T-NORM Y DESAJUSTE LÉXICO Y ACÚSTICO EN RECONOCIMIENTO DE LOCUTOR DEPENDIENTE DE TEXTO

Daniel Hernández López[1], Doroteo Torre Toledano[1], Cristina Esteve Elizalde[1], Joaquín González Rodríguez[1], Rubén Fernández Pozo[2] y Luis Hernández Gómez[2]

[1]ATVS Biometric Recognition Group, Universidad Autónoma de Madrid, España
[2]GAPS, SSR, Universidad Politécnica de Madrid, España

RESUMEN

Este trabajo presenta un estudio extenso sobre T-norm aplicado a Reconocimiento de Locutor Dependiente de Texto, analizando también los problemas del desajuste léxico y acústico. Veremos cómo varían los resultados teniendo en cuenta la dependencia de género y realizando T-norm a nivel de frase, fonema y estado con cohortes de impostores de distintos tamaños. El estudio demuestra que implementar T-norm por fonema o estado puede llegar a conseguir mejoras relativas de hasta un 16% y que realizar una selección de cohorte basada en el género puede mejorar más aún los resultados con respecto al caso independiente de género.

1. INTRODUCCIÓN

El Reconocimiento Automático de Locutor es una disciplina de la biometría que consiste reconocer la identidad de una persona (locutor) a través de la voz. Dentro de ésta hay dos grandes vertientes, el Reconocimiento de Locutor Independiente de Texto y el Reconocimiento de Locutor Dependiente de Texto. La segunda de ellas parece haber quedado en segundo plano comparada con la primera, muy probablemente debido a la ausencia de evaluaciones competitivas como las hay para Reconocimiento de Locutor Independiente de Texto [1].

El Reconocimiento de Locutor Dependiente de Texto tiene la particularidad de que el sistema dispone, tanto para entrenamiento como para test, de las trascripciones de la locución. Esto significa que mediante un diccionario fonético podemos disponer de la trascripción fonética de lo que se dice en la locución, lo que hace que se consigan buenos resultados con menor cantidad de habla que en Reconocimiento de Locutor Independiente de Texto. Como es habitual en este tipo de sistemas, en el nuestro utilizamos Modelos Ocultos de Harkov (HMMs) [2] para modelar las características fonéticas de los locutores. Utilizar HMMs permite tener modelos independientes de cada fonema para cada locutor, donde cada uno de los fonemas estará modelado como una serie de probabilidades de transición entre estados, y cada estado estará representado mediante un Modelo de Mezclas de Gaussianas (GMM) [3]. Con estas herramientas y disponiendo de la trascripción fonética, se puede realizar un reconocimiento fonético, utilizando el algoritmo de Viterbi, que proporcione una trascripción fonética con los instantes de comienzo y fin de cada uno de los fonemas y de sus correspondientes estados.

Esta serie de características suponen varias ventajas al Reconocimiento de Locutor Dependiente de Texto frente al Independiente de Texto, pero sin duda la mayor de ellas es poder trabajar, tanto en entrenamiento como en reconocimiento, con niveles por debajo de la frase (palabra, fonema y estado). Dicha ventaja ha sido utilizada múltiples veces en esta disciplina tanto en entrenamiento como en reconocimiento. Entonces ¿porqué no utilizarla en T-norm?

En este trabajo veremos cómo se pueden mejorar los resultados finales del sistema mediante la conocida técnica de T-norm. Hasta ahora esta técnica ha sido muy utilizada en Reconocimiento de Locutor Independiente de Texto y, aunque en menor medida, también en Reconocimiento de Locutor Dependiente de Texto. Sin embargo en todos los casos en que se ha utilizado, T-norm ha sido aplicado a la puntuación global de la locución de test. En el caso de Reconocimiento de Locutor Independiente de Texto parece lógico que se haga así, pero en el caso de Reconocimiento de Locutor Dependiente de Texto parece mejor aprovecharse de la ventaja de poder trabajar con niveles inferiores. Además estudiaremos cómo influye el género y el tamaño de la cohorte de impostores de T-norm.

