Quality Assessment for Speaker Diarization and Its Application in Speaker Characterization

Carlos Vaquero, Alfonso Ortega, Antonio Miguel, and Eduardo Lleida, *Member, IEEE*

***Abstract*—There are many applications related to speaker characterization, specially in telephone environments, where large datasets are available but not directly useful since there are two speakers involved in every recording. Even with very accurate speaker diarization systems, we can expect to find some recordings with low diarization accuracy. The use of these recordings may reduce the accuracy of any speaker characterization technology. Therefore, it is highly desirable to detect those recordings where the speakers are correctly segmented, in order to discard or process manually the remaining ones before feeding them into the application. In this work we propose a set of confidence measures to assess the quality of a hypothetical diarization output, in order to detect those recordings that are correctly segmented. We show that these confidence measures enable us to retrieve most of the desired recordings from a given dataset, discarding those recordings that degrade the overall accuracy of an application that make use of speaker characterization technologies.**

***Index Terms*—Speaker diarization, confidence measures, speaker characterization, telephone conversations.**

I. Introduction

SPEAKER Characterization is the task of describing the particular and distinctive peculiarities of a person's speech. Speaker Characterization technologies have evolved significantly in the last decade, in part motivated by the NIST Speaker Recognition Evaluations (SRE) [1]. These evaluations focus on security and forensic applications, but there are many other applications of these technologies, including automatic speaker clustering for dataset partitioning, or speaker identification for improved Automatic Speech Recognition (ASR). Most of these applications work reasonably well as far as the operating conditions are close to those considered for development, so usually the development datasets are selected or even collected to match the expected conditions.

However, in the environment of telephone conversations, there are several situations where, even when it is possible to collect data in the desired conditions, both sides of the conversation are present in the available recording. In these situations, the recordings obtained will be almost useless for the purpose of speaker characterization unless a reliable speaker diarization hypothesis is available.

In this work we address the problem of retrieving a subset as representative and reliable as possible for speaker characterization applications from a given dataset composed of two-speaker telephone conversations. The reliability of the subset is related to the accuracy of the diarization hypotheses of the recordings in the subset, given an application. The representativeness of the subset is related to the fraction of information that the subset keeps from the given dataset. For this purpose, two complementary objectives must be pursued. Firstly, the highest speaker diarization accuracy is desired, in order to obtain a high number of correctly segmented recordings in the dataset. Secondly, quality assessment is mandatory, in order to detect as many correctly segmented recordings as possible within the available dataset, avoiding the use of diarization hypotheses with high error.

Accurate speaker diarization systems have been an aim for the research community in the recent years, in part motivated by the NIST Rich Transcription (RT) Evaluations, that evaluate Speaker Diarization Systems in meeting or broadcast news environments. Currently, the best performing systems proposed in these evaluations are mostly based on Agglomerative Hierarchical Clustering (AHC) algorithms using Bayesian Information Criterion (BIC) [2] as distance metric [3].

On the other hand, the recent advances in the field of speaker recognition have motivated new approaches for speaker diarization, developed specifically for two-speaker telephone conversations. Most of them are based on the use of Factor Analysis (FA) models [4] and include the use of eigenvoice modeling [5], [6], soft-clustering [7], or Variational Bayes methods [8]. All these approaches have shown better performance than the traditional BIC based AHC, but only in two-speaker telephone conversations.

However, there is little work in the field of assessing the quality of diarization hypotheses. In [6], a small set of confidence measures is proposed aiming at predicting the performance of a speaker diarization system for telephone conversations. A degree of correlation between the confidence measures and the performance of a speaker diarization system, and also between the confidence measures and the performance of a speaker verification system that faces two-speaker conversations is shown. In [9], it is shown that these confidence measures enable us to improve the performance when used to select the best diarization hypothesis among several hypotheses available for every recording.

In this work, a novel methodology for quality assessment of speaker diarization hypotheses, including an expanded set of confidence measures is presented. It is shown that the presented

Manuscript received May 03, 2012; revised August 04, 2012 and December 03, 2012; accepted December 08, 2012. Date of publication December 24, 2012; date of current version January 18, 2013. This work was supported in part by the MICINN of the Spanish government under Project TIN2011-28169-C05-02 and the FPU program. The associate editor coordinating the review of this manuscript and approving it for publication was Prof. Haizhou Li.

C. Vaquero was with the Aragon Institute for Engineering Research (I3A), University of Zaragoza, 50018 Zaragoza, Spain. He is now with Agnitio S.L., 28013 Madrid, Spain (e-mail: cvaquero@agnitio-corp.com).

A. Ortega, A. Miguel, E. Lleida are with the Aragon Institute for Engineering Research (I3A), University of Zaragoza, 50018 Zaragoza, Spain (e-mail: ortega@unizar.es; amiguel@unizar.es, lleida@unizar.es).

Digital Object Identifier 10.1109/TASL.2012.2236317

1558-7916/$31.00 © 2012 IEEE

confidence measures are suitable for detecting reliable diarization hypotheses, obtaining a blind segregation of useful recordings from a given dataset.

This paper is organized as follows: In Section II we present a figure of merit related to the usefulness of a given dataset for speaker characterization purposes, while in Section III our approach for speaker diarization is described. In Section IV a set of confidence measures and a method to assess the quality of an hypothetical diarization is presented, while in Section V the confidence measures and the proposed method for quality assessment are validated. In Section VI we present a strategy for diarization hypothesis generation and selection that combined with the quality assessment method presented in this work enables us to increase the diarization accuracy. Finally, in Section VII the conclusions of this work are presented, and future work is proposed.

II. Dataset Usefulness for Speaker Characterization

Given a dataset Ω composed of two-speaker telephone conversations, our objective is to extract a subset $\Omega' \subset \Omega$ as reliable and representative as possible for a speaker characterization application. The concept of reliability and representativeness for the defined task are explained next, and a figure of merit that involves both concepts is also introduced.

A. Reliability of a Subset

The reliability of a recording is related to the accuracy of its diarization hypothesis. Thus, to measure the reliability, we need a measure of the accuracy of a diarization hypothesis. As accuracy measure for speaker diarization, we use the Diarization Error Rate (DER), a well known metric used in the NIST RT evaluations to measure the percentage of the time to evaluate that has been incorrectly assigned by a speaker diarization system. It comprises four error types:

- Speaker Error: it accounts for the speech time that has been assigned to an incorrect speaker over the total time to evaluate.
- False Alarm Speech: it accounts for the non-speech time that has been labeled as speech over the total time to evaluate.
- Missed Speech: it accounts for the speech time that has been labeled as non-speech over the total time to evaluate.
- Overlapped Speech Error: it is related to the errors in the speaker detection and assignment in overlapped speech segments. The errors in an overlapped speech segment always fall in one of the three previous categories (Speaker Error when the speakers detected in the overlapped segment are incorrect, False Alarm Speech when the system detects more speakers than the actual number of overlapped speakers in the segment, and Missed Speech when the system detects fewer speakers than the actual number of speakers in the segment).

