Procesamiento del Lenguaje Natural, Revista nº 51, septiembre de 2013, pp 161-168 recibido 18-04-2013 revisado 16-06-2013 aceptado 21-06-2013

Nueva técnica de fusión de clasificadores aplicada a la mejora de la segmentación de audio

Applying a new classifier fusion technique to audio segmentation

David Tavarez[1], Eva Navas[1], Daniel Erro[1,2], Ibon Saratxaga[1], Inma Hernaez[1]
[1] AHOLAB, University of the Basque Country (UPV/EHU), Bilbao, Spain
[2]Basque Foundation for Science (IKERBASQUE), Bilbao, Spain
{david, eva, derro, ibon, inma}@aholab.ehu.es

Resumen: Este artículo presenta un nuevo algoritmo de fusión de clasificadores a partir de su matriz de confusión de la que se extraen los valores de precisión (*precision*) y cobertura (*recall*) de cada uno de ellos. Los únicos datos requeridos para poder aplicar este nuevo método de fusión son las clases o etiquetas asignadas por cada uno de los sistemas y las clases de referencia en la parte de desarrollo de la base de datos. Se describe el algoritmo propuesto y se recogen los resultados obtenidos en la combinación de las salidas de dos sistemas participantes en la campaña de evaluación de segmentación de audio Albayzin 2012. Se ha comprobado la robustez del algoritmo, obteniendo una reducción relativa del error de segmentación del 6.28 % utilizando para realizar la fusión el sistema con menor y mayor tasa de error de los presentados a la evaluación.
Palabras clave: Fusión de clasificadores, clasificación y segmentación de audio

Abstract: This paper presents a new classifier fusion algorithm based on the confusion matrixes of the classifiers which are used to extract the corresponding precision and recall values. The only data needed to be able to apply this new fusion method are the classes or labels assigned by each of the classifiers as well as the reference classes in the development part of the database. The proposed algorithm is described and it is applied to the fusion of two audio segmentation systems that took part in Albayzin 2012 evaluation campaign. The robustness of the algorithm has been assessed and a relative improvement of 6.28 % has been achieved when combining the results of the best and worst systems presented to the evaluation.
Keywords: Classifier fusion, audio classification and segmentation

1. Introducción

La segmentación de audio consiste en dividir una grabación en regiones homogéneas de acuerdo a su contenido, asignando a cada segmento la etiqueta de la clase a la que pertenece. En función de la aplicación para la que se realice, el objetivo de la segmentación de audio puede ser muy diferente: separar la voz de la música y el ruido (Lu, Zhang, y Jiang, 2002), separar las voces masculinas de las femeninas (Ore, Slyh, y Hansen, 2006), separar los segmentos que corresponden a distintos locutores (Moattar y Homayounpour, 2012), etc. Tiene muchas aplicaciones y comúnmente se utiliza como primer paso de pre-procesado para mejorar los resultados de otros sistemas como los de reconocimiento automático de habla (Rybach y Gollan, 2009), identificación de locutores (Reynolds y Torres-Carrasquillo, 2005), recuperación de información e indexado de audio basada en su contenido (Meinedo y Neto, 2003) (Aguilo et al., 2009), etc.

Las campañas competitivas de evaluación son una herramienta muy adecuada para determinar de manera objetiva la validez de los algoritmos desarrollados. En estas campañas distintos grupos de investigación prueban sus algoritmos sobre una base de datos común, lo que permite comparar el rendimiento de los mismos e identificar las técnicas más adecuadas para cada etapa del sistema. La Red

ISSN 1135-5948

Temática en Tecnologías del Habla [1] organiza las campañas competitivas de evaluación Albayzin que se celebran cada dos años y evalúan distintos aspectos relacionados con las tecnologías del habla. La segmentación de audio se ha incluido en las dos últimas campañas realizadas, Albayzin 2010 (Butko y Nadeu, 2011) y 2012 [2].

