Experiencia del I3A en la Evaluación de Reconocimiento de Locutor NIST 2008

Jesús A. Villalba, Carlos Vaquero, Eduardo Lleida, Alfonso Ortega, Antonio Miguel, José E. García, Luís Buera, Óscar Saz

Grupo de Tecnologías de las Comunicaciones (GTC), Instituto de Investigación en Ingeniería de Aragón (I3A), Zaragoza, España
{villalba,cvaquero,lleida,ortega,amiguel,jegarlai,lbuera,oskarsaz}@unizar.es

Abstract. En este artículo se describe el sistema de reconocimiento de locutor implementado por el I3A para la evaluación del NIST 2008. Se dispone de dos sistemas básicos: GMM-UBM likelihood ratio y GMM-SVM. Las señales proporcionadas por el NIST para la evaluación han sido adquiridas a través de diferentes micrófonos y canales de comunicación, se discutirá como afectan las diferentes técnicas de compensación de canal al funcionamiento del sistema. Se presentan los resultados obtenidos durante el desarrollo del sistema sobre la base de datos de 2006 y los obtenidos en 2008.

Palabras Clave: Speaker Recognition, Gaussian Mixture Model (GMM), Support Vector Machine (SVM), Feature Warping, Nuisance Attribute Projection (NAP), NIST.

1 Introducción

Periódicamente el Instituto Nacional de Estándares y Tecnología Americano (NIST) lleva a cabo una evaluación de sistemas de reconocimiento de locutor con el fin de comparar las prestaciones de diferentes técnicas sobre un corpus de test común. Como en años anteriores, la tarea en la que se basa la evaluación 2008 [1] es detección de locutor, es decir, determinar si un determinado locutor está presente en un segmento de voz dado.

Mientras que en años anteriores la evaluación estaba basada básicamente en señales de canal telefónico, en 2008 se incluyen también dentro de la condición principal conversaciones telefónicas y entrevistas grabadas usando diferentes tipos de micrófonos. Para cada uno de los intentos se conoce, para el segmento de entrenamiento y test, si es canal telefónico o micrófono, si es conversación telefónica o entrevista, el idioma y el sexo de locutor.

En la preparación de esta evaluación se han implementado dos sistemas básicos basados en el modelado de las características cepstrales (MFCC) de la señal de voz mediante modelos de mezclas de gaussianas. El primero de ellos es el clásico GMM-UBM [2] consistente en evaluar el log-ratio de verosimilitud entre el modelo del locutor de test y el modelo universal (UBM) que representa al locutor medio. El segundo es el GMM-SVM [5], que utiliza la capacidad discriminativa de las support vector

machines para comparar los supervectores obtenidos concatenando las medias de los GMM de los segmentos de entrenamiento y test, consiguiendo resultados que mejoran a los del GMM-UBM. La presencia en la evaluación de diferentes canales telefónicos y micrófonos degrada considerablemente las prestaciones de estos sistemas, por ello se hace necesaria la aplicación de técnicas de compensación de canal y normalización de scores entre las que se encuentran: sustracción de la media del cepstrum (CMS), Feature Warping [3], Nuisance Attribute Projection (NAP) [6], T-Norm [4] y Z-Norm.

Este artículo se divide en las siguientes secciones: en la sección 2 se describen los sistemas implementados incluyendo la extracción de características, los tipos de clasificadores y las diferentes técnicas de compensación de canal; en la sección 3 se describen los experimentos realizados y bases de datos utilizadas para desarrollar el sistema y los resultados conseguidos con la base de datos NIST 2006 y en la presente evaluación NIST 2008; finalmente en la sección 4 se exponen la conclusiones y los pasos a dar en el futuro.