Nevertheless, the reliability of a recording does not only depend on the accuracy of its diarization hypothesis, but also on the application that will use the recording. Depending on the application, it may be useful to consider diarization hypotheses with little error. In general, we consider that we can define an application dependent threshold ϵ, so that every diarization hypothesis obtaining a DER below the threshold will probably be reliable and useful for the application. In addition and independently of the selected threshold, the lower the DER the better the diarization hypothesis for our application.

Therefore, we can measure the reliability of a recording n for a given diarization hypothesis as:

$$L(n) = \frac{\epsilon - DER(n)}{\epsilon}, \tag{1}$$

where $\mathrm{DER}(n)$ is the DER obtained for the recording n. The term in the numerator is the distance between the DER obtained for n and the application dependent threshold, and will be higher as better is the diarization hypothesis. Note that $L(n) > 0$ if $\mathrm{DER}(n) < \epsilon$, so correct diarization hypotheses obtain a positive reliability while incorrect diarization hypotheses obtain negative reliability. The denominator is just a normalization term, so that the maximum value of the reliability is 1. A recording n will be completely reliable $L(n) = 1$ only if $\mathrm{DER}(n) = 0\%$.

This way, the reliability of a subset Ω' is defined as the mean of the reliabilities of the recordings in Ω':

$$L(\Omega') = \frac{\sum_{n \in \Omega'} L(n)}{|\Omega'|}, \tag{2}$$

where $|\Omega'|$ is the cardinality (number of recordings) of the subset Ω'. Therefore, the reliability is higher as the recordings in Ω' are better segmented. Note that $L(\Omega')$ can be negative if many recordings in Ω' are incorrectly segmented. Actually, the negative values in the reliability means that an incorrectly segmented recording can be "destructive" in the sense that the performance of the application can be degraded severely when considering such recordings.

B. Representativeness of a Subset

Nevertheless, increasing the reliability of a subset Ω' does not guarantee an increase in the accuracy of a given speaker characterization application. For example, note that the reliability of a subset Ω' containing only a single recording which is perfectly segmented $(\mathrm{DER} = 0\%)$ will be 100%, but this might not be the best development subset to train a speaker characterization system. We also need Ω' to be as representative of Ω as possible.

Representativeness has to do with the fraction of information of the dataset that is kept in Ω', being Ω' fully representative of Ω only if $\Omega' = \Omega$. Therefore, the definition of representativeness will depend on the application. In general, we can define the representativeness of Ω' as the fraction of total net speech available in Ω that is kept in Ω'. However, some applications may require to keep in Ω' as many speakers as possible or as many recordings as possible from the original dataset Ω (for example if every recording represents a different speaker, environment or condition).

In this work, we define the representativeness of $\Omega' \subseteq \Omega$ as:

$$P(\Omega') = \frac{|\Omega'|}{|\Omega|}, \tag{3}$$

where $|\Omega|$ is the cardinality of the dataset Ω.

We do not include the net speech in the definition since the techniques proposed in this work are evaluated on a database where the net speech in the recordings does not vary significantly. However, a more general definition of representativeness should take into account the net speech of the recordings in the dataset.

C. Dataset Usefulness for a Given Subset

As mentioned before, we need Ω' to be as reliable and representative of Ω as possible, so we define a figure of merit ξ, the Dataset Usefulness for a given subset Ω', $\xi(\Omega')$, that involves both the representativeness and reliability of Ω':

$$\xi(\Omega') = P(\Omega') \times L(\Omega') = \frac{\sum\limits_{n\in\Omega'} L(n)}{|\Omega|} = \frac{\sum\limits_{n\in\Omega'} \frac{\epsilon - \mathrm{DER}(n)}{\epsilon}}{|\Omega|} \tag{4}$$

The figure of merit $\xi(\Omega')$ increases as more representative and reliable, that is, as more useful is the subset Ω' for the application given the diarization hypotheses obtained. In this work we study techniques to obtain a subset Ω' that maximizes $\xi(\Omega')$.

There are two ways to increase $\xi(\Omega')$. First, $\xi(\Omega')$ can be increased obtaining diarization hypotheses with lower DER, that is, increasing the performance of the speaker diarization system. This way we increase the reliability of the whole dataset Ω and also the number of recordings that are correctly segmented in Ω. Note that $\xi(\Omega')$ always increases when adding correctly segmented recordings $(L(n) > 0)$ to Ω', so the higher the number of correctly segmented recordings the higher $\xi(\Omega')$. We consider a state-of-the-art speaker diarization system (see Section III) in order to obtain as many correctly diarized recordings as possible in Ω.

Secondly, $\xi(\Omega')$ can be increased selecting Ω' so that $\forall n \in \Omega',\ \mathrm{DER}(n) < \epsilon$, which implies detecting those recordings correctly segmented, and discarding those incorrectly segmented. Since the reliability is negative for incorrectly segmented recordings, such recordings decrease $\xi(\Omega')$, and removing them from Ω' will increase $\xi(\Omega')$. To detect those recordings correctly segmented we study several confidence measures and a method to assess the quality of the diarization output for every recording in Section IV.

III. Diarization System

A. Factor Analysis for Speaker Segmentation

The study and modeling of inter-speaker variability, that is, the variability present among different speakers, have shown to be very useful in the field of speaker recognition [4]. Consequently, in the last years, many approaches for speaker segmentation based on inter-speaker variability modeling have been proposed [5]–[8]. Most of these approaches build a factor analysis model using prior knowledge on inter-speaker variability to obtain a compact representation of a single speaker. This compact representation is usually a low dimension vector y, whose components are known as speaker factors. Such a representation has the advantage that, compact as it is, does not need much data to be estimated.

Most of the mentioned approaches share the way speakers are modeled. Let us assume that a set of feature vectors $X = \{x(1), x(2), \ldots x(F)\}$ (for example, Mel-Frequency Cepstral Coefficients, MFCC) of dimension D has been extracted from a recording or set of recordings that belongs to a single speaker s, and that a Universal Background Model (UBM), trained on a large and rich dataset (containing a wide variety of speakers), is available. The UBM is a Gaussian Mixture Model (GMM) of C components whose component mean vectors and covariance matrices can be represented with the pair $(M_{\mathrm{UBM}}, \Sigma_{\mathrm{UBM}})$, where M_{UBM} is the UBM GMM-supervector, obtained concatenating all its component means, and its associated covariance matrix Σ_{UBM} is a diagonal matrix whose diagonal blocks are the diagonal covariance matrices of the GMM components. Then, every speaker is modeled using a GMM whose means are adapted from the UBM, using an eigenvoice approach [4], [10], according to:

$$M_s = M_{\mathrm{UBM}} + V y_s, \tag{5}$$

where M_s is the speaker s GMM-supervector of dimension CD, y_s is a set of R speaker factors that represents the speaker s, and V is a $CD \times R$ low rank eigenvoice matrix that models inter-speaker variability, capturing those directions of maximum variability among different speakers. We try to describe this variability using a small set of variables, i.e., speaker factors y, that follow a Normal Standard distribution $N(y|0, I)$ a priori.

In order to train this factor analysis model, we need to obtain the parameters $\{M_{\mathrm{UBM}}, \Sigma_{\mathrm{UBM}}, V\}$. We usually estimate the pair $\{M_{\mathrm{UBM}}, \Sigma_{\mathrm{UBM}}\}$, training the UBM GMM using a large dataset and assuming that these estimates are good enough so we will not reestimate them. Once the UBM is obtained, the eigenvoice matrix V is trained using the factor analysis paradigm described in [4].