En los problemas de clasificación en los que se comparan diferentes métodos, el que obtiene los mejores resultados suele ser el sistema seleccionado para realizar la clasificación. Sin embargo, en general se observa que los errores cometidos por el resto de los sistemas no son comunes y sus resultados podrían utilizarse para mejorar el rendimiento general del sistema seleccionado, mediante técnicas de fusión de clasificadores (Kittler y Hatef, 1998), (Xu, Krzyzak, y Suen, 1992). De hecho, en diferentes campañas de evaluación con objetivos de clasificación muy distintos, la fusión de varios sistemas funciona mejor que cualquiera de ellos por separado (Schuller, 2012). La fusión de clasificadores puede realizarse a varios niveles (Ruta y Gabrys, 2000):

- *a nivel de datos*: se combinan datos provenientes de diferentes fuentes para realizar la clasificación, como sucede cuando se combinan diferentes rasgos biométricos (voz, huella dactilar, imagen facial,...) en la identificación de personas (Jain y Ross, 2004).
- *a nivel de características*: se combinan distinto tipo de características extraídas a partir de los datos de que se dispone para realizar la clasificación. Un típico ejemplo se produce en los sistemas de verificación de locutor que utilizan información segmental y prosódica extraída a partir de la voz de los locutores (Reynolds et al., 2003).
- *a nivel de decisión*: se combinan directamente los resultados de los clasificadores. Esta combinación puede realizarse bien a nivel de etiqueta, cuando únicamente se dispone como dato de la clase asignada por cada clasificador (Asman y Landman, 2011) o bien a nivel de confianza o *score* cuando se dispone no sólo de la clase a la que pertenece cada segmento sino también de la confianza con la que el clasificador ha tomado la decisión (Ross y Jain, 2003).

En este artículo se propone una nueva técnica de fusión de clasificadores a nivel de etiqueta que se ha aplicado con éxito a la fusión de los resultados de dos sistemas de segmentación de audio participantes en la campaña de Albayzin 2012.

La sección 2 del artículo presenta la técnica de fusión de clasificadores propuesta. En la sección 3 se describen resumidamente los datos más relevantes de la campaña Albayzin 2012 de evaluación de sistemas de segmentación de audio. Los resultados obtenidos aplicando la técnica propuesta a dos sistemas participantes en dicha evaluación se presentan y analizan en la sección 4 y finalmente en la sección 5 se exponen las conclusiones del trabajo.

2. Técnica de fusión propuesta

En esta sección se describe el algoritmo propuesto para realizar la fusión de los resultados de dos clasificadores diferentes, en base a su matriz de confusión. Es un algoritmo de fusión que funciona a nivel de etiqueta en el que la única información requerida para poder ser aplicado son las clases asignadas por cada uno de los sistemas y las clases de referencia en la parte de desarrollo de la base de datos

Dados dos clasificadores, c_1 y c_2, en un escenario multiclase, el algoritmo propuesto trata de evaluar la confiabilidad de las decisiones de cada uno de los clasificadores, basándose en sus valores de precisión (*precision*) y cobertura (*recall*) para las clases emitidas. Supongamos que para un caso concreto tenemos como salidas las clases a y b para los clasificadores c_1 y c_2 respectivamente. El algoritmo propuesto plantea que para estimar la confiabilidad de la decisión tomada por el clasificador c_1 hay que tener en cuenta no sólo su precisión para la clase a, sino también la probabilidad de que el clasificador c_2 confunda la clase a con la clase b. Es decir, siendo c_1 y c_2 independientes, la probabilidad de que la respuesta real sea a equivale al producto de la probabilidad de que c_1 haya acertado al elegir a por la probabilidad de que c_2 se haya equivocado al elegir b. Estimaremos esta probabilidad por medio de la tasa de falsos negativos para este segundo clasificador, cal-

[1] http://www.rthabla.es/

[2] http://iberspeech2012.ii.uam.es/index.php/call-for-evalproposals-2/audio-segmentation

culada en la parte de desarrollo de la base de datos.

