Speaker Verification on Summed-Channel Conditions with Confidence Measures

Verificación de locutor en condiciones de canal sumado con medidas de confianza

Carlos Vaquero Avilés Casco, Jesús Villalba López, Alfonso Ortega Giménez, and Eduardo Lleida Solano
Communications Technology Group (GTC), Aragón Institute for Engineering Research (I3A),
University of Zaragoza, Spain
{cvaquero, villalba, ortega, lleida}@unizar.es

Article received on July 30, 2010; accepted on January 15, 2011

Abstract. This paper addresses the problem of speaker verification in two speaker conversations, proposing a set of confidence measures to assess the quality of a given speaker segmentation. We study how these measures can be used to estimate the performance of a state-of-the-art speaker verification system, the I3A submission for the core-summed condition in the NIST SRE 2010. We present a Factor Analysis based speaker segmentation system, along with three confidence measures that are fused to obtain a single measure that we show to constitute a good estimation of the segmentation accuracy, when evaluated on the summed-channel telephone data of the NIST SRE 2008. Finally we present speaker verification results obtained with the I3A submission for the NIST SRE 2010 on several conditions of this evaluation, involving summed-channel. We show that the confidence measure also predicts the performance of a state-of-the art speaker verification system when it faces two speaker conversations.
Keywords. Confidence measures, speaker segmentation, speaker verification and telephone conversations.

Resumen. Este artículo trata el problema de verificación de locutor en conversaciones con dos locutores, proponiendo un conjunto de medidas de confianza para evaluar la calidad de una segmentación de locutores dada. Estudiamos cómo estas medidas pueden ser utilizadas para estimar el rendimiento de un sistema de verificación del locutor del estado del arte, el sistema del I3A para la evaluación de reconocimiento del locutor NIST SRE 2010. Presentamos un sistema de segmentación de locutor basado en Análisis Factorial y tres medidas de confianza que son combinadas en una medida que constituye una buena estimación de la calidad de la segmentación, cuando se evalúa en las grabaciones de canal sumado de la NIST SRE 2008. Finalmente presentamos resultados de verificación de locutor obtenidos con el sistema del I3A en distintas condiciones de canal sumado de la NIST SRE 2010. Se demuestra que las medidas de confianza también predicen el rendimiento de un sistema de verificación del locutor cuando se enfrenta a conversaciones de dos locutores.
Palabras clave. Medidas de confianza, segmentación de locutor, verificación de locutor y conversaciones telefónicas.

1 Introduction

Recently, there has been a great advance in the field of speaker identification, in part motivated by the NIST Speaker Recognition Evaluations (SRE). One of the main breakthroughs of the last years has been the formulation of the Joint Factor Analysis (JFA) for speaker verification [Kenny, *et al.*, 2008]. Nowadays most state of the art speaker verification systems are based on this approach. Since then, researchers have explored its application to different areas, especially to study new speaker diarization methods. One of the most interesting of these methods is the one presented in [Castaldo *et al.*, 2008], a novel approach for streaming speaker diarization, which shows several differences with traditional diarization systems. This method makes use of a simple Factor Analysis (FA) model composed only of eigenvoices [Kuhn *et al.*, 2000] to obtain high accuracy in a two

speaker segmentation task on telephone conversations. However, performance decreases significantly when the number of speakers is unknown.

Consequently, the speaker identification community has focused on improving the performance in the two speaker segmentation task on telephone conversations, a task quite related to speaker verification. In [Reynolds, *et al.*, 2009] several approaches using JFA and Variational Bayes are proposed and compared to a traditional Bayesian Information Criterion (BIC) based Agglomerative Hierarchical Clustering (AHC) system [Reynolds and Torres Carrasquillo, 2005]. In that study, results are reported in terms of segmentation error on the NIST SRE 2008 summed dataset. Most approaches show higher accuracy than the classical AHC system, including the streaming eigenvoice based approach; however, this last system is outperformed by two Variational Bayes based systems. The first one is a classical AHC system that makes uses of Variational Bayes to perform a final resegmentation. The second one applies Variational Bayes to build iteratively eigenvoice based speaker models.