To perform speaker segmentation, given a recording that may contain speech from different speakers, we estimate the posterior distribution of $y(i)$ for small overlapped segments $i = 1, \ldots, T$ as in [6], according to the factor analysis model presented in [4]. Only one speaker is assumed to be active in every segment i, and that speaker will be represented as a point estimate given by the mean of the posterior of $y(i), m_{y(i)}$. Therefore, $m_{y(i)}$ will be a compact representation of the speaker present in every segment i. This way, the problem of speaker segmentation reduces to a clustering problem, where the speaker factors associated to the same speaker should be clustered together. Since we know that for a given speaker s the posterior distribution of y_s is normal, the problem of two-speaker segmentation reduces to finding the two Gaussian models that generated the obtained stream of speaker factors $Y = m_{y(1)}, \ldots, m_{y(T)}$.

B. System Architecture

We present a speaker diarization system for two-speaker telephone conversations. In order to perform speaker diarization given a recording containing two speakers, we follow the steps shown in Fig. 1 which are described below:

1) Front End: The first stage extracts a sequence of feature vectors from the input audio signal, and then it computes the sequence of speaker factors. The feature vectors considered are 19 MFCC including C0 plus Δ features, estimated over a 25

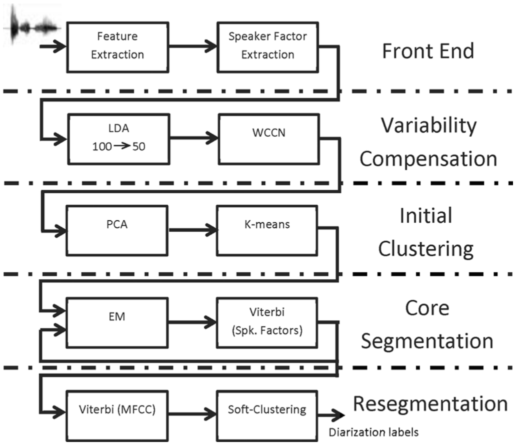

Fig. 1. Block diagram of the proposed segmentation system.

ms window every 10 ms. No compensation is considered for the MFCC features. To extract the speaker factors we consider a gender independent UBM of $C = 1024$ components trained on telephone speech from NIST SRE04, SRE05 and SRE06 databases. The eigenvoice matrix V of rank $R = 100$ is trained on the same databases, and for every input recording we estimate the speaker factors every 10 ms over a 1 sec. window (99% overlap). In our case, M_s will be a supervector of dimension $CD = 1024 \times 38$, and the speaker factor vectors obtained will have a dimension of $R = 100$.

2) Variability Compensation: It has been shown that the speaker factors for a single speaker extracted over short windows present intra-session variability [9]. To compensate for this variability we apply the techniques proposed in [11] for inter-session variability compensation in speaker verification: LDA and Within Class Covariance Normalization (WCCN) [12]. These techniques have shown to be useful for intra-session variability compensation in speaker diarization [9]. We consider LDA to reduce the dimensionality of the speaker factor vectors from 100 to 50, and then WCCN is applied on the 50-dimensional vectors.

3) Initial Clustering: The system applies PCA on the compensated speaker factors to obtain the best direction to separate the two speakers. Two clusters are obtained using uniquely this direction, and then refined using K-means on the fully dimension speaker factor vectors. This stage is based on the assumption that the covariance of the compensated speaker factor vectors for a single speaker will be close to the identity. There are two main reasons to believe this: First, the covariance matrix of the Gaussian prior for the speaker factor vectors is the identity matrix, and we estimate the posterior over small segments, so we do not expect the covariance matrix to change significantly in the posterior. Second, WCCN is applied as part of the variability compensation stage.

4) Core Segmentation: The clusters previously obtained serve to build an initial two-component GMM for the whole recording. Then a two stage iterative process is applied until convergence. First the two Gaussians that best fit the stream of compensated speaker factors are estimated by means of Expectation-Maximization (EM) iterations over the initial GMM. Each one of these Gaussians is assigned to a single speaker. Second, an HMM including models for the two speakers and a silence model is built. Every speaker is modeled with a left-to-right sequence of 10 tied-states, whose observation distribution is the speaker Gaussian model previously obtained. Transitions to/from a speaker model are only allowed through its first/last state. Transitions between two non-consecutive states within a speaker model are not allowed. The compensated speaker factor vectors are reassigned to the speakers using Viterbi decoding. Silence frames are modeled with a single-state, but no observation distribution is considered since the decoding process is forced to go through the silence state according to the speech/non-speech labels. We assume that the speech/non-speech labels are obtained previously using a VAD or speech/non-speech segmentation strategy.

5) Resegmentation: Since the compensated speaker factors enable easy speaker separation, the output of the core segmentation system gives accurate speaker labels in most cases, but these labels can be refined by means of Viterbi resegmentation. In this case we model every speaker using a left-to-right sequence of 10 tied-states sharing a 32-component GMM as observation distribution, with the same restrictions in the transitions as explained in Section III.B4. Again we use a single state for all silence frames. The features considered are 12 MFCC including C0 with no compensation and no Δ features. We have observed that this set of features obtain better accuracy in this resegmentation step than the complete set of 19 MFCC plus Δ features considered for speaker factor extraction. The use of MFCC for resegmentation is motivated by the fact that they provide a much higher temporal resolution than the speaker factors, since the MFCC features are estimated over windows of 25 ms, while the speaker factors are estimated over windows of 1 second length.

After this Viterbi resegmentation we retrain the GMM speaker models and run a soft-clustering pass [7].

IV. Assessing Diarization Quality

In the previous sections, we have presented a state-of-the-art speaker diarization system for two-speaker telephone conversations that is expected to obtain a very good diarization accuracy for most of the recordings of a given dataset Ω. However, this system or any other diarization system will be much more useful if for every diarization hypothesis given as output, a measure or set of measures to assess the quality of such segmentation could be provided, in order to retrieve those recordings correctly diarized. These measures are then confidence measures for speaker diarization, and enables us to assess the quality of a diarization hypothesis.

A. Confidence Measures

In this work we consider a set of four confidence measures. The first three confidence measures proposed are speaker separability indicators, that is, for a given input recording and its hypothetical segmentation, these measures are distance metrics between two speaker models built from the diarization hypothesis. We expect the distance between both models to be high if the segmentation is correct, i.e., the speaker models are different and pure. The last confidence measure presented is a well-conditioned data indicator that, for a given input recording and its hypothetical two-speaker segmentation, tries to determine whether or not the input data are well conditioned for our prior models and whether or not our system assumptions are fulfilled. We expect that well-conditioned recordings will be easier to diarize.

We have analyzed additional confidence measures that are not presented in this work, including the KullBack-Leibler (KL) divergence between the Gaussian speaker models in the speaker factor space and the number of iterations that the Core Segmentation stage of the proposed speaker diarization system (see Section III) needs to reach convergence. These two confidence measures were presented in [6], but they were not helpful in this case, since they do not seem to provide additional information over the four presented next.