2 Descripción del Sistema

2.1 Extracción de Características

El Front-End extrae los 16 primeros Mel Frequency Cepstral Coefficients (MFCC) incluyendo el C0 a los que se añaden sus primeras y segundas derivadas tomando tramas de 25 msg. con un desplazamiento de 10 msg. El C0 se elimina manteniendo sólo sus derivadas quedando finalmente un vector de dimensión 47. El banco de filtros triangulares se modifica para encajar con el ancho de banda de canal telefónico 0.3-3.4 kHz y limitar la influencia del ruido de baja frecuencia presente en la mayoría de señales. Para las condiciones que incluyen señal de micrófono, mucho más ruidosa que la telefónica, ya sea en entrenamiento o en test se utiliza el Advanced Front-End del ETSI (AFE) [9], que a las características anteriores añade un filtrado de Wiener.

La selección de tramas de voz se realiza mediante un umbral sobre la log-energía. Para estimar el umbral se modela la distribución de la log-energía de la señal mediante un modelo bigaussiano, La gaussiana de mayor energía se supone asociada al habla del locutor y la de menor energía asociada al ruido. Buscando el valle entre las mismas se puede determinar el umbral de energía que decidirá si una trama es de voz o es ruidosa. Además se selecciona otro umbral 30dB por debajo de la energía de pico del segmento, por si el nivel de ruido es demasiado bajo para estimar correctamente la gaussiana de menor energía. El umbral seleccionado es el mayor de ambos. Este algoritmo funciona correctamente con SNR aceptables pero puede fallar en el caso de grabaciones con micrófonos de campo lejano mucho más ruidosas, en el caso de que el tanto por ciento de tramas escogidas no sea suficiente para modelar a un locutor se selecciona el umbral que deja pasar el 30 % de las tramas de más energía. A esta selección de tramas se aplica un filtro de mediana, para eliminar tramas de voz aisladas, que puedan deberse a ruidos impulsivos. Dicho filtro es asimétrico: Descarta tramas de voz rodeadas de silencio, pero nunca recupera una trama de silencio convirtiéndola en voz.

2.2 GMM-UBM

Como se ha dicho anteriormente, cada locutor se modela mediante una mezcla de gaussianas. Dicho modelo se obtiene mediante adaptación MAP de las medias de un UBM [2].

$$\mu_k = \alpha_k E_k[x] + (1-\alpha_k)\mu_k^{UBM} \quad \textbf{(1)}$$

$$\alpha_k = \frac{c_k}{c_k + \tau} \quad c_k = \sum_t c_{kt} \quad E_k[x] = \frac{\sum_t c_{kt} x_t}{c_k} \quad \textbf{(2)}$$

siendo c_{kt} la probabilidad a posteriori de que la muestra x_t pertenezca a la gaussiana k y α_k el coeficiente de adaptación MAP.

El UBM se estima previamente mediante el algoritmo EM utilizando gran cantidad de señal de diferentes locutores, y representa el modelo del locutor medio. Como en la evaluación del NIST no hay test cruzados hombre-mujer se obtienen modelos universales de hombre y mujer por separado.

La evaluación de cada intento se realiza calculando el valor del ratio de verosimilitud entre el modelo Target y el UBM.

$$LLR = \log[p(O \mid TARGET)] - \log[p(O \mid UBM)] \quad \textbf{(3)}$$

Para evitar tener que evaluar todas las gaussianas de UBM y Target se aprovecha que las gaussianas de ambos modelos son correspondientes para obtener del UBM las N gaussianas de mayor probabilidad y solo evaluar esas en el Target. Esta técnica se conoce con el nombre de Fast-Scoring. Las pruebas indican que es necesario evaluar al menos 10 gaussianas para no perder prestaciones.

2.3 GMM-SVM

Un SVM es un clasificador binario formado por sumas de una función kernel:

$$f(x) = \sum_{i=1}^{L} \alpha_i y_i K(x_i, x) + b \quad \textbf{(4)}$$

El proceso de optimización coloca un hiperplano capaz de separar ambas clases en el espacio de alta dimensionalidad definido por el kernel. Los supervectores de entrenamiento que se encuentran en la frontera de separación constituyen los vectores soporte. El proceso de entrenamiento consiste en la obtención de estos vectores soporte que modelan la frontera de separación.