In this work we address the problem of speaker verification in two speaker conversations and how a set of confidence measures that assess the quality of a given speaker segmentation can be used to estimate the performance of a speaker verification system, enabling us to identify those test recordings that will give good results on speaker verification. We use the eigenvoice based approach for two speaker segmentation and the confidence measures presented in [Vaquero *et al.*, 2010], and the state-of-the-art speaker verification system presented in [Villalba *et al.* 2010]. Results are presented on the NIST SRE 2010, and such results, combined with those presented in [Vaquero, *et al.*, 2010] show that the proposed approaches are valid across different datasets.

In Section 2 we describe the proposed segmentation system, and three reliable confidence measures to estimate the segmentation performance are presented in Section 3. In Section 4, we evaluate the speaker segmentation system and the confidence measures for speaker segmentation, while in Section 5 we analyze the performance of the mentioned speaker verification system when using the segmentation system and confidence measures proposed. Finally, in Section 6 we summarize the conclusions of this study.

2 Speaker Segmentation

The proposed speaker segmentation system is described in [Vaquero, *et al.*, 2010]. We use a factor analysis approach to model the desired sources of variability. As a starting point we try to capture the variability present among different speakers. For this purpose, we model every speaker by a Gaussian Mixture Model (GMM) adapted from a Universal Background Model (UBM) using an eigenvoice approach [Kuhn *et al.*, 2008], according to

$$M_s = M_{UBM} + Vy \quad (1)$$

where M_s is the speaker GMM supervector, obtained concatenating all Gaussian means, M_{UBM} is the UBM supervector, V is the low rank eigenvoice matrix, and y is the set of speaker factors, which follows a standard normal distribution $N(y|0,I)$ a priori. This way, every speaker is represented by a GMM supervector in a high dimensional space, and in such space we allow the speakers to lie in the low dimensional subspace generated by the column vectors of V, which point to the directions of maximum variability among speakers. We refer to this variability as inter-speaker variability and to the low rank subspace as the speaker subspace.

In our approach we use a 256 Gaussian UBM, and as feature vectors we use 12 Mel Frequency Cepstral Coefficients (MFCC) including C0, computed every 10 ms over a 25 ms window. The dimension of the speaker subspace is 20, compared to the dimension of the supervector space that is $256x12=3072$. This way every point estimate for a given speaker is defined by a set of 20 speaker factors.

To perform speaker segmentation given a sequence of feature vectors, as in [Castaldo *et al.*, 2008], we estimate the speaker factors for

every frame over a 100 frame window, with an overlap of 990 ms, an we estimate a 2-Gaussian GMM to model the stream of speaker factors obtained, after removing silence frames according to a Voice Activity Detector (VAD). Each one of these Gaussians will be assigned to a single speaker. In contrast to [Castaldo *et al.*, 2008], we estimate the GMM using all available data in the recording, rather than processing 1 minute slices and applying a clustering technique. The later allows stream processing with 1 minute latency but the former yields better results. A block diagram of the proposed segmentation system is shown in Figure 1.

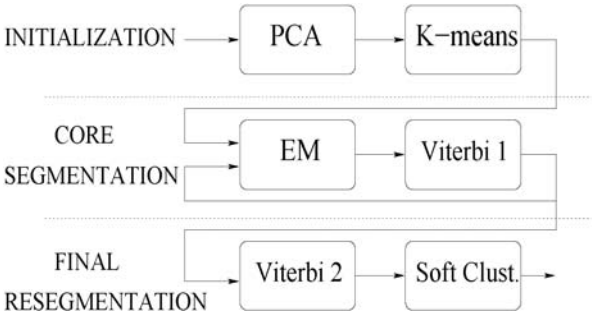

Fig. 1. Block Diagram of the proposed segmentation system

2.1 Initialization

We have detected that a good initialization is quite important to ensure that every Gaussian in the GMM corresponds to a single speaker. In our approach, we use prior knowledge about speaker factors proposed in [Kenny *et al.*, 2008]: A priori, speaker factors are assumed to be distributed according to the standard normal distribution $N(y|0,I)$. Since we obtain speaker factors from a small data sample (100 frames, which is small compared to the number of frames that speaker recognition systems usually manage, around 10000), using MAP estimation, we can expect the posterior distribution of speaker factors for a single speaker to keep some properties of the prior. Assuming that the posterior variance is close to I, we can perform PCA to obtain the direction of maximum variability in the speaker factor space. Such direction should be the best one to separate speakers, since both are supposed to have a variance close to I and a different mean.