1) Bayesian Information Criterion (BIC): Given two sequences of acoustic feature vectors X_1, X_2 segregated 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 (joint hypothesis). The confidence measure is the difference between both BIC values. To avoid adjusting BIC penalty parameters, we force the models for both hypotheses to have the same complexity. This way the confidence measure is defined as:

$$C_{\mathrm{BIC}} = \Delta\mathrm{BIC} = \log\left(\frac{\mathcal{L}(X_1|\theta_1)\mathcal{L}(X_2|\theta_2)}{\mathcal{L}(X_{1,2}|\theta_{1,2})}\right), \qquad (6)$$

where $\mathcal{L}$ denotes likelihood, and θ_s is the model obtained from every vector sequence X_s. We use 12 MFCC including C0 as feature vectors, and every speaker model is a 32-component GMM, while the global model is a 64-component GMM. This measure was presented in [6] as confidence measure for speaker segmentation showing good performance. We expect C_{BIC} to be higher for better diarization hypotheses.

2) i-Vector Based Speaker Verification Score: Given two segments, we want to determine whether every segment was uttered by a single speaker. Otherwise, at least one speaker will be present in both segments. Therefore, this problem is related to speaker verification, in the sense that we want to detect whether the same speaker is present in both segments, and a speaker verification system could give a measure of how probable are both segments to be uttered by the same speaker. Such measure can be used as confidence measure for this task. As speaker verification system we use an i-vector system [11]. In this case, we replace the i-vectors with the speaker factors obtained directly with the same eigenvoice matrix V used for segmentation, with $R = 100$, and we consider all available data to compute the speaker factor vector for every speaker. Intra-session variability is compensated using LDA + WCCN. This way the confidence measure is defined as:

$$C_{iv} = -score(\tilde{y_1}, \tilde{y_2}) = -\frac{\tilde{y_1}^t \tilde{y_2}}{\|\tilde{y_1}\|\|\tilde{y_2}\|}, \qquad (7)$$

where $\tilde{y_s}$ is the compensated speaker factor vector obtained for speaker s. Note that the minus sign has been added to ensure that C_{iv} increases as the hypothetical speakers are more separable.

3) Probabilistic LDA (PLDA) Based Speaker Verification Score: Recently, a new approach for speaker verification when working with i-vectors has been proposed in [13]. Such approach is similar to that presented in [11] but instead of using LDA to find the subspace that best discriminate among different speakers, it uses PLDA to model such subspace, and optionally an inter-session variability subspace. The main advantage of this approach against traditional LDA is that, since it is probabilistic, it provides a way to compute the likelihood of a speaker model, and thus the score can be computed in a similar way to BIC, as a likelihood ratio defined by:

$$C_{\mathrm{PLDA}} = -\log\frac{\mathcal{L}_{\mathrm{PLDA}}(w_1, w_2)}{(\mathcal{L}_{\mathrm{PLDA}}(w_1)\mathcal{L}_{\mathrm{PLDA}}(w_2))}, \qquad (8)$$

where $\mathcal{L}_{\mathrm{PLDA}}$ is the likelihood computed on the PLDA model, and w_s is the i-vector obtained for speaker s. Again, the minus sign has been added to ensure that C_{PLDA} is higher as the hypothetical speakers are more separable.

To compute C_{PLDA} we consider i-vectors of dimension 400 extracted using the same UBM and the same features considered in the speaker diarization system in Section III.B1. The total variability matrix is trained on the same databases as the UBM. The PLDA model considered has a speaker subspace with dimension 100 and a full rank channel covariance matrix.

Note that this confidence measure is very similar to C_{iv}. The main motivation of considering both C_{iv} and C_{PLDA} is that each measure compensates for a different type of variability. When computing C_{iv} only intra-session variability is compensated using LDA + WCCN. Thus it works under the assumption that the channel observed for each speaker will remain constant in the complete session. For the computation of C_{PLDA} we consider a PLDA model trained to compensate for inter-session variability, which comprises mainly channel variability. Therefore, C_{PLDA} should be able to detect diarization errors due to channel variations for a single speaker in a recording (for example, if the relative position between the speaker and the microphone varies).

4) Normalized Eigenvalue Spread of the Speaker Factors: The proposed speaker diarization system works under the assumption that the speaker factor vectors for a given speaker follow a Gaussian distribution whose covariance is close to the identity. This assumption is critical for the initial clustering performed by the diarization system, since it makes use of PCA and K-means. If this assumption is not fulfilled, we can expect certain degradation in the accuracy of the speaker diarization system. So an indicator of how close our speaker models are to fulfill this assumption is a good candidate for confidence measure. We propose a normalized eigenvalue spread estimation defined as:

$$C_{\mathrm{spread}} = \log\frac{\frac{\max(\lambda_{1,2})}{\mathrm{median}(\lambda_{1,2})}}{\frac{\max(\lambda_1)}{\mathrm{median}(\lambda_1)}\frac{max(\lambda_2)}{\mathrm{median}(\lambda_2)}} \qquad (9)$$

where λ_s are the eigenvalues of the covariance matrix of speaker s estimated using $\tilde{y_s}$. In all eigenvalue spread computation the median of the eigenvalues have been used in the denominator rather than the minimum, since the minimum may be noisier. The term in the numerator is the eigenvalue spread considering the speaker factors from both speakers (joint hypothesis), and should increase as the speakers are more separable, while the term in the denominator is the product of the eigenvalue spread for every speaker and should be close to one if the mentioned assumption is fulfilled.

B. Detecting Correctly Segmented Recordings

Obtaining a reliable confidence measure can be useful to predict the performance of the segmentation system for a segmentation hypothesis, so that a given application can decide how to deal with the current recording. This usually means that the confidence measure is compared to a threshold or set of thresholds in order to classify the recording into different classes that will be processed differently. Then, given a dataset Ω, a partition of Ω is created according to the predicted accuracy of the diarization system so that the application can deal properly with every class in the partition. Therefore, the partition will be application dependent. For example, a semi-supervised segmentation system can be built, so that the user only needs to check the segmentation hypotheses for a small subset of the whole dataset. This subset will be composed of those recordings that the diarization system labels as unreliable.

In this study we limit our problem to two classes, assuming that our application will use only the subset Ω_c of correctly segmented recordings and discard the rest of the dataset (Ω_i). This way, we want to solve the detection problem where only those segmentation outputs obtaining an accuracy below the threshold ϵ are desired.

In order to detect those recordings correctly segmented, we can train a linear logistic regression [14] model fusing the four proposed confidence measures. This model will give as output the certainty of a hypothetical segmentation for a given recording to be correct. This certainty can be considered as a fused confidence measure C_{fused} which is defined as follows:

$$C_{\text{fused}} = \frac{1}{e^{-(\boldsymbol{\alpha}^t \boldsymbol{C} + \beta)} + 1} \tag{10}$$

$$\boldsymbol{\alpha} = \begin{bmatrix} \alpha_{\text{BIC}} \\ \alpha_{iv} \\ \alpha_{\text{PLDA}} \\ \alpha_{\text{spread}} \end{bmatrix}, \boldsymbol{C} = \begin{bmatrix} C_{\text{BIC}} \\ C_{iv} \\ C_{\text{PLDA}} \\ C_{\text{spread}} \end{bmatrix} \tag{11}$$

where $\boldsymbol{\alpha}$ is the vector of fusion weights, and β is an offset so that the term $\boldsymbol{\alpha}^t\boldsymbol{C} + \beta$ can be interpreted as a well-calibrated log-likelihood ratio in order to make the decision.