Como aparece en [5] a partir de la aproximación de la distancia KL se puede obtener el siguiente kernel que es un producto escalar de dos supervectores:

$$K(i,j) = \sum_k w_k (\mu_k^i)^T \Sigma_k^{-1} \mu_k^j = \sum_k (\sqrt{w_k}\Sigma_k^{-1/2}\mu_k^i)^T (\sqrt{w_k}\Sigma_k^{-1/2}\mu_k^j) = \phi(i)\phi(j) \quad \textbf{(5)}$$

De este modo para cada modelo construimos un supervector concatenando sus medias normalizadas por su desviación típica, y ponderadas por la raíz cuadrada del peso de su gaussiana. El kernel es el producto escalar de los supervectores que queremos comparar.

Se entrena un SVM para cada modelo target utilizando la librería SVMTorch [8]. Se pasa a la librería el modelo del locutor como único ejemplo positivo y varios modelos de locutores de background como ejemplos negativos. Para evaluar se utiliza la siguiente función:

$$f(x) = \sum_{i=1}^{L} \alpha_i y_i K(x_i, x) + b = \left(\sum_{i=1}^{L} \alpha_i y_i \phi(x_i) \right)^T \phi(x) + b = w^T \phi(x) + b \quad (\mathbf{6})$$

Donde $\phi(x)$ es el supervector obtenido a partir del modelo que se estima mediante adaptación MAP al fichero de test, $\phi(x_i)$ son los vectores soporte que da como resultado el algoritmo de optimización cuadrática implementado por SVMTorch y $\alpha_i y_i$ los pesos de los mismos. Al ser el kernel un producto escalar no es necesario almacenar todos los vectores soporte como ocurre con otros kernels sino que podemos limitarnos a calcular el vector del hiperplano que separa ambas clases y la evaluación consiste simplemente en el producto escalar del vector del plano por el supervector de test.

2.4 Normalización de Parámetros

Sustracción de la Media del Cepstrum (CMS).
Una gran ventaja de los MFCC es que constituyen una transformación homomórfica, de forma que, en teoría, las convoluciones y efectos de filtrado en el dominio temporal se convierten en sumas en el dominio cepstral. Suponiendo que el canal no varía, la contribución del mismo al valor de los MFCC se convierte, principalmente, en una constante aditiva. Dicha constante se puede eliminar fácilmente restando directamente la media temporal del valor de los MFCC a cada vector de los mismos [10].

Feature Warping.
En el caso de que además de ruido convolucional, haya presente distorsión no lineal y ruido en el canal, el CMS no es capaz de compensar estos efectos. Una técnica que se probado eficaz en estos casos consiste en aplicar una ecualización de histograma a cada componente de tal manera que tenga una distribución fija, generalmente una gaussiana de media cero y varianza unidad. Se ha demostrado experimentalmente [3] que para la tarea de verificación de locutor, la ventana de análisis óptima es de 3 segundos, de forma que dicha ventana se desplaza trama a trama y únicamente se aplica la transformación a la trama situada en el centro de la ventana, considerando únicamente las tramas de voz.