El resto del artículo está organizado de la siguiente manera: en la Sección 2 describiremos el sistema del que se parte y que ha evolucionado a lo largo de los experimentos, en la Sección 3 explicamos como se implementa T-norm para los experimentos realizados, en la Sección 4 se muestran las bases de datos utilizadas para los experimentos de las Secciones 5 y 6 y por último se presentan las conclusiones en la Sección 7.

2. DESCRIPCIÓN DEL SISTEMA DE PARTIDA

Se parte de un sistema de Reconocimiento de Locutor Dependiente de Texto basado en HMMs. La parametrización que se ha aplicado al audio usado en

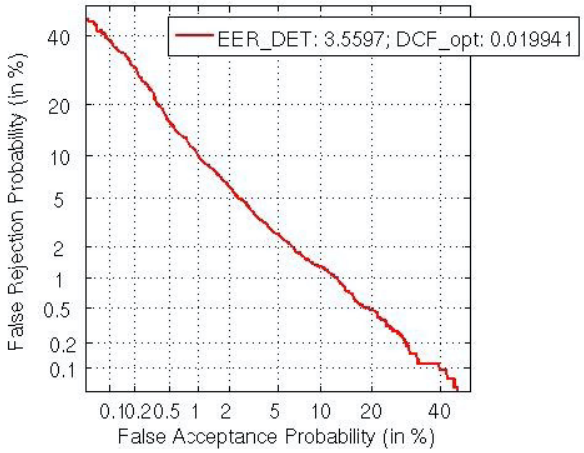

Figura 1. *Curva DET para el sistema de partida.*

entrenamiento y test se basa en la extracción de coeficientes cepstrales mediante filtros de Mel (MFCC, *Mel Frequency Cepstral Coefficients*), tomados en formato $(13+\Delta+\Delta\Delta)$. El sistema de partida realiza alineamiento no completamente forzado de la trascripción tanto para entrenamiento como para verificación (no es totalmente forzado porque se incluyen silencios opcionales entre palabras). Esto es porque aunque se tenga una trascripción textual de lo que se ha pronunciado en la locución no se sabe si hay silencio entre palabras ni de qué duración es éste. De esta forma, realizado un reconocimiento fonético con un modelo de silencio opcional se mejora el alineamiento temporal, lo cual favorece tanto a la etapa de entrenamiento como posteriormente la de reconocimiento.

Dada la trascripción fonética correctamente alineada en el tiempo y el audio parametrizado, el sistema realiza la adaptación al locutor del modelo acústico independiente del locutor. Para ello la adaptación se realiza en tres fases.

En una primera fase se adaptan las medias de las Gaussianas de los Modelos de Mezclas de Gaussianas de cada uno de los estados de cada fonema de forma global. Esto quiere decir que se adaptan todas las medias de forma conjunta. Esta adaptación se hace según el algoritmo MLLR [4] (*Maximum Likelihood Linear Regression*) de forma global, sin clases de regresión. De esta adaptación se obtiene el modelo de transformación lineal, que consiste en una matriz para transformar los modelos fonéticos independientes del locutor en modelos adaptados al locutor. De esta forma no es necesario guardar un modelo de cada locutor, sino simplemente el modelo de transformación. Posteriormente se adaptan los modelos resultantes empleando MLLR, ahora con 2 clases de regresión, obteniendo un nuevo modelo de transformación lineal. Finalmente se aplica adaptación MAP (*Maximum A Posteriori*) [4] a los modelos después de haber sido transformados con el modelo de adaptación MLLR global y posteriormente con 2 clases de regresión.

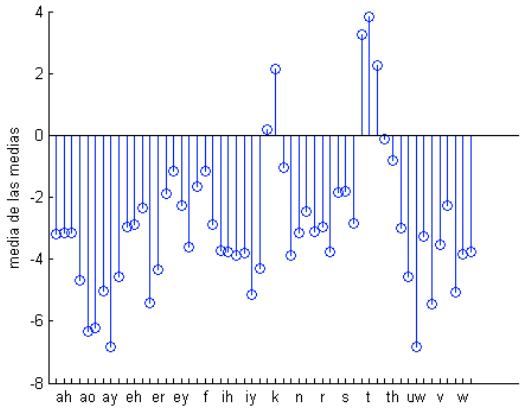

Figura 2. *Medias de las puntuaciones medias de los fonemas (compuestos por 3 estados) y estados en tests de impostores.*