Given a development dataset, the weights $\boldsymbol{\alpha}$ and the offset β are trained to minimize a Detection Cost Function (DCF) usually defined as:

$$\text{cost}(c_m, c_{fa}) = c_m p_m p_c + c_{fa} p_{fa} (1 - p_c) \tag{12}$$

where c_m and c_{fa} are the cost of a false negative or a miss and the cost of a false positive or a false alarm respectively, p_m and p_{fa} are the rate of miss and false alarm respectively, and p_c is the probability of segmenting a recording correctly, or the prior rate of correctly segmented recordings. Note that p_c can be extracted from a dataset and it is identical to the representativeness of the correct subset: $p_c = P(\Omega_c)$.

The main problem of a linear logistic regression model is that it considers that all recordings are equally important for the detection task. That is, the logistic regression model does not take into account the reliability of every recording, only whether its DER is over or below ϵ. To overcome this problem we can use a weighted linear logistic regression model, which is a logistic regression model where every recording has a weight or an importance within the training algorithm. The weight for every recording is related to the cost of misclassifying that particular recording, and thus to the reliability of the recording. Therefore, the weights for the correctly (w_c) and incorrectly (w_i) segmented recordings are defined as follows:

$$w_c(n) = \frac{c_m L(n)}{P(\Omega_c) c_m L(\Omega_c) - P(\Omega_i) c_{fa} L(\Omega_i)} \tag{13}$$

$$w_i(n) = \frac{c_{fa} L(n)}{P(\Omega_c) c_m L(\Omega_c) - P(\Omega_i) c_{fa} L(\Omega_i)} \tag{14}$$

where $P()$ denotes representativeness as defined in (3), and $L()$ denotes reliability as defined in (2). The minus sign in the denominator is used since $L(\Omega_i)$ will be negative (see Section II). This way, every recording is weighted by the absolute value of the reliability of its segmentation hypothesis, which is the distance between the DER obtained for that hypothesis and the application dependent threshold ϵ. The use of these weight definitions for training the logistic regression minimizes the following cost function:

$$\begin{aligned}\text{cost}(c_m, c_{fa}) = P(\Omega_c) c_m L(\Omega_m) p_m \\ - P(\Omega_i) c_{fa} L(\Omega_{fa}) p_{fa},\end{aligned} \tag{15}$$

where Ω_m and Ω_{fa} are the subsets composed by the missed and false alarms recordings respectively. Note that this expression is analogous to that in (12) and also to that used in the NIST SRE [1] to define the DCF, but in this case the prior probability of target and non-target trials are $P(\Omega_c)$ and $P(\Omega_i)$ respectively, and the cost of every miss and false alarm depends on every recording and it is given by its reliability.

In the definition of our objective figure of merit $\xi(\Omega')$ we have considered that $c_m = c_{fa} = 1$. Taking this into account, and realizing that $P(\Omega_c)L(\Omega_m)p_m = \xi(\Omega_m)$ and $P(\Omega_i)L(\Omega_{fa})p_{fa} = \xi(\Omega_{fa})$, the cost can be defined as:

$$cost(c_m = 1, c_{fa} = 1) = \xi(\Omega_m) - \xi(\Omega_{fa}), \tag{16}$$

It can be seen that minimizing this cost is equivalent to detecting the subset Ω' that maximizes the Dataset Usefulness $\xi(\Omega')$, since $\xi(\Omega')$ can be expressed as:

$$\begin{aligned}\xi(\Omega') &= \xi(\Omega_c) - \xi(\Omega_m) + \xi(\Omega_{fa}) \\ &= \xi(\Omega_c) - cost(c_m = 1, c_{fa} = 1).\end{aligned} \tag{17}$$

V. Validation of Quality Assessment

A. Experimental Setup

To evaluate the described approach for assessing speaker diarization quality, we consider the NIST SRE08 summed channel test condition, a well known set of 2213 five minute telephone conversations. This dataset has been used to report speaker diarization accuracy in [6], [7] and [8]. In addition, to validate the logistic regression model, which was developed on the previous dataset, we consider a different dataset composed of a total of 7044 summed channel telephone conversations extracted from the NIST SRE10.

Results are presented in terms of the overall DER and the percentage of recordings (Representativeness) in the subset Ω' considered and also in terms of the usefulness for the subset Ω' measured using the figure of merit ξ defined. We consider that the speaker characterization application that makes use of the output of the diarization system in general operates correctly when $\mathrm{DER} < 10\%$, so $\epsilon = 10\%$. This ϵ value has shown to be valid for a speaker verification application in [6] and [7]. However, depending on the application, other values for ϵ should be considered.

As reference diarization labels for DER computation, we consider the output labels of the Automatic Speech Recognition (ASR) system that the NIST runs on the recordings to provide a speech/non-speech segmentation to the participants of the SRE. Such labels are obtained on every side of every telephone conversation separately, providing a very good estimate of the real speaker segmentation labels in the summed conversation. For every conversation, we assume that the speech/non-speech segmentation is provided and correct, and we obtain such segmentation combining the ASR labels from both sides of the conversations. To mitigate the effect of possible errors in the reference labels, a collar of 0.25 seconds is considered in the DER computation, as it is usual in the evaluation of speaker diarization systems. Overlapped speech segments are not considered in DER computation. Therefore, the DER in this study is identical to the Speaker Error, or the rate of speech time that is incorrectly assigned to a speaker.

We also validate the quality assessment approach in a speaker verification task. We evaluate an extended short2-summed condition of the NIST SRE 2008 considering the proposed diarization system and quality assessment approach. The extended short2-summed condition evaluates all possible trials comparing a total of 1788 models to the 2213 summed recordings. Results are presented in terms of Equal Error Rate (EER) and the normalized minimum of the DCF (minDCF) as defined by NIST [1] in the SRE 2008.

B. *Performance of the Detection Task*

To analyze the performance of the weighted linear logistic regression for retrieving the correctly segmented recordings in a dataset, we first consider the NIST SRE08 summed channel condition as dataset Ω, and we obtain the segmentation hypotheses using the proposed speaker segmentation system. We train the logistic regression on the same dataset Ω, considering that we need to retrieve recordings whose DER is below $\epsilon = 10\%$.

Table I shows the accuracy of the diarization system and the detection task on the NIST SRE08 dataset. If we consider the whole dataset Ω, we obtain an overall DER of 1.31% and a Dataset Usefulness ξ of 85.63%. Analyzing the detection task, we can see that 96.75% (2141 out of 2213) of the recordings are correctly detected as true positives ($\Omega_{true\ positives}$), and the overall DER and the standard deviation of the DER for such recordings are as low as 0.75% and 1.30% respectively. Only 0.63% (14 out of 2213) of the recordings are false alarms or false positives, which is a low number compared to the number of correctly segmented recordings detected. The most important result is the Dataset Usefulness for the selected subset $\Omega' = \{\Omega_{true\,positives} \cup \Omega_{false\,positives}\}$, $\xi(\Omega')$. We can see that $\xi(\Omega')$ is over 2.5% higher than $\xi(\Omega)$, which means that the subset Ω' of detected recordings is more useful for the application than the whole dataset Ω, since it is much more reliable. This can be noted also in the DER of the subset Ω', 0.86%, which is lower than that obtained for the whole dataset. Actually, the DER of the complementary subset $\Omega\backslash\{\Omega'\}$ is very high (19.96%), since we are discarding many incorrectly segmented recordings, and $\xi(\Omega\backslash\{\Omega'\})$ is negative, which means that it is better to discard the subset for the application.