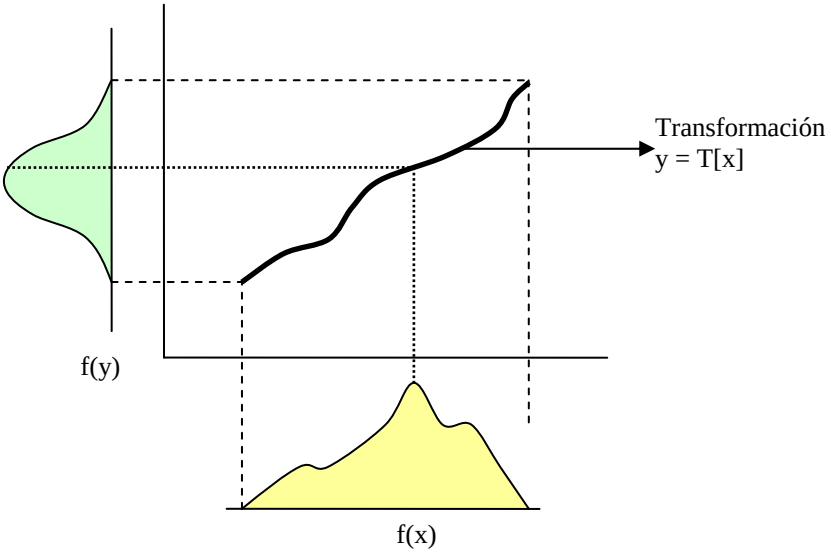

Fig. 1. Feature Warping

2.5 Nuisance Attribute Projection (NAP)

Como técnica de compensación de canal para el sistema GMM-SVM se ha implementado Nuisance Attribute Projection [5][6]. Esta técnica consiste en encontrar una matriz de transformación del tipo $P = I - vv^T$ que proyecte los supervectores en un subespacio vectorial más resistente a las variaciones intra-locutor y de canal. Las columnas de **v** son los k vectores que representan las direcciones de variación del supervector con los cambios de canal o de otros factores. El criterio para obtener P consiste en:

$$v^* = \arg\min_{v, \|v\|_2=1} \sum_{i,j} W_{ij} \left\| P\phi(x_i) - P\phi(x_j) \right\|_2^2 \quad (7)$$

donde W_{ij} debe escogerse de diferente manera en función de las direcciones de variación que se quieran eliminar, W_{ij}=1 si se quiere $\phi(x_i)$ se parezca a $\phi(x_j)$, W_{ij}=-1 si se quiere que sean distintos y W_{ij}=0 si no importa. Utilizando que P^2=P se puede demostrar que los vectores de v se obtienen resolviendo el siguiente problema de autovalores:

$$A(diag(W1) - W)A^T v = \Lambda v \quad (8)$$

donde ***A*** es la matriz cuyas columnas son los supervectores $\phi(x_i)$, **1** es el vector de todo unos y $\boldsymbol{W}=(W_{ij})$.

Para eliminar las direcciones de variación debidas a cambios en el locutor o en el canal entre sesiones se fija W_{ij}=1 si $\phi(x_i)$ pertenece al mismo locutor que $\phi(x_j)$ y W_{ij}=0

en caso contrario. Para este caso concreto podemos desarrollar la formula anterior de tal manera que la matriz de la que hay que hallar los autovalores es:

$$S = \sum_{i=1}^{L} n_i \sum_{j=1}^{n_j} (\phi_i^j - \overline{\phi_i})(\phi_i^j - \overline{\phi_i})^T \quad (\mathbf{9})$$

donde n_i es el número de supervectores del locutor i, ϕ_i^j es el supervector del locutor i en la sesión j $\overline{\phi_i}$ es el supervector medio del locutor i y L es el numero de locutores. Para obtener los autovectores de estas matrices de una forma eficiente se ha seguido el siguiente procedimiento. Se crea una matriz ***B*** cuyas filas son los supervectores de cada locutor menos su media multiplicada por la raiz cuadrada del número de vectores del locutor de tal forma que:

$$B = \left(\sqrt{n_i}(\phi_i^j - \overline{\phi_i})\right)^T; \quad S = B^T B \quad (\mathbf{10})$$

Es inviable calcular la matriz S por su tamaño (varios GB) pero se pueden obtener los autovectores a partir de la descomposición en valores singulares de ***B*** sin necesidad de hacer el producto.

2.6 Normalización de Scores

Los scores que se obtienen de la evaluación del modelo pueden presentar gran variabilidad, debida fundamentalmente al desajuste existente entre la fase de entrenamiento y funcionamiento del sistema, pero también debida a otros factores difíciles de eliminar como el locutor en sí.