This strategy gives two clusters that can be seen as first speaker segmentation, and then, K-means clustering is performed to reassign frames to the two clusters and a single Gaussian is trained on each of them. Using this frame assignment as segmentation output gives reasonably good results, as we will see later, in Section 4.

2.2 Core Segmentation

The 2 Gaussians previously trained serve as initial GMM of the whole recording. Then a two stage iterative process is applied until convergence: first several Expectation-Maximization (EM) iterations are used and then, every Gaussian is assigned to a single speaker and a Viterbi segmentation is performed (Viterbi 1 in Figure 1). According to this new frame assignment, 2 Gaussian models are trained and the iterative process restarts again. Convergence is reached when the segmentation of the current iteration is identical to that obtained in the previous one.

To avoid false fast speaker changes, in the Viterbi segmentation, we modify the speaker turn duration distribution using a sequence of tied-states [Levinson, 1986] for every speaker model. This way, we avoid the state duration to follow a geometric distribution that cannot accurately model real speaker turn durations. Each speaker model is composed of 10 states that share the same observation distribution, a single Gaussian in this case. Tied-states are not considered for the silence, but a single state without an observation distribution is used, since the algorithm is forced to go through the silence state according to the VAD labels. We have observed that this way of modeling speaker turn duration yields better results than modifying the transition probability.

2.3 Viterbi Segmentation and Soft Clustering

The output of the core segmentation system gives accurate speaker labels in most cases, but

these labels can be refined by means of Viterbi resegmentations (Viterbi 2 in Figure 1).

In this case we model every speaker with a 32 component GMM according to the output of the core segmentation system using as features 12 MFCC including C0. Again we use 10 tied-states for speaker models and a single state for all silence frames.

After this resegmentation we retrain the GMM models and run a forward backward decoding to perform a soft reassignment of the frames to the two speakers. GMM models are retrained according to the soft reassignment and a final Viterbi resegmentation is performed. This approach was first presented in [Reynolds *et al.*, 2009] as soft-clustering.

3 Confidence Measures

In the following section we describe a set of confidence measures that aims at determining the performance of the segmentation system explained in the previous section for a given audio recording. These set of confidence measures is described and analyzed in [Vaquero *et al.*, 2010].

3.1 Bayesian Information Criterion

BIC has been successfully applied to the task of speaker diarization, both for speaker segmentation and speaker clustering. Currently, most speaker diarization systems rely on BIC to perform AHC [Reynolds and Torres Carrasquillo, 2005]. In such systems, BIC is used both to decide the next pair of closest clusters to merge and as a stopping criterion, to decide the final number of speakers in the current audio recording. In our task the number of speakers is priorly known, so we do not need a stopping criterion to make that decision. However, BIC can be used as a measure of the accuracy of a given segmentation.

In this approach, given two sequences of acoustic feature vectors obtained by the segmentation system, we compute the BIC for two hypotheses: Each sequence belongs to a different speaker or both sequences belong to the same speaker. The confidence measure is the difference between BIC values. To avoid adjusting BIC penalty parameters, we force the models for both hypotheses to have the same complexity. That is, we model every speaker in the first hypothesis with a GMM of *N* Gaussians, and the global model in the second hypothesis with a GMM of *2N* Gaussians. In our experiments we set *N* to 32 Gaussians.

3.2 Kullback-Leibler Divergence in the Speaker Factor Space

Another way to measure the accuracy of a given segmentation is to compute the symmetric Kullback-Leibler (KL) divergence between the Gaussian speaker models obtained in the speaker factor space. In this approach we use the hypothetic segmentation labels to obtain two sequences of speaker factors, and Gaussian models are trained for each sequence. We can expect higher KL divergences between both Gaussian models when the segmentation is correct (i.e. the models are pure).

3.3 Core Segmentation System Convergence

Previous measures were based on the principle that if the segmentation is accurate we can build pure and separate models for every speaker, so both measures will be quite correlated. In Section 2 we saw that the core segmentation runs until convergence. A way to estimate the quality of the output of the core segmentation system is to study how long it took to converge. We can expect the system to converge fast when it can easily find the correct segmentation and to converge slowly otherwise. This measure is probably less correlated with the previous measures described.

4 Speaker Segmentation Experiments

4.1 Experimental Setup

To evaluate the proposed segmentation system and the confidence measures, the 2213 five minute telephone conversations from the NIST

SRE 2008 summed channel condition are considered. Performance is measured in terms of segmentation error rate, obtained as specified in the NIST SRE 2000 speaker segmentation task. In all cases speech/non-speech and reference segmentation labels are derived from Automatic Speech Recognition (ASR) transcripts provided by NIST as in [Reynolds *et al.*, 2009].

