locutores GMM de 32 gausianas entrenados con 8 conversaciones y HMM de silencio de tres estados con una gausiana por estado. El UBM (*Universal Background Model*) está constituido por 116 conversaciones correspondientes a 64 locutores diferentes. Los parámetros extra Frequency Filtering (FF): 20 parámetros estáticos más la primera y la segunda derivadas

temporales. La longitud de la ventana de análisis es de 30 ms con un desplazamiento de 10 ms. Cada locutor se ha entrenado con ocho conversaciones diferentes y se ha obtenido un EER del 10.1 %.

4.4. Fusión

Para la fusión de los dos sistemas se ha utilizado, igual que en la fusión de características prosódicas, la regla de la suma combinada con los dos tipos de normalización. En la regla de la suma se ha realizado un barrido de ponderaciones de los dos niveles de información. En la figura siguiente pueden observarse los EER obtenidos en función de la ponderación dada a cada nivel y para dos conjuntos diferentes de características prosódicas: un primer conjunto que contempla las características de duración de segmentos y de la F0 (excluyendo la información de las pausas), y un segundo conjunto que contempla las 14 características utilizadas.

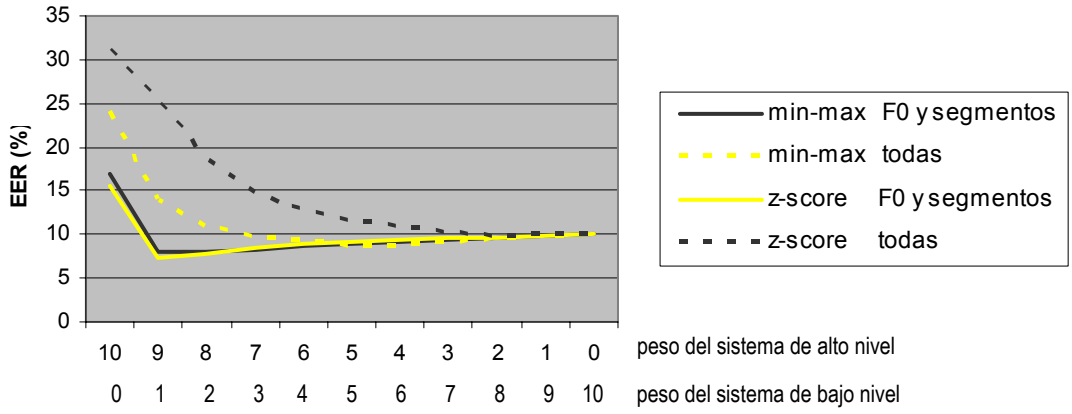

Fig. 1. EER de la fusión de los niveles ponderados

En la tabla 3 se presentan los mejores resultados obtenidos en las diferentes ponderaciones para cada tipo de normalización y diferentes conjuntos de características prosódicas:

conjunto de características	**normalización**	**EER (%)**		
		alto nivel	bajo nivel	fusión
3 duraciones segmentos + 6 características F0	min-max	16.9	10.1	8.0
las 14 características		24.0	10.1	8.7
3 duraciones segmentos + 6 características F0	z-score	15.6	10.1	7.4
las 14 características		31.0	10.1	9.9

Tabla 3. EER de las diferentes combinaciones de características prosódicas fusionadas con el sistema acústico de bajo nivel.

5. Conclusiones

Con la fusión basada en la regla de la suma se ha logrado una mejora del funcionamiento del sistema, tanto en la fusión de las características prosódicas individuales -a excepción de las pausas-, como en la fusión del sistema basado en características de alto nivel con el sistema acústico de bajo nivel. Así pues, los dos sistemas utilizados se complementan mutuamente, eso es, la información contenida en un sistema puramente prosódico no es redundante frente a la obtenida con el sistema acústico de bajo nivel. Sin embargo, los resultados obtenidos no difieren significativamente en función de la normalización utilizada, y los mejores resultados obtenidos con un tipo u otro de normalización varían según la información fusionada.