TABLE I
PERFORMANCE OF THE DETECTION TASK USING WEIGHTED LINEAR LOGISTIC REGRESSION ON THE NIST SRE 2008

Subset	DER	σ_{DER}	Represent.	ξ
Ω	1.31%	4.58%	100.00%	85.63%
Ω_c	0.77%	1.34%	97.51%	89.55%
Ω_i	25.36%	12.93%	2.49%	-3.92%
$\Omega_{true\,positives}$	0.75%	1.30%	96.75%	89.05%
$\Omega_{true\,negatives}$	26.78%	12.91%	1.85%	-3.18%
$\Omega_{false\,positives}$	20.84%	12.57%	0.63%	-0.73%
$\Omega_{false\,negatives}$	3.67%	2.73%	0.77%	0.50%
Ω'	**0.86%**	**2.34%**	**97.38%**	**88.31%**
$\Omega\backslash\{\Omega'\}$	19.96%	15.38%	2.62%	-2.69%

The high DER value obtained for the complementary subset $\Omega\backslash\{\Omega'\}$ shows that there is a small set of recordings very poorly diarized. A deeper analysis on the causes of the poor accuracy for these recordings shows that there is a wide variety of causes that may produce a poor diarization: the presence of high overlapped speech rate, highly unbalanced speakers (one dominant speaker), conversations composed of very short turns and, in some cases, the language can be an issue since the diarization system is trained mostly on English recordings.

The accuracy of the quality assessment approach has also been tested considering the reference diarization labels as segmentation hypotheses. In this case, the DER is 0% for the whole dataset Ω, and thus all the recordings should be detected as correctly segmented. Only 10 out of 2213 recordings are classified as incorrectly segmented, and most of the dataset would be considered for the application in this case.

In order to check how the improvement in the Dataset Usefulness is reflected in a speaker characterization application we run a speaker verification experiment on the extended short2-summed condition of the NIST SRE 2008 considering the proposed diarization system and quality assessment approach. Fig. 2 shows the DET curves obtained for the extended short2-summed condition considering the reference diarization labels and the proposed diarization system. For the proposed diarization system, the curves are presented for the datasets Ω, Ω' and $\Omega\backslash\{\Omega'\}$ obtained considering the quality assessment approach. The DET curve obtained for the selected dataset Ω' is not significantly better than that obtained for the whole dataset Ω, since both datasets are quite similar. However, it is interesting to observe the DET curve obtained for the discarded subset $\Omega\backslash\{\Omega'\}$: The discarded subset is actually obtaining a very poor accuracy in the speaker verification task, so the quality assessment approach is working as expected.

The same effect can be observed in Table II, where the EER and the normalized minDCF are presented. Note that there are

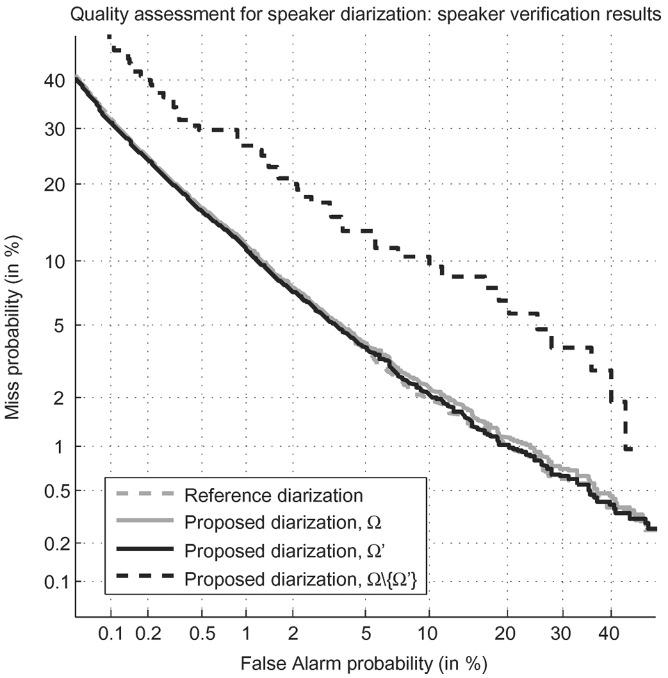

Fig. 2. DET curves for the extended short2-summed condition considering the reference diarization labels and the proposed diarization system. For the proposed diarization system, the curves are presented for the datasets Ω, Ω' and $\Omega\backslash\{\Omega'\}$.

TABLE II
EER AND THE NORMALIZED MINDCF FOR THE EXTENDED SHORT2-SUMMED CONDITION CONSIDERING THE REFERENCE DIARIZATION LABELS AND THE PROPOSED DIARIZATION SYSTEM. FOR THE PROPOSED DIARIZATION SYSTEM, THE RESULTS ARE PRESENTED FOR THE DATASETS Ω, Ω' AND $\Omega\backslash\{\Omega'\}$

Subset	EER	minDCF norm
Reference	4.23%	0.2102
Ω	4.39%	0.2097
Ω'	**4.26%**	**0.2056**
$\Omega\backslash\{\Omega'\}$	9.44%	0.3429

not significant differences among the results obtained considering the reference labels, the proposed system for Ω and for Ω'. However, the results obtained for $\Omega\backslash\{\Omega'\}$ are quite poor and might not be acceptable depending on the application.

Note that the improvement observed when considering the selected subset Ω' instead of Ω is not significant, due to the fact that both datasets are almost identical, since most recordings are correctly diarized. Assuming that a dataset with much poorer diarization accuracy is available, the size of Ω' will be reduced and we expect a significantly better accuracy in a speaker verification task for the subset Ω' than for Ω. In fact, the proposed quality assessment approach has been considered for a speaker clustering task using the same dataset (NIST SRE 2008 summed dataset) in [15]. Since the DER threshold considered in that work is smaller (1%), the number of recordings accepted as reliable is also smaller and the accuracy of the clustering task for the subset Ω' is significantly higher than for the dataset Ω.

We have observed that the proposed quality assessment approach selects most of those recordings correctly diarized and it is potentially capable of increasing the accuracy of a speaker characterization application. However, we need to validate it on a different dataset. To validate the quality assessment approach, we run the detection task on the NIST SRE10 dataset. In this case, Ω is composed of 7044 conversations not considered during the development of the weighted linear logistic regression model.