4.2 Segmentation Performance

As we explained in Section 2, the proposed segmentation system comprises several steps, including PCA initialization, K-means clustering, iterative EM and Viterbi segmentation in the speaker factor space, a Viterbi resegmentation using MFCC features and a last soft-clustering resegmentation. Table 1 shows the results obtained by the segmentation system after every step.

Table 1. Block Diagram of the proposed segmentation system

Segmentation System	Segmentation Error	Typical dev
PCA	20.2%	14.3%
+K-means	4.9%	8.8%
+Core segmentation	3.1%	6.6%
+Viterbi resegmentation	2.3%	6.2%
+Soft-clustering	2.2%	6.1%

Given these results we can extract several conclusions. First, speaker factors enable easy separability between speakers. Just with PCA and using one dimension to classify the frames we get 20.2% segmentation error. Compared to the eigenvoice based system presented in [Castaldo *et al.*, 2008] and evaluated on the same dataset in [Reynolds *et al.*, 2009], we can see that our approach outperforms that one just using PCA initialization and K-means clustering.

Note that at that point, frames are assigned to one speaker or the other assuming statistical independence, no context or temporal information is used. Completing the core system gives great improvement and results are comparable to those obtained with the best systems presented in [Reynolds *et al.*, 2009]. Moreover, after resegmentations results improve further. We believe that the key improvements to outperform the system in [Castaldo *et al.*, 2008] are the novel PCA initialization and the modification on the speaker turn duration distribution.

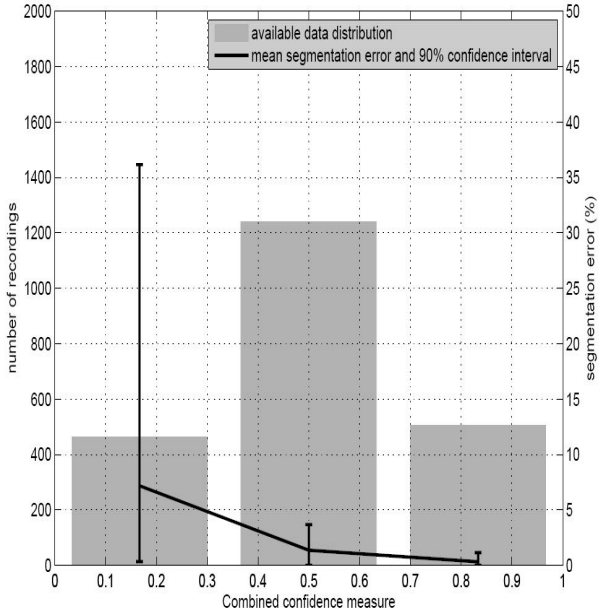

Fig. 2. Segmentation error and data distribution for the fused confidence measure

4.3 Confidence Measures

To analyze the proposed confidence measures, first we normalize them to be in the range [0,1] and then we divide the dataset into 3 subsets according to a uniform division of the confidence measure range. We combine the confidence measures applying Linear Logistic Regression using the FoCal toolkit [Brummer]. For this purpose we optimize the weights in order to detect those recordings that have less than 5% segmentation error, since it has been suggested that low segmentation errors does not impact in speaker verification performance [Reynolds *et al.*, 2009]. Both normalization and Linear Logistic

Regression is made using the NIST SRE 2008 data, since we only have ground truth segmentation labels for this task. However, we will see in Section 5 that the fused confidence measure performs as expected in a different dataset, at least in terms of Speaker Verification Performance.

Figure 2 represents the distribution of the recordings and the mean segmentation error with the 90% confidence interval (CI) over the previously proposed confidence measure ranges, for the fused confidence measure. We can observe that all confidence measures proposed follow the expected behavior: as they increase, the mean segmentation error decreases and so does the 90% CI. This way, we can assure that given a segmentation output with a high value in its confidence measure there is a high probability of having a good segmentation. However, we cannot assure that given a low confidence measure the segmentation is wrong, since the CI is large is that case. This behavior does not allow us to predict the segmentation error given the confidence measures in all cases, but it is enough to consider them as an indicator of the segmentation quality.