Con el fin de evaluar dicha confiabilidad de la decisión, en primer lugar es necesario obtener las matrices de confusión de los dos clasificadores. Para ello se utilizan los resultados que los clasificadores logran en la parte de desarrollo de la base de datos. Podemos ver un ejemplo de dicha matriz para un caso en que se consideran tres clases diferentes en la tabla 1, en la que se representan de arriba a abajo las clases reales, a, b y c, y de izquierda a derecha las predicciones realizadas por un clasificador, a', b' y c'. De forma general, V_x representa el número de positivos verdaderos o aciertos del clasificador para la clase x, $F_{xx'}$ el número de falsos negativos o errores cometidos por el clasificador al no identificar la clase real x y predecir en su lugar x', T_x el número total de ejemplos de la clase x en la parte de desarrollo de la base de datos y $T_{x'}$ el número total de ejemplos de la base de datos marcados como clase x' por el clasificador correspondiente.

Real / Pred.	a	b	c	Total
a'	V_a	$F_{ba'}$	$F_{ca'}$	$T_{a'}$
b'	$F_{ab'}$	V_b	$F_{cb'}$	$T_{b'}$
c'	$F_{ac'}$	$F_{bc'}$	V_c	$T_{c'}$
Total	T_a	T_b	T_c	-

Tabla 1: Ejemplo de matriz de confusión para un escenario con tres clases

Para evaluar la confiabilidad de la decisión de cada clasificador tendremos en cuenta además de la precisión del mismo para la clase propuesta, el error cometido por el otro clasificador al escoger una salida diferente, por lo que utilizaremos las matrices de confusión de los dos clasificadores implicados en la fusión a la hora de analizar la confiabilidad de la decisión de cada uno.

Por un lado se considerará la precisión del clasificador para cada clase, que de acuerdo con la nomenclatura anterior es:

$$TP_x = \frac{V_x}{T_{x'}} \tag{1}$$

donde x representa la clase propuesta por el clasificador, con V_x y $T_{x'}$ definidos anteriormente. Mediante este valor se estima la probabilidad de que el segmento clasificado pertenezca realmente a la clase a la que ha sido asignado por el clasificador.

Adicionalmente, podemos analizar el error cometido por cada clasificador por medio de la tasa de falsos negativos, obtenida siguiendo el ejemplo anterior como:

$$FN_{xx'} = \frac{F_{xx'}}{T_x} \tag{2}$$

donde x representa la clase real (la de referencia) y x' la clase propuesta por el clasificador, con $F_{xx'}$ y T_x definidos anteriormente. Este valor se utiliza como estimación de la probabilidad de que el clasificador haya confundido la clase real x con la x'.

Si en un caso concreto las salidas de los sistemas originales coinciden, la clase propuesta por ambos es asignada directamente. En caso de obtener salidas diferentes para los sistemas originales, procedemos a evaluar la confiabilidad de la decisión de cada uno con el fin de seleccionar una de las dos clases propuestas.

Aplicando un razonamiento probabilístico, para calcular la confiabilidad de la decisión tomada por el primer sistema multiplicamos su precisión para la clase propuesta y la tasa de falsos negativos del segundo clasificador en función de la salida de ambos. Siguiendo el ejemplo anterior tenemos:

$$r_{xy'}\big|_1 = TP_x|_1 \cdot FN_{xy'}\big|_2 \tag{3}$$

donde x representa la clase propuesta por el primer clasificador e y' la clase propuesta por el segundo clasificador. En este caso suponemos que la clase propuesta por el primer clasificador, x, es correcta, por lo que la confiabilidad de la decisión del primer clasificador dependerá de su precisión para esta clase, $TP_x|_1$, y del supuesto error cometido por el segundo clasificador al escoger la clase y, es decir, la tasa de falsos negativos del segundo clasificador para la clase y cuando la clase real es x, $FN_{xy'}\big|_2$. Se considera que las decisiones de los dos clasificadores son independientes y por ello se multiplican las probabilidades para estimar la probabilidad conjunta que representa la confiabilidad de la decisión.

Una vez obtenida la confiabilidad de la decisión del primer clasificador, procedemos del mismo modo para evaluar la confiabilidad de la decisión del segundo clasificador. En este caso multiplicamos su precisión para la clase que ha seleccionado y la tasa de falsos negativos del otro clasificador en función de la

salida de ambos. Siguiendo el ejemplo anterior tenemos:

$$r_{yx'}|_2 = TP_y|_2 \cdot FN_{yx'}|_1 \quad (4)$$

donde ahora y representa la clase propuesta por el segundo clasificador, que se supone correcta, y x' la clase propuesta por el primer clasificador.