T-Norm
Se normaliza el score con la media y varianza de scores que obtiene la señal de test impostando varios modelos de background. De esta forma se acota el rango dinámico de scores que produce una señal de test. Estos modelos se escogen de manera que sean similares al del locutor objetivo, para seleccionarlos se ha utilizado una aproximación de la distancia Kullback Leibler [5]:

$$KL(i, j) \leq \sum_k w_k (\mu_k^i - \mu_k^j)^T \Sigma_k^{-1} (\mu_k^i - \mu_k^j) \quad (\mathbf{11})$$

Z-Norm
Se normaliza el score con la media y varianza de scores que obtiene el modelo del locutor objetivo al ser impostado por varios modelos de background. En este caso lo que se acota es el rango dinámico de scores que produce el modelo objetivo.

3 Experimentos y Resultados

3.1 Efecto de las Técnicas de Compensación de Canal y Normalización de Scores.

Se han realizado experimentos sobre la condición principal NIST 2006 (tlfn-tlfn) orientados a comprobar la mejora que aportan cada una de las técnicas de compensación de canal expuestas en la sección 2 por separado. Para ello se ha partido de un baseline consistente en el sistema GMM-UBM básico con 256 gaussianas al que se han ido añadiendo cada una de estas técnicas. Los datos de desarrollo utilizados vienen descritos en la tabla 4. Los resultados obtenidos se resumen en la tabla 1 y la figura 2:

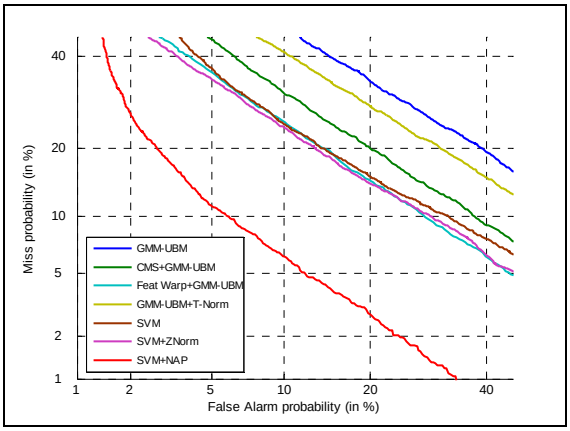

Fig. 2. Curvas DET resumen de técnicas de compensación de canal y normalización de scores.

Tabla 1. Resultados de las Tecnicas de Compensación de Canal.

	ERR(%)	Mejora GMM-UBM (%)	Mejora SVM (%)
Baseline (GMM-UBM)	27.5	0	0
CMS+GMM-UBM	20	27	0
Feat Warp+GMM-UBM	16.8	39	0
GMM-UBM+T-Norm	24.5	11	0
SVM	17.2	37	0
SVM+ZNorm	16.5	40	4
SVM+NAP	7.9	71	54

3.2 Resultados del Sistema Completo en NIST 2006

Una vez comprobado el aporte de cada una de las técnicas por separado se van a proceder a superponer para obtener el sistema completo. Se muestran resultados para la condición principal (tlfno-tlfno) y para la condición cross-channel (tlfno-mic) utilizando modelos de 512 gaussianas.

Tabla 2. Resultados superponiendo diferentes técnicas de compensación y normalización.

Condición	Tlfno-Tlfno		Tlfno-Mic	
	EER (%)	Mejora (%)	EER (%)	Mejora (%)
Warp+GMM-UBM	8.9	0	11.5	0
+TNorm	7.9	+11.2	9.8	+10.6
Warp+SVM	6.9	+12.6	10.7	-9,1
+NAP	5.3	+23.1	5.9	+44.8
+ZNorm	5	+5.6	4.4	+25.4