5 Speaker Verification Experiments

To evaluate the effect of our speaker segmentation system and the confidence measures on the speaker verification performance we have conducted experiments on the core-summed and 8summed-core conditions of the NIST Speaker Recognition Evaluation 2010 [SRE, 2010]. For that purpose, we have used a state-of-the-art JFA system.

5.1 Speaker Verification System Description

As speaker verification system, we use the I3A submission for the NIST SRE 2010 [Villalba *et al.* 2010]. This is a SV system based on JFA [Kenny, *et al.*, 2008]. Feature vectors of 20 MFCC (C0-C19) plus first and second derivatives are extracted. Voice Activity Detection (VAD) is performed computing the Long-Term Spectral Divergence (LTSD) of the signal every 10 ms, and comparing it against a threshold [Ramirez, *et al.*, 2004]. After frame selection and segmentation, every feature stream is short time Gaussianized as in [Pelecanos and Sridharan, 2001].

A gender independent Universal Background Model (UBM) of 2048 Gaussians is trained by EM iterations. Then 300 eigenvoices *v*, 100 eigenchannels *u* and the residual variability matrix *d* are trained by EM ML+MD iterations. We have used all telephone data from SRE04, SRE05 and SRE06 for UBM and JFA training.

Speakers are enrolled using MAP estimates of their speaker factors (*y*, *z*) so that the speaker means super vector is given by

$$M_s = m_{UBM} + vy + dz \tag{2}$$

For the 8 summed-channel training condition, we have clustered the streams belonging to the target speaker prior to the estimation of the model. First, we calculate the speaker factors of each of the streams separately and then, we use a criterion based on the cosine distances between the factors of the different streams for selecting the ones belonging to the same speaker. Given a set o possible stream selections

$$S = \{(s_1, s_2, ..., s_N) \,|\, s_i \in \{1, 2\}\} \tag{3}$$

where *N* is the number of conversations. We choose the stream combination I_{opt} such as

$$I_{opt} = \arg\max_{I \in S} \sum_{i=1}^{N-1} \sum_{j=i+1}^{N} \frac{y_{I(i)}^t y_{I(j)}}{\|y_{I(i)}\| \, \|y_{I(j)}\|} \tag{4}$$

.

Finally, we accumulate the statistics of the selected streams to estimate the target speaker model.

Trial scoring is performed using first order Taylor approximation of the LLR between the

target and the UBM Models like in [Glembek *et al.*, 2009].

$$LLR \approx (vy_{trn} + dz_{trn})^t \Sigma^{-1}(F_{tst} - N_{tst}(ux_{tst} + m_{UBM})) \quad (5)$$

Finally, scores are gender dependent ZT Normalized using data from SRE04, SRE04 and SRE06 (628 male speakers and 858 female speakers with 4 segments by speaker). For the core-summed condition, the maximum score of the two automatic segmented speakers is chosen.

5.2 Results Core-summed

Figure 3 shows Detection Error Trade-off (DET) curves for the NIST SRE10 core-summed det5 condition. On the one hand, we present results for the full trial list and, on the other, for three different subsets of trials split according to the fused confidence measure described in section 4.3. Minimum and actual NIST Detection Cost Function (C_{Miss}=10, C_{FA}=1, P_{Target}=0.01) are marked on the curve with a point and a cross respectively. Output scores have been calibrated to log-likelihood ratios by linear logistic regression using the FoCal package [Brummer] with the matching condition of the NIST SRE08 short2-summed condition. In this manner, actual costs are calculated applying the Bayesian threshold of 2.29. Table 2 presents the EER and cost values for the different confidence intervals, together with the number of trials belonging to each subset.

The performance of the system on the core-summed condition is not far from the performance on the core-core condition in which we have a 2.4% of EER. We can appreciate a fair correlation between the confidence and the performance. The subset with higher confidence is a 48% better than the subset with lower confidence in terms of EER and a 26% better in terms of actual DCF. These results prove that if we have a high confidence on the segmentation of the test speech segment we can expect a good speaker verification performance. If we analyze, the number of trials of each subset, we observe that 85% of the trials have a confidence bigger than 0.33. This implies that trials with lower confidence do not have a big effect on the performance of the full trial list.