Una vez evaluada la confiabilidad de la decisión de cada clasificador, se asigna en cada caso la clase propuesta por el clasificador cuya confiabilidad obtenida resulta mayor.

3. Campaña Albayzin 2012

La campaña de evaluación de sistemas de segmentación de audio Albayzin 2012 consistió en la segmentación de audio *broadcast*, asignando a los segmentos obtenidos etiquetas para indicar la presencia de voz, música y ruido en cada uno de ellos, pudiendo existir solapamiento entre las tres clases en cualquier instante.

3.1. Base de datos

Los organizadores de la campaña proporcionaron dos bases de datos de audio diferentes correspondientes a programas de noticias para ser utilizadas en el desarrollo de los sistemas de segmentación.

La primera, utilizada también en la campaña Albayzin 2010 de evaluación de sistemas de segmentación y diarización, está formada por unas 87 horas de grabaciones de programas emitidos por el canal catalán de televisión 3/24. Estos datos podían ser utilizados para realizar el entrenamiento de los sistemas. La distribución de las clases de audio contenidas en esta base de datos es la siguiente: 37 % de voz limpia, 5 % de música, 15 % de voz con música de fondo, 40 % de voz con ruido de fondo y 3 % de otros. En esta última clase se engloba todo el material que no pertenece a las cuatro clases anteriores, incluyendo el ruido.

La segunda base de datos proporcionada por la organización proviene de la Corporación Aragonesa de Radio y Televisión (CARTV), que donó parte de su archivo de Aragón Radio. Está formada por unas 20 horas de audio con la distribución de clases que se describe a continuación: 22 % de voz limpia, 9 % de música, 31 % de voz con música de fondo, 26 % de voz con ruido de fondo y 12 % de otros. En este caso la clase 'otros' contiene tanto los silencios como el ruido y las combinaciones de clases que no se han mencionado. Aproximadamente 4 horas podían ser utilizadas para el entrenamiento de los sistemas y las 16 horas restantes fueron empleadas por la organización para su evaluación.

3.2. Métrica utilizada

Al igual que en las evaluaciones organizadas por el NIST (*National Institute of Standards and Technology*), la métrica utilizada para evaluar el funcionamiento de los sistemas ha sido el SER (Tasa de Error de Segmentación o *Segmentation Error Rate*), que se corresponde con la fracción de tiempo que no ha sido correctamente atribuida a la clase correspondiente (voz, música y ruido en este caso). En las zonas de solapamiento entre clases la duración del segmento se atribuye a todas las clases presentes en el mismo, por lo que un mismo segmento temporal puede ser considerado más de una vez en los cálculos.

El SER se calcula como la suma de tres tipos de errores: el porcentaje de tiempo que es asignado a una clase incorrecta (Error de Clase o *Class Error Time*), el porcentaje de tiempo en el que una clase está presente pero no ha sido etiquetada (Error de Omisión o *Missed Class Time*) y el porcentaje de tiempo en que se ha etiquetado una clase cuando realmente no estaba presente (Error de Inserción o *False Alarm Time*). Todos estos errores se han calculado mediante las herramientas de evaluación proporcionadas por el NIST (NIST, 2009).

3.3. Resultados de la evaluación

En la campaña de evaluación Albayzin 2012 tomaron parte 6 sistemas desarrollados por 5 grupos de investigación diferentes. La tabla 2 recoge los resultados obtenidos por los distintos sistemas en la parte de evaluacion de la base de datos.

Por respeto a los otros participantes y dado que la técnica de fusión no depende de sistemas concretos sino sólo de las etiquetas que proporcionan, nos referiremos a los sistemas ajenos con los nombres ficticios S2...S6 en atención al puesto que ocuparon en la evaluación.