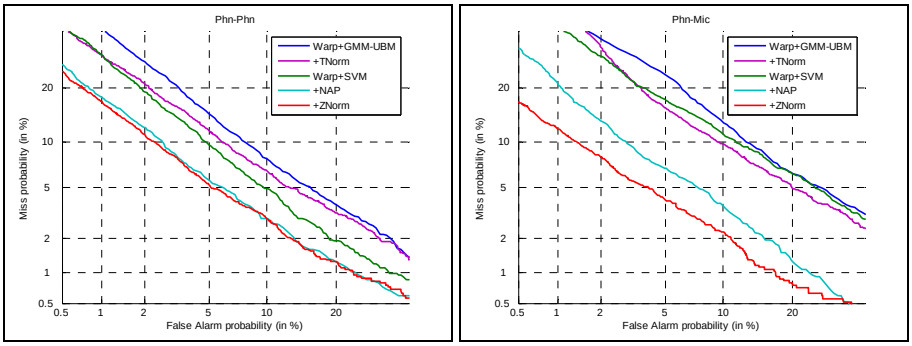

Fig. 3. Curvas DET superponiendo diferentes técnicas de compensación y normalización.

3.3 Resultados del Sistema Completo en NIST 2008

El sistema primario para evaluación 2008 incluye Feature Warping, SVM-NAP y Z-Norm 512 gaussianas. Se ha observado en el caso de que los segmentos tengan pocas tramas, estas no son suficientes para extraer un buen modelo de 512 gaussianas, y se obtienen mejores resultados utilizando modelos de mejor orden. Por ello, uno de los sistemas secundarios enviados consiste en la fusión del sistema primario con 256 y 512 gaussianas utilizando regresión logística lineal [11]. A continuación se presentan los resultados obtenidos en NIST 2008 para las diferentes condiciones existentes.

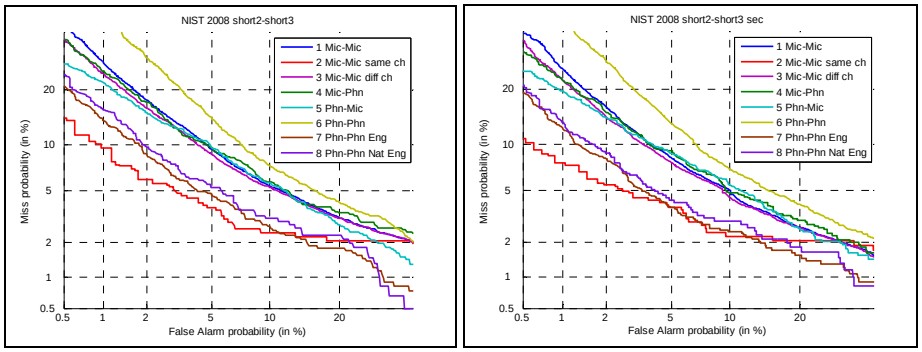

Fig. 4. Curvas DET del sistema en la evaluación NIST 2008 para todas las condiciones.

Tabla 3. Resultados del sistema en la evaluación NIST 2008 para todas las condiciones.

Condición	1	2	3	4	5	6	7	8
EER(%) Primario	7.1	4.1	6.8	7.6	7.2	8.5	4.8	5.2
EER(%) Secundario	6.6	4.2	6.4	6.9	6.9	8.2	4.3	4.6

3.4 Datos de Desarrollo

En función del tipo de condición de entrenamiento y test de cada intento se han utilizado diferentes datos de desarrollo para entrenar modelos de background, proyecciones NAP, etc. Todos los datos empleados pertenecen a evaluaciones previas del NIST. En la tabla siguiente se resumen los datos utilizados en cada caso.

Tabla 4. Datos de desarrollo para las diferentes condiciones.