Table 2. Performance on SRE10 core-summed condition for different confidence intervals

Confidence range	0-1	0-0.33	0.33-0.67	0.67-1
EER(%)	3.38	4.35	3.59	2.25
min DCF(x10)	0.192	0.230	0.177	0.148
act DCF(x10)	0.193	0.238	0.188	0.174
Target trials	633	94	390	137
Non target trials	26487	3884	18041	4533

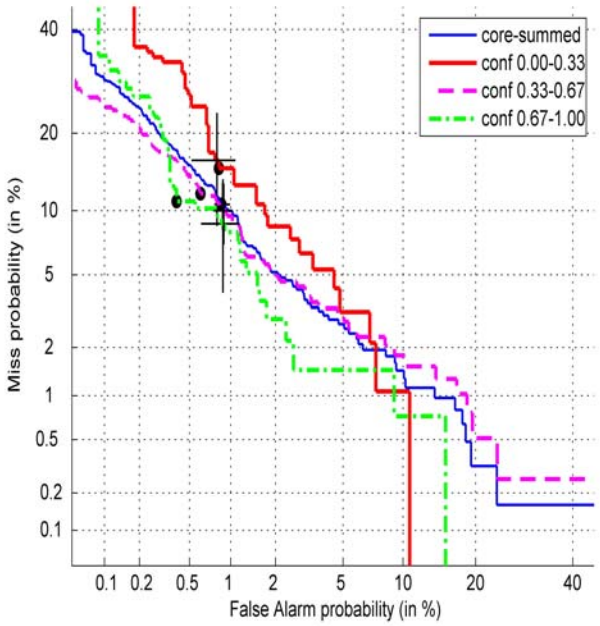

Fig. 3. DET plot for the SRE10 core-summed condition

5.3 Results 8summed-core

Figure 4 compares the DET curves of the NIST SRE10 8conv-core and 8summed-core det5 conditions. Minimum NIST detection costs are marked on the curves with a point. Table 3 presents the EER and cost values, and the number of trials of each condition. According to these results, having perfect segmentation and knowledge of the streams where the target speaker is present leads to an improvement of 28% in terms of EER and 37% in terms of DCF. However, we must take into account that we have achieved very low error rates for both conditions (under 1% of EER). With the number of trials available, this means that, in the EER operating point, we have only 3 absolute misses for both conditions. For the min DCF operating point, we have 6 misses and 32 false alarms on the 8conv-core condition, and 10 misses and 22 false alarms for 8summed core. Doddington 's "rule of 30" [Doddington, 2000], affirms that to be 90% confident that the true error rate is ±30% of the true error rate there needs to be at least 30 errors. Therefore, the degradation between 8conv and 8summed condition is inside the confidence range of the estimated error rates so we would need a much bigger number of trials for being able to measure it precisely. The fact that, in absolute terms, 8summed and 8conv performance is quite similar makes us think that we are very near of achieving the same results as with the perfect segmentation.

Table 3. Performance on SRE10 8conv-core and 8summed-core conditions

	8conv-core	8summed-core
EER(%)	0.67	0.93
min DCF(x10)	0.028	0.045
Target trials	442	322
Nontarget trials	21093	15010

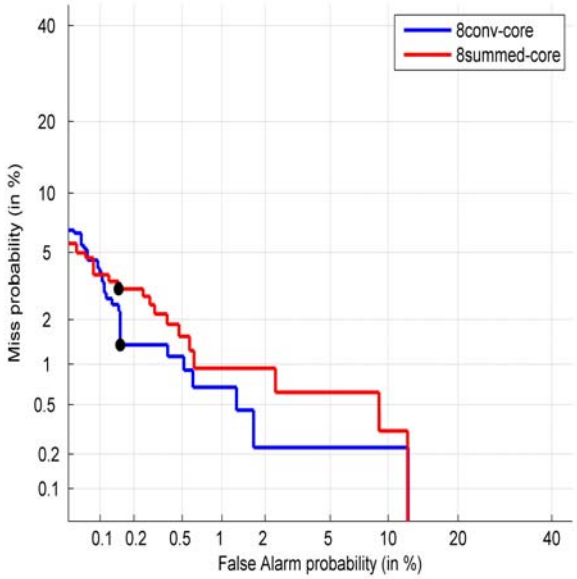

Fig. 4. DET plots for the SRE10 8conv-core and 8summed-core conditions

6 Conclusions

In this work, we have presented an eigenvoice based speaker segmentation system and a speaker verification system that, together, produce performances that are among the very best of the state-of-the-art systems on speaker verification tasks that involve summed channel segments in the enrollment or the testing sets. We have shown a set of confidence measures of the segmentation that can be fused together into a unique measure. We can use this measure to estimate the level of confidence that we can have on the speaker verification performance on a given test segment. This can be useful to apply back-off strategies on the segments with low segmentation confidences. These strategies include using other segmentation approaches on the segment or even human inspection. On the other hand, we have presented results on the NIST SRE10 8summed enrollment condition that proves that our system can produce a performance very near to the one we get having perfect segmentation. Besides, we think that

bigger trial lists should be needed to measure performance on 8summed and 8conv conditions precisely.