3.4. Sistemas seleccionados para la fusión

Para comprobar los resultados de la técnica de fusión propuesta se seleccionaron los sistemas que mejor y peor resultados obtuvieron

Sistema	SER (Test)
AHOLAB-EHU	25.78 %
S2	26.53 %
S3	28.12 %
S4	33.30 %
S5	39.55 %
S6	40.01 %

Tabla 2: Resultados de los sistemas de segmentación de audio presentados en la campaña de Albayzin 2012

en la evaluación Albayzin 2012, suponiendo que si de este modo se consiguen mejoras en los resultados, realizando la fusión entre sistemas con menor tasa de error la mejora sería mayor. Por lo tanto se utilizarán en los experimentos de fusión el sistema presentado por Aholab (Tavarez et al., 2012) y el sistema S6. Además, con el fin de comprobar los resultados cuando los dos clasificadores originales presentan bajas tasas de error, se seleccionaron dos de los sistemas que mejores resultados obtuvieron en la evaluación, Aholab y el sistema S3.

4. Experimentos y resultados

En esta sección se muestran los resultados obtenidos al aplicar el algoritmo de fusión descrito anteriormente a los sistemas de segmentación de audio seleccionados.

En primer lugar, se ha realizado un mapeo de las clases para evitar el solapamiento de las mismas. Como se ha comentado en la sección 3, en el problema de segmentación planteado en Albayzin 2012 las clases podían solaparse. Sin embargo, para llevar a cabo la fusión de los sistemas con el método propuesto, la salida de cada sistema debe ser única en cada segmento, por lo que las clases originales (voz, música y ruido), han sido sustituidas por las diferentes combinaciones posibles entre ellas: voz limpia, música, ruido, voz con música, voz con ruido, etc.

A continuación se calcula la matriz de confusión de cada sistema usando la parte de desarrollo de la base de datos de Aragón Radio. Posteriormente se realiza la combinación de las salidas de los sistemas originales mediante la técnica de fusión propuesta y se obtienen las marcas finales para cada segmento deshaciendo el mapeo realizado inicialmente. La tabla 3 muestra el resultado, en términos de SER, de la fusión de los dos sistemas seleccionados tanto en la parte de desarrollo y como en la de evaluación de la base de datos.

Sistema	Desarrollo	Evaluación
AHOLAB	19,97 %	25,78 %
S6	35,93 %	40,01 %
Fusión	**18,71 %**	**24,16 %**

Tabla 3: Resultado de la fusión de los sistemas Aholab y S6

Podemos observar cómo se logra una mejora de los resultados obtenidos en ambos casos, con una reducción relativa del SER del 6.3 % en la parte de desarrollo y del 6.28 % en la parte de evaluación respecto al mejor de los sistemas. Se trata de una mejora significativa teniendo en cuenta que uno de los sistemas utilizados parte con un SER del 40 %, el mayor de la campaña de evaluación. El método de fusión propuesto es capaz de obtener información suficiente de éste sistema y de utilizarla para mejorar los resultados del sistema propuesto por AHOLAB, lo que demuestra la robustez del algoritmo desarrollado.

Con el fin de analizar el origen de la mejora de los resultados, se ha estudiado el comportamiento de la fusión en cada una de las clases consideradas. Para ello se ha evaluado el error cometido para cada evento original (voz, música y ruido) considerados individualmente. Se incluyen además el error de omisión (MC) y el error de inserción (FA) de cada clase, referidos al tiempo total asignado a dicha clase en la referencia, tal y como son calculados por las herramientas de evaluación de NIST.

Sistema		MC	FA	SER
AHOLAB	Desarrollo	3,9 %	0,8 %	4,63 %
S6		0,8 %	1,6 %	2,34 %
Fusión		**4,3 %**	**0,4 %**	**4,12 %**
AHOLAB	Evaluación	3,3 %	0,9 %	4,19 %
S6		0,6 %	3,0 %	3,56 %
Fusión		**3,6 %**	**0,5 %**	**4,12 %**

Tabla 4: Detalle del error cometido por los sistemas Aholab y S6 y la fusión de ambos para la clase 'voz'

La tabla 4 muestra el resultado obtenido para la clase de voz en la parte de desarrollo y de evaluación de la base de datos. Se pue-

de observar cómo el resultado de la fusión en este caso es prácticamente nulo, debido principalmente a los buenos resultados para esta clase en cada uno de los sistemas originales, con un SER en torno a sólo el 4 % en el sistema AHOLAB, el peor sistema en este caso.