	Tlfno-Tlfno	Tlfno-Mic	Mic-Tlfno	Mic-Mic
UBM	Entrenamiento 2004	Entrenamiento 2004 + xchannel 2006 y 2005		xchannel 2006 y 2005
NAP	Locutores 2004 con más de una sesión (1500 sesiones)	Tlfno+xchannel 2005 (3000 sesiones)		xchannel 2006 y 2005 (2500 sesiones)
T-Norm/SVM Background	120 Loc. entrenamiento 2004		50 loc x 8 canales xchannel 2005	
Z-Norm	120 Loc. Test 2004	50 loc x 8 canales xchannel 2005	120 Loc. Test 2004	50 loc x 8 canales xchannel 2005

4 Conclusiones

Se ha presentado la evolución del sistema de reconocimiento de locutor del I3A para la evaluación NIST 2008 junto con las prestaciones de diferentes métodos de compensación de canal y normalización de scores. Se ha comprobado que el sistema es robusto al cambio de bases de datos dando tasas de error comparables entre 2006 y 2008. También lo es entre las distintas condiciones de entrenamiento y test de 2008, a pesar de que las señales de micrófono son bastante más ruidosas que las telefónicas, el Feature Warping y el NAP han sido capaces de compensar una gran parte de la variación inter-canal.

En los resultados de 2008 se aprecia una degradación considerable de las prestaciones entre la condición tlfno-tlfno genérica y las que solo incluyen voz en ingles. Esto se debe a que en esta evaluación se han introducido multitud de idiomas que no estaban presentes en evaluaciones anteriores y que por tanto no tienen modeladas sus características en el UBM, NAP, etc. La problemática del idioma tendrá que ser muy tenida en cuenta de cara a próximas evaluaciones a la hora de diseñar el sistema y de escoger los datos de desarrollo, ya que, como muestran los resultados, puede degradar el funcionamiento del sistema tanto o más que las diferencias de canal. También se ha

visto que apenas hay diferencia entre los resultados de la condición ingles genérica y la de ingles sólo hablado por nativos.

Una de las dificultades que se encuentran al construir un sistema de estas características es contar con datos de desarrollo suficientes para entrenar todos modelos de background, NAP y normalizaciones necesarios. Es necesario dedicar esfuerzos a la optimización en la selección de estos datos.

5 Agradecimientos

Este trabajo ha sido financiado por el Ministerio de Educación y Ciencia español a través del Proyecto Nacional TIN 2005-08660-C04-01.

Referencias

1. http://www.nist.gov/speech/tests/sre/2008/sre08_evalplan_release4.pdf
2. D. Reynolds, T. Quatieri, R. Dunn, “Speaker Verification Using Adapted Gaussian Mixture Models” Digital Signal Processing 10, pp 19-41 (2000).
3. Jason Pelecanos, Sridha Sridharan, “Feature Warping for Robust Speaker Verification”, Odissey 2001.
4. D. Ramos, D. Garcia, I. Moreno, J. Gonzalez, “Speaker Verification Using Fast Adaptive TNorm Based On Kullback Leibler Divergence” Third Cost 275 Worshop, Biometrics On The Internet (2005).
5. W. M. Campbell, D. Sturim, and D. Reynolds, “Support vector machines using GMM supervectors for speaker verification,” *IEEE Signal Processing Letters*, vol. 13, no. 5, May 2006.
6. W.M. Campbell, D.E. Sturim, D.A. Reynolds, and A. Solomonoff, “SVM based speaker verification using a GMM supervector kernel and NAP variability compensation,” in *IEEE-ICASSP*, Toulouse, France, 2006.
7. Alex Solomonoff, W. M. Campbell, and I. Boardman, “Advances in channel compensation for SVM speaker recognition,” in *Proceedings of ICASSP*, 2005.
8. R. Collobert and S. Bengio, “SVMTorch: Support vector machines for large-scale regression problems,” *J. Mach. Learn. Res.*, vol. 1, pp.143–160, 2001.
9. ETSI ES 202 050 recommendation, 2002. Speech processing, transmission and quality aspects (STQ); distributed speech recognition; advanced front-end feature extraction algorithm; compression algorithms.
10. Furui Sadaoki. ”Cepstral analysis technique for automatic speaker verification”. IEEE Transactions on speech and audio processing, Vol. ASSP-29, No.2. April 1981
11. http://www.dsp.sun.ac.za/~nbrummer/focal