Una vez adaptados los modelos se procede a realizar la fase de evaluación con intentos tanto *target* (donde la locución a verificar es del locutor representado por el modelo) como *non-target* (donde la locución a verificar es de un locutor distinto al representado por el modelo). Se obtienen puntuaciones para cada estado de cada fonema enfrentando la locución de test con el modelo adaptado al locutor y restándola de la puntuación obtenida de enfrentarla al modelo independiente del locutor.

Por último se eliminan las puntuaciones obtenidas por los silencios y se promedia la puntuación general de la locución de test. El resultado obtenido de esta forma para la base de datos YOHO (descrita en la Sección 4), se representa en forma de curva DET en la Figura 1.

3. T-NORM A DISTINTOS NIVELES

Lo que se propone en este artículo es un estudio sobre T-norm, que básicamente consiste en tomar las puntuaciones obtenidas por una cohorte de impostores y calcular la media (μ) y la desviación típica (σ) de dichas puntuaciones. De esta forma se calcula la nueva puntuación ($score_{T-norm}$) como la puntuación obtenida por el locutor que realiza el intento de acceso ($score$) menos la media, dividido entre la desviación típica, como se muestra en la Fórmula 1.

$$score_{T-norm} = \frac{score - \mu}{\sigma} \qquad (1)$$

La clave de este estudio consiste en que se realizará este proceso no sólo a nivel de locución (como suele ser habitual) sino también a nivel de fonema y de estado. Por otra parte se implementará también T-norm dependiente de género. Esto significa que la cohorte de impostores estará compuesta por locutores del mismo género que el locutor *target* para experimentos de este tipo.

La idea de realizar este tipo de T-norm surge de un estudio analítico de las puntuaciones. Como podemos observar en la Figura 2 las puntuaciones de impostor de los estados de un mismo fonema tienen una cierta correlación mientras que entre fonemas las puntuaciones son muy dispares. Esto nos induce a pensar que realizar T-norm a nivel de fonema o estado puede reportarnos buenos resultados debido a que de este modo alinearemos las puntuaciones obtenidas por cada fonema y estado, que parecen desalineadas (Fig. 2).

4. DESCRIPCIÓN DE LAS BASES DE DATOS

4.1. YOHO

Se ha usado la base de datos YOHO [5] para este experimento. Esta base de datos tiene 138 locutores, de los cuales 106 son hombres y 32 son mujeres. Cada locutor presenta 96 locuciones de entrenamiento repartidas en 4 sesiones y 40 locuciones de test repartidas en 10 sesiones. Se han utilizado 6 locuciones de entrenamiento de la primera sesión para realizar el entrenamiento y todas las locuciones de test del locutor como intentos *target* para la etapa de verificación, tomando una locución al azar de cada uno de los demás locutores como intentos *non-target*. Cabe destacar que el léxico de esta base de datos consiste en frases de pares de dígitos (p.e. 32-98-64) y que no hay ninguna relación entre los dígitos pronunciados en entrenamiento y test, con lo cual tenemos un importante desajuste léxico.

4.2. BioSec

Para otro experimento realizado se ha usado la base de datos BioSec Baseline [6]. En esta base de datos hay 150 locutores cuyo idioma nativo es el castellano. Cada locutor ha grabado 2 sesiones con 4 locuciones cada una de un número aleatorio asignado al usuario (el mismo para todas las locuciones de las 2 sesiones). Se han utilizado las 4 frases de la primera sesión para entrenar los modelos acústicos del locutor (se entrena un modelo con cada frase) y las 4 de la segunda sesión como intentos *target* para la fase de test, siendo los intentos *non-target* la primera frase de la primera sesión del resto de impostores (sin enfrentamientos simétricos). Todas las locuciones descritas anteriormente se han realizado de forma idéntica para 4 escenarios. Castellano grabado con un micrófono de unos auriculares (cercano), castellano grabado con un micrófono integrado en una webcam (lejano) y otros 2 escenarios equivalentes en inglés.