La base de datos utilizada está basada en conversaciones telefónicas. De ahí que las características relacionadas con las pausas -a diferencia de las otras características- pueda depender no solamente del locutor analizado sino del otro hablante que participa en la conversación, en función de sus comentarios, de la longitud de su intervención o de su capacidad para interrumpir al otro locutor. No es casualidad, pues, que las características individuales referentes a las pausas proporcionen un error tan elevado, e incluso lleguen a empeorar el sistema al fusionarse con el resto de características prosódicas. Es de suponer, pues, que un análisis realizado a partir de monólogos espontáneos en vez de diálogos aportaría más información sobre las características de cada locutor.

BIBLIOGRAFÍA

CAMPBELL, J.P.; REYNOLDS, D.A.; "Corpora for the Evaluation of Speaker Recognition Systems", *Proceedings of the ICASSP*, pp. 829-832, 1999.

CAMPBELL, J.P.; REYNOLDS, D.A.; DUNN, R.B., "Fusing High- and Low-Level Features for Speaker Recognition", *Proceedings of the Eurospeech*, 2003.

CAREY, M.J.; PARRIS, E.S.; LLOYD-THOMAS, H.; BENNETT, S., "Robust Prosodic Features for Speaker Identification", *Proceedings of the ICSLP*, 1996.

FIÉRREZ-AGUILAR, J.; ORTEGA-GARCÍA, J.; GONZÁLEZ-RODRÍGUEZ, J., "Fusion Strategies in Multimodal Biometric Verification", *Proceedings of the IEEE International Conference on Multimedia and Expo,* vol. 3, pp. 5-8, 2003.

GARCÍA-ROMERO, D.; FIÉRREZ-AGUILAR, J.; ORTEGA-GARCÍA, J.; GONZÁLEZ-RODRÍGUEZ, J., "Support Vector Machine Fusion of Idiolectal and Acoustic Speaker Information in Spanish Conversational Speech", *Proceedings of the ICASSP,* vol. 2, pp. 229-232, 2003.

GODFREY, J.J.; HOLLIMAN, E.C.; McDANIEL, J., ”SWITCHBOARD: Telephone Speech Corpus for Research and Development”, *Proceedings of the ICASSP,* vol. 1, pp. 517-520, 1992.

HIGGINS, A.; BHALER, L.; PORTER, J., “Voice Identification using Nearest Neighbor Distance Measure”, *Proceedings of the ICASSP*, pp. 375-378, 1993.

JAIN, A.K.; NANDAKUMAR, K.; ROSS, A., “Score Normalization in Multimodal Biometric Systems”, *Pattern Recognition*, 2005 (to appear).

KAJAREKAR, S.; FERRER, L.; VENKATARAMAN, A.; SÖNMEZ, K.; SHRIBERG, E.; STOLCKE, A.; BRATT, H.; GRADE, R.R., “Speaker Recognition using Prosodic and Lexical Features”, *Proceedings of the IEEE Speech Recognition and Understanding Workshop,* pp. 19-24, 2003.

NEUMAN, M.; GILLICK, L.; ITO, Y.; MacALLASTER, D.; PESKIN, B., “Speaker Verification through Large Vocabulary Continuous Speech Recognition”, *Proceedings of the ICSLP,* pp. 2419-2422, 1996.

PESKIN, B.; NAVRATIL, J.; ABRAMSON, J.; JONES, D.; KLUSACEK, D.; REYNOLDS, D.A.; XIANG, B., “Using Prosodic and Conversational Features for High-Performance Speaker Recognition: Report from JHU WS'02”, *Proceedings of the ICASSP*, 2003.

REYNOLDS, D.A. *et al*., “The SuperSID Project: Exploiting High-Level Information for High-Accuracy Speaker Recognition”, *Proceedings of the ICASSP*, 2003.

ROSS, A..; JAIN, A.K.; QIAN, J.Z.; "Information Fusion in Biometrics", *Proceedings of the 3rd Audio and Video-Based Person Authentication*, pp. 354-359, 2001.

VERLINDE, P.; CHOLLET, G.; ACHEROY, M., "Multi-Modal Identity Verification using Expert Fusion", *Information Fusion*, no. 1, pp. 17-33, 2000.