Acknowledgements

This work was supported by project TIN2008-06856-C05-04 and FPU program of MEC of the Spanish government.

References

1. **Bogert, B. P., Healy, M. J. R. & Tukey, J. W. (1963).** The quefrency alanysis of time series for echoes: Cepstrum, pseudo-autocovariance, cross-cepstrum and saphe cracking. *Symposium on Time Series Analysis*, New York, USA, 209–243.
2. **Burget, L., Fapso, M. Hubeika, V., Glembek, O., Karafiát, M., Kockmann, M., Matějka, P., Schwarz, P., & Černocký, J. (2009).** But system for nist 2008 speaker recognition evaluation. *Interspeech 2009.* Brighton, Great Britain, 2335–2338.
3. **Chen, S. S., & Gopinath, R. A. (2001).** Gaussianization. In Todd K. Leen, Thomas G. Dietterich,Volker Tresp (Eds.). *Advances in neural information processing systems 13*, (423-429), Massachusetts, USA, The MIT Press.
4. **Davis, S. & Mermelstein, P. (1980).** Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 28(4), 357-366.
5. **Dempster, A. P., Laird, N. M., & Rubin, D. B. (1977)**. Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society, Series B*, 39 (1), 1–38.
6. **Duda, R. O.** & **Hart, P. E. (1973).** *Pattern classification and scene analysis*. New York: Wiley.
7. **Furui, S. (1981).** Cepstral analysis techniques for automatic speaker verification. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 29 (2), 254–272.
8. **Gauvain, J. L. & Lee, C. H. (1994).** Maximum a posteriori estimation for multivariate Gaussian mixture observations of Markov chains. *IEEE Transactions on Speech and Audio Processing*, 2 (2), 291–298.
9. **Hermansky, H.**, **Morgan, N.**, **Bayya, A.**, **& Kohn, P. (1992).** RASTA-PLP speech analysis technique. *IEEE International Conference on Acoustics, Speech and Signal Processing ICASSP-92*, San Francisco, USA, 1, 121–124.
10. **Lee, C.-H. (1997).** A unified statistical hypothesis testing approach to speaker verification and verbal information verification. *Proceedings COST,Workshop on Speech Technology in the Public Telephone Network: Where are we today?,* Rhodes, Greece, 63–72.
11. **Marcel, S., McCool, C., Matejka, P., Ahonen, T., Cernocky, J. (2010).** *Mobile biometry (mobio) face and speaker verification evaluation.* Retrieved from http://publications.idiap.ch/index.php/publications/show/1848
12. **Mariéthoz, J. & Bengio, S. (2005).** A unified framework for score normalization techniques applied to text-independent speaker verification. *IEEE Signal Processing Letters*, 12 (7), 532-535.
13. **Martin, A.F**. **& Greenberg, C.S. (2009).** NIST 2008 Speaker Recognition Evaluation: Performance across Telephone and Room Microphone Channels. *Interspeech 2009*, Brighton, United Kingdom, 2579-2582.
14. **McCool, C. & Marcel, S. (2010).** *Mobio database for the ICPR 2010 face and speech competition*. Retrieved from http://publications.idiap.ch/index.php/publications/show/1757
15. **Navratil, J. & Ramaswamy, G.N. (2003).** The awe and mystery of t-norm. *8th European Conference on Speech Communication and Technology*, Geneva, Switzerland, 2009-2012.
16. **Pelecanos, J. & Sridharan, S. (2001).** Feature warping for robust speaker verification. *A Speaker Odyssey-The Speaker Recognition Workshop*, Crete, Greece, 213-218.
17. **Petrovska-Delacrétaz, D.**, **Hannani, A. E.**, **& Chollet, G. (2007).** Text-independent speaker verification: state of the art and challenges. *Progress in nonlinear speech processing*. *Lecture Notes in Computer Science*, 4391, 135–169.
18. **Reynolds, D.A. (1992).** *A Gaussian mixture modeling approach to text-independent speaker identification*. Ph.D. dissertation, Georgia Institute of Technology, Atlanta, Georgia, USA.
19. **Reynolds, D.A. (1995)**, Speaker identification and verification using Gaussian mixture speaker models. *Speech Communication*, 17 (1-2), 91–108.
20. **Reynolds, D.A.**, **Quatieri, T. F. & Dunn, R. B. (2000).** Speaker verification using adapted Gaussian mixture models. *Digital Signal Processing*, 10 (1-3), 19–41.
21. Speaker Recognition Evaluation. Retrieved from http://www.itl.nist.gov/iad/mig/tests/sre/
22. *Speech processing, Transmission and Quality aspects (STQ); Distributed speech recognition; Front-end feature extraction algorithm; Compression algorithms*. ETSI ES 201 108 V1.1.2 (2000-04), 2000.
23. **Viikki, O. & Laurila, K. (1998).** Cepstral domain segmental feature vector normalization for noise robust speech recognition. *Speech Communication-Special issue on robust speech recognition*, 25 (1-3), 133–147.
24. **Wald, A. (1947).** *Sequential analysis*. New York: John Wiley and Sons.