5. EXPERIMENTOS CON T-NORM EN FUNCIÓN DEL NIVEL, COHORTE Y GÉNERO

Para poder implementar T-norm, se ha realizado un reconocimiento de cada locución de test con los modelos de locutores de la cohorte de impostores por cada enfrentamiento tanto *target* como *non-target*. De esta forma se han realizado 3 tipos de T-norm en función de la cohorte de impostores: una con una cohorte fija de 20 locutores, 10 hombres y 10 mujeres, a la que llamaremos TN10; otra con una cohorte variable de 60 locutores, 30 hombres y 30 mujeres, a la que llamaremos TN30; una última con una cohorte masculina variable que incluye como impostores todos aquellos locutores que no sean ni el *target* ni el *non-target* (en el caso de que se trate de una prueba *non-target*), a la que llamaremos TNMale.

Para el caso sin T-norm y TN10 hemos realizado experimentos tanto dependientes de género como independientes de género, mientras que para TN30 y TNMale únicamente hemos realizado experimentos dependientes de género. Para los experimentos independientes de género anteriormente descritos se han obtenido los resultados expresados en EER (*Equal Error Rate*) mostrados en la Tabla 1.

T-norm\Nivel	Frase	Fonema	Estado
No	3.56		
TN10	3.91	2.98	3.04

Tabla 1. *EERs (%) obtenidas para distintos tipos de T-norm independiente de género en función del nivel.*

Como podemos ver en la Tabla 1 resulta mucho mejor realizar T-norm a nivel de estado o fonema que a nivel de frase, de hecho podemos ver que es incluso mejor no implementar T-norm que hacerlo a nivel de frase para este experimento en concreto. A continuación vemos en la Tabla 2 como evolucionan los resultados al incrementar el número de impostores de la cohorte y realizar una selección por género de la cohorte de impostores a utilizar.

T-norm	Género	Frase	Fonema	Estado
No	Masc	3.54		
	Fem	3.72		
TN10	Masc	3.32	2.64	2.80
	Fem	3.57	3.15	2.45
	Ambos	3.64	2.97	2.91
TN30	Masc	2.69	2.48	2.46
	Fem	3.99	3.67	3.67
	Ambos	3.10	2.98	2.96
TNMale		2.57	2.41	2.53

Tabla 2. *EERs (%) obtenidas para distintos tipos de T-norm en función del nivel y género.*

En la Tabla 2 vemos cómo varían las tasas de error obtenidas en función del género y el número de impostores de la cohorte. En líneas generales podemos observar que parece ser que cuanto mayor es la cohorte de impostores mejor funciona el sistema, debido probablemente a que al tener un número mayor de impostores tenemos más probabilidades de encontrarnos con modelos próximos al del locutor *target*. Sin

embargo esto no se cumple para todos los casos y es debido, muy probablemente, a que también nos encontraremos más modelos que se alejen del modelo del locutor *target*. Por otra parte también vemos que se generaliza la suposición de que es mejor realizar T-norm a nivel de fonema o estado que a nivel de frase. Además vemos que no hay mucha diferencia entre realizarlo a nivel de estado o fonema, ya que hay casos en los que resulta mejor uno que otro y viceversa.

6. OTROS EXPERIMENTOS

A fin de extender nuestra experimentación al idioma castellano realizamos también experimentos con la base de datos BioSec Baseline [7]. Esta base de datos, aparte de permitirnos comparar resultados en inglés y castellano con el mismo entorno experimental nos permite comparar la influencia del canal de grabación (micrófono de habla cercana frente a micrófono de habla lejana) y la influencia de la coincidencia léxica entre entrenamiento y test (cosa que ocurre en BioSec pero no en YOHO). Otra diferencia con los resultados presentados anteriormente es que en los resultados con BioSec se ha empleado únicamente MLLR y no MAP.