TABLE III
PERFORMANCE OF THE DETECTION TASK USING WEIGHTED LINEAR LOGISTIC REGRESSION ON THE NIST SRE 2010

Subset	DER	σ_{DER}	Represent.	ξ
Ω	2.45%	4.87%	100.00%	72.79%
Ω_c	1.76%	1.84%	95.83%	77.57%
Ω_i	21.72%	11.15%	4.17%	-4.78%
$\Omega_{true\,positives}$	1.75%	1.81%	95.34%	77.35%
$\Omega_{true\,negatives}$	29.37%	10.94%	1.46%	-2.82%
$\Omega_{false\,positives}$	17.04%	8.75%	2.71%	-1.96%
$\Omega_{false\,negatives}$	5.33%	3.02%	0.48%	0.21%
Ω'	**2.08%**	**3.41%**	**98.06%**	**75.39%**
$\Omega\backslash\{\Omega'\}$	23.48%	14.07%	1.94%	-2.60%

Table III shows the performance of the detection task on the NIST SRE10 dataset. In this case, the detection task enables us to select a more reliable subset Ω' so that $\xi(\Omega')$ is 2.5% higher than that obtained for the whole dataset Ω. This increase is similar to that obtained for the NIST SRE08 dataset. Actually, the conclusions that can be extracted from Table III are, in general, those extracted for Table I.

Comparing results for both datasets, we can see that the accuracy of the diarization system is higher for NIST SRE08 than for NIST SRE10 (1.31% against 2.45%), and so the percentage of correctly segmented recordings and the Dataset Usefulness is higher for NIST SRE08 (97.51% against 95.83% and 85.63% against 72.79% respectively). This difference is mainly due to a lower accuracy of the diarization system on the NIST SRE10 dataset. Exploring the causes of this reduced accuracy, we have detected that the percentage of overlapped speech over the net speech is much higher for NIST SRE10 than for NIST SRE08, 12.7% against 5.8%. This high percentage of overlapped speech is one of the reasons for the degradation. Note that even though we do not consider overlapped speech for DER computation, it is considered as speech, and the diarization system will include the overlapped speech in the speaker models that it builds for segmentation purposes. The presence of a substantial amount of overlapped speech will reduce the purity of these speaker models and thus the accuracy of the segmentation, even in those segments where a single speaker is speaking (which are those evaluated for DER computation). Note also that the accuracy of the confidence measures can be affected due to the presence of a high rate of overlapped speech, since overlapped speech segments are also considered to compute them.

Since the logistic regression model has been trained and calibrated on the NIST SRE08, and the proportion of incorrectly segmented recordings is notably different in both datasets (2.49% against 4.17%), the logistic regression may not be optimal for NIST SRE10. The dependency of the $\xi(\Omega')$ on the calibration for both datasets is analyzed in Fig. 3. We can see that the optimum threshold for the output of the logistic regression is different for both datasets. However, $\xi(\Omega')$ does not change significantly for variations on this threshold, so this approach is robust against calibration errors. Particularly, for

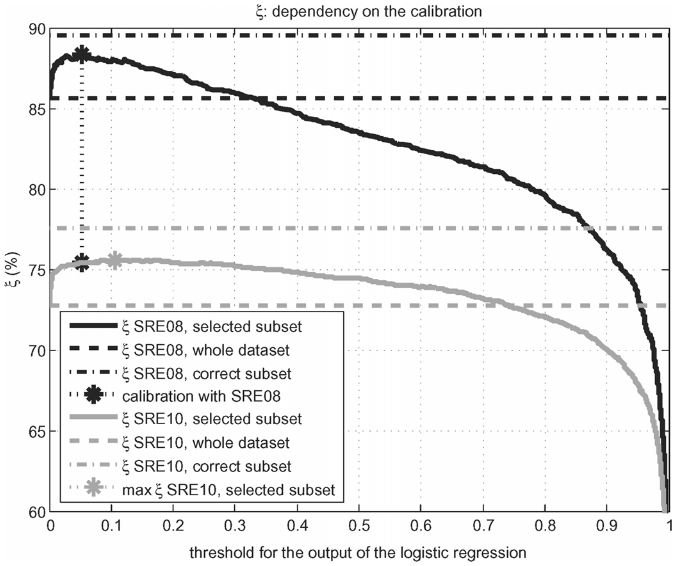

Fig. 3. Dependency of ξ on the calibration.

the NIST SRE10 dataset, $\xi(\Omega')$ is higher that $\xi(\Omega)$ for most of the threshold range.

We have presented results considering a threshold of 10% for the DER which has been shown to be suitable for speaker verification applications [7], [8]. However, depending on the speaker characterization application that will use the diarization hypotheses, other thresholds can be considered. In fact, in [15] it is shown that the proposed methodology for quality assessment is useful when considering a speaker clustering task, and that the clustering tasks operates with little degradation with those recordings obtaining a DER below 1%.

VI. Using Quality Assessment for the Generation and Selection of Diarization Hypotheses

In this section we present an approach that makes use of the proposed segmentation system, the confidence measures and the logistic regression modeling, in order to generate and select reliable segmentation hypotheses for a given recording, decreasing the DER and thus increasing ξ. In addition, the reliable subset of correctly segmented recordings is detected. The selected subset could be used, for example, for training speaker models in a speaker verification system from unlabeled data in an unsupervised or semi-supervised way, improving the accuracy of the system.

A. Hypothesis Generation and Selection

It is known that a good initialization is critical for the correct operation of most speaker diarization systems, including the proposed one [6]. Here, we propose a method to generate several segmentation hypotheses for a recording, using the confidence measures to select the best fragments to initialize the segmentation system, and also to select the final segmentation hypothesis.

In order to generate different segmentation hypotheses for a given conversation, we split the conversation into halves iteratively until we obtain a set of non-overlapping slices of sufficient length, as shown in Fig. 4. Thus, we obtain 2^{l-1} slices in every level l, and every slice is processed independently using the complete two-speaker segmentation system presented in this work. This way, for every conversation and level we obtain several slices with independent segmentation hypotheses, and we can expect some of them to be correct even if the segmentation hypothesis obtained for $l = 1$ (i.e., for the whole conversation) is not correct. The idea is to use those slices correctly segmented at every level as a good initialization to segment the whole conversation, obtaining a unique segmentation hypothesis for every level.

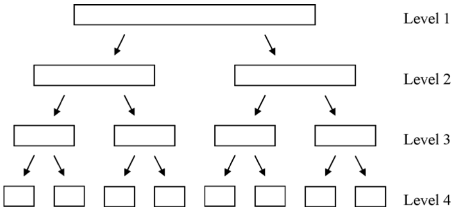

Fig. 4. Slice partition diagram.

To do so, for every recording, and for every level l, we select the best segmented slices at level l, we agglomerate the segmented speakers in every slice using BIC based AHC and modeling every speaker with a full covariance gaussian until we obtain two clusters from the selected slices, and the whole recording is resegmented using 32-component GMM speaker models trained on those clusters, as in [9]. A final soft-clustering resegmentation on the whole recording is done to refine the speaker boundaries. This process is done for every level l separately.

This way, we generate a segmentation hypothesis for every level l and then the best segmentation hypothesis will be selected as final segmentation for the recording. Of course, both to select the best slices for each recording at each level and later to select the best segmentation hypothesis for each recording, the proposed confidence measures and logistic regression model are used. For this purpose, the logistic regression model is trained as explained in Section IV, so that for every segmentation hypothesis and its set of confidence measures, a fused confidence measure C_{fused} is obtained as output of the regression model. The selected segmentation hypotheses will be those with maximum fused confidence measure.