Sistema		MC	FA	SER
AHOLAB	Desarrollo	26,8 %	4,1 %	30,86 %
S6		56,6 %	1,8 %	58,41 %
Fusión		**25,4 %**	**4,5 %**	**29,94 %**
AHOLAB	Evaluación	36,9 %	6,7 %	43,59 %
S6		64,3 %	1,4 %	65,76 %
Fusión		**33,9 %**	**5,4 %**	**39,34 %**

Tabla 5: Detalle del error cometido por los sistemas Aholab y S6 y la fusión de ambos para la clase 'música'

El resultado obtenido para la clase de música se muestra en la tabla 5. En este caso se consigue una reducción del SER del 10 % en la parte de test de la base de datos. El algoritmo desarrollado permite utilizar las diferencias para esta clase (en realidad cuatro clases tras el mapeo realizado: música, voz con música, música con ruido y música con voz y ruido) entre los dos sistemas para mejorar el resultado final, a pesar de que los resultados del sistema S6 son considerablemente inferiores a los del sistema propuesto por el grupo AHOLAB.

Por último, la tabla 6 muestra el resultado obtenido para la clase de ruido. Se puede observar cómo, al igual que en el caso de la voz, el resultado de la fusión en este caso es casi inapreciable, debido principalmente a los resultados del sistema S6 con un 162,18 % de SER (recordemos que el porcentaje está referido al tiempo total asignado a cada clase en la referencia, por lo que puede superar el 100 %). En este caso no es posible extraer información de utilidad con la que mejorar los resultados. Sin embargo, también cabe resaltar que el resultado final no se ve comprometido a pesar de estas tasas de error superiores al 100 % y se mantiene del orden de lo logrado por el mejor de los sistemas, lo que demuestra la robustez del método desarrollado.

Tras realizar este estudio del comportamiento respecto a cada etiqueta, se puede observar cómo la mejora obtenida al aplicar el método de fusión propuesto se debe al resultado obtenido en la clase 'música', que fue la que más dificultades de clasificación planteó en la campaña Alabyzin 2012.

Sistema		MC	FA	SER
AHOLAB	Desarrollo	33,3 %	9,5 %	42,85 %
S6		52,7 %	61,3 %	113,95 %
Fusión		**36,0 %**	**7,5 %**	**43,57 %**
AHOLAB	Evaluación	34,8 %	28,2 %	63,03 %
S6		42,5 %	119,6 %	162,18 %
Fusión		**38,8 %**	**24,8 %**	**63,61 %**

Tabla 6: Detalle del error cometido por los sistemas Aholab y S6 y la fusión de ambos para la clase 'ruido'

A continuación, este algoritmo se ha aplicado también a la fusión entre el sistema propuesto por el grupo AHOLAB y el sistema S3 (tercer mejor sistema de la campaña de evaluación de Albayzin 2012). El objetivo de este experimento es comprobar los resultados de la técnica de fusión propuesta cuando los dos clasificadores originales presentan bajas tasas de error. La tabla 7 muestra el resultado, en términos de SER, de la fusión de los dos sistemas.

Sistema	Desarrollo	Evaluación
AHOLAB	19,97 %	25,78 %
S3	21,24 %	28,12 %
Fusión	**16,13 %**	**18,86 %**

Tabla 7: Resultado de la fusión de los sistemas Aholab y S3

Podemos observar cómo en este caso, en el que se cuenta con buenos resultados de partida en ambos sistemas, el algoritmo desarrollado obtiene una reducción relativa del SER del 19.5 % en la parte de desarrollo y del 26.8 % en la parte de evaluación respecto al mejor de los sistemas, lo que demuestra la validez del método de fusión propuesto, cuando se utilizan como datos de partida los de un clasificador tipo.