Canal\Idioma	Castellano	Inglés
Mic. cercano	1.68	2.17
Mic. lejano	17.24	12.72

Tabla 3. *EER (%) obtenidas para distintos tipos de micrófono e idioma.*

Como podemos observar en la Tabla 3 la calidad del micrófono supone una gran contribución a la eficiencia del sistema de Reconocimiento de Locutor Dependiente de Texto. Vemos que los resultados obtenidos con el micrófono de la webcam son mucho peores que con el micrófono cercano integrado en los auriculares. Se ha de indicar que en estos experimentos no se han utilizado técnicas de compensación de canal de ningún tipo (salvo CMN).

Sin desajuste	7.02
SNR	7.47
Canal	9.76
Léxico (2 dígitos en común)	8.23
Léxico (1 dígito en común)	13.4
Léxico (0 dígitos en común)	36.3

Tabla 4. *EER (%) obtenida para distintos tipos de desajuste para el estudio realizado en [8].*

Por otra parte si nos fijamos en los resultados para micrófono cercano observamos que son mucho mejores para esta base de datos que con YOHO (utilizando la misma técnica sólo con MLLR en YOHO el EER resultante es de 4.82%). La principal diferencia entre ambas bases de datos es el desajuste léxico existente en YOHO e inexistente en BioSec. El problema del desajuste léxico ya se ha analizado con anterioridad [8] y se ha comprobado (ver Tabla 4 con resultados publicados en [8]) que el desajuste léxico puede ser el tipo más perjudicial de desajuste, incluso peor que el de canal.

7. CONCLUSIONES

Dados los resultados obtenidos en los diferentes experimentos se puede concluir que el desajuste léxico tiene una gran influencia en la eficiencia de un sistema de Reconocimiento de Locutor Dependiente de Texto. La principal razón es que en este campo se entrenan modelos de unidades léxicas por debajo de la locución completa (palabra, fonema, tri-fonema, estado…). Y el hecho de que se intente reconocer al locutor con modelos de unidades léxicas que hemos podido no entrenar previamente (desajuste léxico), o que hemos entrenado en contextos léxicos distintos, hace que los resultados empeoren de forma muy abultada.

Demostrado esto, el principal objetivo de la técnica de T-Norm a nivel de fonema y estado era tratar de reducir la influencia del desajuste léxico, para así ponderar la influencia de cada fonema en el proceso de verificación. Aunque los resultados obtenidos con la normalización a nivel de estado y fonema son positivos, superando los resultados a nivel de frase, el problema del desajuste léxico sigue sin estar resuelto y sigue teniendo una influencia importante en los resultados.

8. BIBLIOGRAFÍA

[1] "National institute of standard and technology. Speaker Recognition Evaluation Home Page", http://www.nist.gov/speech/tests/sre/

[2] L. R. Rabiner, "A tutorial on Hidden Markov Models and selected applications in speech recognition", Proceedings of the IEEE, vol 77, no 2, pp. 257-286, Febrero 1989.

[3] D. A. Reynolds, T. F. Quatieri, R. B. Dunn, "Speaker verification using adapted gaussian mixture models", Digital Signal Processing, vol 10, no 1, pp. 19-41, Enero 2000.

[4] D. T. Toledano, D. Hernandez-Lopez, C. Esteve-Elizalde, J. Gonzalez-Rodriguez, R. Fernandez, L. Hernandez, "MAP and sub-word level T-norm for text-dependent speaker recognition", to appear in Interspeech 2008.

[5] J. P. Campbell, "Testing with the YOHO CD-ROM voice verification corpus", Proc. ICASSP, vol 1, pp. 341-344, 1995.

[6] J. Fierrez, J. Ortega-Garcia, D. T. Toledano, J. Gonzalez-Rodriguez, "Biosec baseline corpus: A multimodal biometric database", Pattern Recognition, vol 40, no 4, Abril 2007.

[7] D. T. Toledano, D. Hernandez-Lopez, C. Esteve-Elizalde, J. Fierrez, J. Ortega-Garcia, D. Ramos, J. Gonzalez-Rodriguez, "BioSec Multimodal Biometric Database in Text-Dependent Speaker Recognition", Proc. LREC, Mayo 2008.

[8] D. Boies, M. Hébert, L. P. Heck, "Study of the effect of lexical mismatch in text-dependent speaker verification", Proc. Odyssey Speaker Recognition Workshop, vol 1, pp. 135-140, Junio 2004.