Carlos Vaquero Avilés Casco

He received the M.Sc. in Telecommunication Engineering from University of Zaragoza (Spain) in 2006. Since he graduated he has been researching on speech and speaker recognition in the Group of Speech Technologies of the Aragon Institute for Engineering Research (I3A). Recently he joined Agnitio S.L. a Spanish company that provides speech recognition and voice biometrics solutions. He is currently is doing his Ph.D. on speaker diarization for speaker characterization. He has done research internships in SRI International and ICSI in California. His current interests are speaker diarization, clustering and recognition.

Jesús Villalba López

He received the M.Sc. in Telecommunication Engineering from University of Zaragoza (Spain) in 2004. Since he graduated he has been researching on speech and speaker recognition in the Group of Speech Technologies of the Aragon Institute for Engineering Research (I3A). He is doing his Ph.D. about robustness on speaker verification systems. He has leaded the I3A submissions to the NIST speaker recognition evaluations since 2006 achieving competitive results. He has done research internships in Brno University of Technology (BUT) and Agnitio Labs in South Africa. His current interests are speech quality measures, attacks to speaker verification systems and i-vectors based speaker verification.

Alfonso Ortega Giménez

He was born in Teruel, Spain. He received the Telecommunication Engineering and the Ph. D. degrees from the University of Zaragoza in 2000 and 2005, respectively. His Ph. D. Thesis, advised by Dr. Eduardo Lleida, received the PhD Extraordinary Award and the Telefónica Chair Award to the best technological Ph. D. He has participated in more than 20 research projects funded by national or international public institutions and more than 15 research projects for private companies. He is author of more than 40 papers published in international journals or conference proceedings and international patents. He is presently Associate Professor in the Department of Electronic Engineering and Communications in the University of Zaragoza. His research interests span the areas of digital speech processing, analysis and modeling of speech and speaker, and robust automatic speech recognition.

Eduardo Lleida Solano

He was born in Spain in 1961. He received the M.Sc. degree in telecommunication engineering and the Ph.D. degree in signal processing from the Universitat Politècnica de Catalunya (UPC), Spain, in 1985 and 1990, respectively. From 1986 to 1988, he was involved in his doctoral work at the Department of Signal Theory and Communications at the Universitat Politècnica de Catalunya, Spain. From 1989 to 1990 he worked as assistant professor and from 1991 to 1993, he worked as associated professor in the Department of Signal Theory and Communications at the Universitat Politècnica de Catalunya, Spain. From February 1995 to January 1996, he was with AT&T Bell Laboratories, Murray Hill, NJ as a consultant in Speech Recognition. Currently, he is a full professor of signal theory and communications in the Department of Electronic Engineering and Communications at the University of Zaragoza (Spain), where he is heading a research team in speech

recognition and signal processing. He has more than 50 speech-related projects in Spain. He has co-authored more than 150 technical papers in the field of speech and speaker recognition, speech enhancement and recognition in adverse acoustic environments, acoustic modeling, confidence measures, and spoken dialogue systems.