Al igual que en el caso anterior, se ha estudiado el comportamiento de la fusión en cada una de las clases consideradas. Para ello evaluamos de nuevo el error cometido para cada evento original considerados individualmente.

Sistema		MC	FA	SER
AHOLAB	Desarrollo	3,9 %	0,8 %	4,63 %
S3		0,1 %	5,5 %	5,65 %
Fusión		**1,5 %**	**0,8 %**	**2,32 %**
AHOLAB	Evaluación	3,3 %	0,9 %	4,19 %
S3		0,2 %	8,5 %	8,69 %
Fusión		**1,7 %**	**1,4 %**	**3,11 %**

Tabla 8: Detalle del error cometido por los sistemas Aholab y S3 y la fusión de ambos para la clase 'voz'

La tabla 8 muestra el resultado obtenido para la clase de voz en la parte de desarrollo y de evaluación de la base de datos. Se puede observar cómo en este caso sí se obtiene cierta mejora, a pesar de contar con peores resultados para esta clase que en el caso de utilizar los resultados de S6. El reparto diferente de clases entre los dos sistemas permite mejorar el resultado final.

Sistema		MC	FA	SER
AHOLAB	Desarrollo	26,8 %	4,1 %	30,86 %
S3		25,9 %	4,1 %	29,93 %
Fusión		**10,3 %**	**6,3 %**	**16,59 %**
AHOLAB	Evaluación	36,9 %	6,7 %	43,59 %
S3		37,5 %	5,4 %	42,91 %
Fusión		**19,2 %**	**6,7 %**	**25,92 %**

Tabla 9: Detalle del error cometido por los sistemas Aholab y S3 y la fusión de ambos para la clase 'música'

En la tabla 9 se presenta el resultado obtenido para la clase de música. Al igual que en el caso anterior, se consigue una reducción considerable del SER, 40 % en la parte de test de la base de datos. En este caso el error original de los sistemas es menor y la mejora obtenida tras aplicar el algoritmo de fusión es más elevada.

La tabla 10 muestra el resultado obtenido para la clase de ruido. Se puede observar cómo, al igual que en el caso de la voz y la música, se ha conseguido una importante reducción del SER, 22 % en la parte de test. En este caso los dos clasificadores aportan información de utilidad, a pesar del elevado error de ambos para esta clase, y el algoritmo de fusión propuesto es capaz de mejorar el resultado de ambos.

Sistema		MC	FA	SER
AHOLAB	Desarrollo	33,3 %	9,5 %	42,85 %
S3		14,3 %	20,8 %	35,09 %
Fusión		**30,3 %**	**4,1 %**	**34,42 %**
AHOLAB	Evaluación	34,8 %	28,2 %	63,03 %
S3		19,0 %	65,9 %	84,87 %
Fusión		**29,5 %**	**19,6 %**	**49,07 %**

Tabla 10: Detalle del error cometido por los sistemas Aholab y S3 y la fusión de ambos para la clase 'ruido'

Realizado el estudio respecto a cada etiqueta, se puede observar cómo hemos obtenido una mejora considerable de los resultados en todas las clases, con una reducción importante del SER, salvo en la clase de voz en la que los resultados originales eran buenos y el margen de mejora era menor. Esto demuestra la validez del método de fusión propuesto.

5. Conclusiones

Se ha descrito un nuevo sistema de fusión de clasificadores a nivel de etiqueta en base a las matrices de confusión de cada clasificador, obtenidas a partir de la parte de desarrollo de la base de datos utilizada. Además se ha probado con éxito su uso en la combinación de las salidas de sistemas de segmentación de audio propuestos en la campaña de evaluación Albayzin 2012.

Asimismo, se ha comprobado la robustez del algoritmo, obteniendo una mejora de los resultados aún cuando uno de los sistemas utilizados presenta originalmente una tasa de error considerablemente elevada, sin empeorar los resultados del mejor de los dos sistemas considerados.

Este algoritmo de fusión puede ser utilizado para combinar los resultados de más de dos sistemas de clasificación, sin más que aplicarlo de manera jerárquica. También es posible ampliar el algoritmo a la fusión de varios clasificadores considerando a la hora de valorar la confiabilidad de la decisión la tasa de falsos negativos de más de un clasificador.

6. Agradecimientos

Este trabajo ha sido financiado parcialmente por la UPV/EHU (Ayudas para la Formación de Personal Investigador), el Gobierno Vasco (proyecto Ber2Tek, IE12-333) y el Ministerio de Economía y Competitividad (Proyecto SpeechTech4All,

http://speechtech4all.uvigo.es/, TEC2012-38939-C03-03).

Bibliografía

Aguilo, M., T. Butko, A. Temko, y C. Nadeu. 2009. A hierarchical architecture for audio segmentation in a broadcast news task. En *Proc. I Iberian SLTech*, páginas 17–20, Porto Salvo, Portugal.

Asman, A. J. y B. A. Landman. 2011. Robust statistical label fusion through consensus level, labeler accuracy, and truth estimation (COLLATE). *IEEE Transactions on Medical Imaging*, 30(10):1779–1794.

Butko, T. y C. Nadeu. 2011. Audio segmentation of broadcast news in the Albayzin-2010 evaluation: overview, results, and discussion. *EURASIP Journal on Audio, Speech, and Music Processing*, 2011(1):1–10.

Jain, A. K. y A. Ross. 2004. Multibiometric systems. *Communications of the ACM*, 47(1):34–40, Enero.

Kittler, J. y M. Hatef. 1998. On combining classifiers. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 20(3):226–239.

Lu, L., H. J. Zhang, y H. Jiang. 2002. Content analysis for audio classification and segmentation. *IEEE Transactions on Speech and Audio Processing*, 10(7):504–516.

Meinedo, H. y J. Neto. 2003. Audio segmentation, classification and clustering in a broadcast news task. En *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, volumen 2, páginas 5–8, Hong-Kong, China.

Moattar, M. H. y M. M. Homayounpour. 2012. A review on speaker diarization systems and approaches. *Speech Communication*, 54(10):1065–1103, Junio.

NIST. 2009. The 2009 (RT-09) Rich Transcription Meeting Recognition Evaluation Plan. Online: http://www.itl.nist.gov/iad/mig/tests/rt/2009/docs/rt09-meetingeval-plan-v2.pdf, accessed on 15 April 2013.

Ore, B. M., R. E. Slyh, y E. G. Hansen. 2006. Speaker Segmentation and Clustering using Gender Information. En *Proceedings IEEE Odyssey'06 Conference*, páginas 1–8.

Reynolds, D., W. Andrews, J. Campbell, J. Navratil, B. Peskin, A. Adomi, D. Kluracek, J. Abramson, R. Mihaescu, J. Godfrey, D. Jones, y S. Xiang. 2003. The SuperSID Project: Exploiting High-level Information. En *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, volumen 4, páginas 784–787, Hong-Kong, China.

Reynolds, Douglas A y P. Torres-Carrasquillo. 2005. Approaches and applications of audio diarization. En *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, páginas 953–956, Philadelphia, USA.

Ross, A. y A. K. Jain. 2003. Information fusion in biometrics. *Pattern Recognition Letters*, 24(13):2115–2125, Septiembre.

Ruta, D. y B. Gabrys. 2000. An overview of classifier fusion methods. *Computing and Information systems*, 7:1–10.

Rybach, D. y C. Gollan. 2009. Audio segmentation for speech recognition using segment features. En *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, páginas 4197 – 4200, Taipei, Taiwan.

Schuller, B. 2012. The Computational Paralinguistics Challenge. *Signal Processing Magazine, IEEE*, (July):97–101.

Tavarez, D., E. Navas, D. Erro, y I. Saratxaga. 2012. Audio Segmentation System by Aholab for Albayzin 2012 Evaluation Campaign. En *Iberspeech*, páginas 577–584, Madrid, Spain.

Xu, L., A. Krzyzak, y C. Y. Suen. 1992. Methods of combining multiple classifiers and their applications to handwriting recognition. *IEEE Transactions on Systems, Man, and Cybernetics*, 22(3):418–435.