B. Experimental Setup

To evaluate our approach for hypothesis generation and selection, we use NIST SRE08 and NIST SRE10 summed channel datasets. We consider that the NIST SRE08 dataset is available and labeled to train our logistic regression model, our approach will be validated on the NIST SRE10 summed channel dataset. Again we consider that speech/non-speech segmentation is given by the ASR transcripts provided by NIST. The segmentation configuration is again the best available configuration using 100 speaker factors and LDA $100 \rightarrow 50 + \text{WCCN}$ for intra-session variability compensation. To select the best segmented slices and the best segmentation hypothesis for every recording we use the confidence measures and the logistic regression model proposed in Section IV.

TABLE IV
PERFORMANCE OF THE SEGMENTATION SYSTEM WITH HYPOTHESIS GENERATION AND SELECTION FOR THE NIST SRE08 DATASET

Segmentation	DER	$\%DER < \epsilon$	$\xi(\Omega)$
$l = 1$	1.31%	97.51%	85.63%
$l = 2$	1.23%	97.70%	86.43%
$l = 3$	1.45%	97.24%	83.97%
$l = 4$	1.73%	96.29%	81.70%
Max conf hypotheses	**1.00%**	**98.69%**	**88.89%**
Min DER hypotheses	0.70%	99.14%	92.27%

TABLE V
PERFORMANCE OF THE DETECTION TASK AND USEFULNESS OF THE DETECTED SUBSET FROM THE NIST SRE08 WITH HYPOTHESIS GENERATION AND SELECTION

Subset	DER	σ_{DER}	Represent.	ξ
Ω	1.00%	2.98%	100.00%	88.89%
Ω_c	0.79%	1.33%	98.69%	90.41%
Ω_i	21.47%	10.98%	1.31%	-1.52%
Ω'	**0.92%**	**2.38%**	**99.28%**	**89.35%**
$\Omega\backslash\{\Omega'\}$	13.39%	10.98%	0.72%	-0.46%

TABLE VI
PERFORMANCE OF THE SEGMENTATION SYSTEM WITH HYPOTHESIS GENERATION AND SELECTION FOR THE NIST SRE10 DATASET

Segmentation	DER	$\%_{DER<\epsilon}$	$\xi(\Omega)$
$l = 1$	2.45%	95.83%	72.79%
$l = 2$	2.43%	95.88%	73.07%
$l = 3$	2.46%	95.71%	72.64%
$l = 4$	2.73%	96.29%	69.93%
Max conf hypothesis	**2.12%**	96.96%	76.17%
Min DER hypothesis	1.73%	98.11%	80.56%

TABLE VII
PERFORMANCE OF THE DETECTION TASK AND USEFULNESS OF THE DETECTED SUBSET FROM THE NIST SRE10 WITH HYPOTHESIS GENERATION AND SELECTION

Subset	DER	σ_{DER}	Represent.	ξ
Ω	2.12%	3.78%	100.00%	76.17%
Ω_c	1.74%	1.82%	96.96%	78.73%
Ω_i	18.24%	9.97%	3.04%	-2.56%
Ω'	**2.06%**	**3.49%**	**99.49%**	**76.53%**
$\Omega\backslash\{\Omega'\}$	15.84%	14.72%	0.51%	-0.46%

TABLE VIII
PERFORMANCE OF THE DETECTION TASK AND USEFULNESS OF THE DETECTED SUBSET FROM THE NIST SRE08 WITH HYPOTHESIS GENERATION AND SELECTION, ASSUMING THAT $\epsilon = 1\%$

Subset	DER	σ_{DER}	Represent.	ξ
Ω	1.00%	2.98%	100.00%	-11.15%
Ω_c	0.25%	0.27%	73.29%	54.79%
Ω_i	3.34%	5.04%	26.71%	-65.94%
Ω'	**0.38%**	**0.61%**	**68.10%**	**42.52%**
$\Omega\backslash\{\Omega'\}$	2.56%	4.83%	31.90%	-53.67%

TABLE IX
PERFORMANCE OF THE DETECTION TASK AND USEFULNESS OF THE DETECTED SUBSET FROM THE NIST SRE10 WITH HYPOTHESIS GENERATION AND SELECTION, ASSUMING THAT $\epsilon = 1\%$, ESTIMATING THE PRIOR PROBABILITY OF A RECORDING TO BE CORRECTLY SEGMENTED ON THE NIST SRE08 AND ON THE NIST SRE10

Subset	DER	σ_{DER}	Represent.	ξ
Ω	2.12%	3.78%	100.00%	-138.31%
Ω_c	0.46%	0.29%	38.60%	20.66%
Ω_i	3.47%	4.41%	61.40%	-158.97%
Ω'	**1.65%**	**2.42%**	**73.50%**	**-54.16%**
$\Omega\backslash\{\Omega'\}$	4.03%	5.77%	26.50%	-84.15%
$\Omega'_{p_c(SRE10)}$	**0.94%**	**0.76%**	**34.25%**	**1.60%**
$\Omega\backslash\{\Omega'_{p_c(SRE10)}\}$	2.93%	4.44%	65.75%	-139.91%

C. Experiments

Table IV shows the results obtained for every level and for the proposed approach for hypothesis selection on NIST SRE08. As we can see, the performance in terms of DER, the number of correctly segmented recordings and thus $\xi(\Omega)$ using our approach for the selection of correct segmentation hypotheses (Max Conf hypotheses) is better than that obtained at every level. However, the hypothesis selection could be much better if we could always select the best segmentation hypothesis (Min DER hypotheses). To increase the Dataset Usefulness further, we can try to detect the subset Ω' of correctly segmented recordings, i.e., those whose DER is below 10%.

Table V show the results for the detection task on the selected hypotheses for NIST SRE08. This time the detection of correctly segmented recordings is helpful, but not as significantly as in Table I, the ξ only increases 0.46% compared to the increase of 2.68% obtained when considering only $l = 1$ (Table I). This is due to the fact that, after the hypothesis selection, most recordings are reliable and those that are not, are not far from the threshold. However, comparing these results to those obtained at $l = 1$ with no detection, the increase is quite significant, from 85.63% to 89.35%,and even considering detection at $l = 1$ there is a significant increase (from 88.31% to 89.35%).

In order to validate the hypothesis generation and selection strategy, we test it on the NIST SRE10 dataset. Again we can see in Table VI that this approach is useful to reduce the overall DER and thus to increase the reliability and usefulness of the dataset Ω, obtaining an increase of more than 3% in $\xi(\Omega)$. We could go further selecting always the best hypothesis.

Table VII shows that after hypothesis generation and selection, as in NIST SRE08, we do not get great improvement by using the proposed approach to detect the subset Ω' containing correctly segmented recordings, but again, the improvement in ξ is significant compared to that obtained for $l = 1$. Using hypothesis generation and selection and detecting a reliable subset Ω', we obtain a $\xi(\Omega')$ of 76.53% while for $l = 1$ we obtained a $\xi(\Omega)$ of 72.79% for the whole dataset Ω, and 75.39% after detecting the reliable subset Ω'.

As commented in Section V.A, the value of $\epsilon = 10\%$ was selected since it has been shows that it is optimal for speaker verification task [7]. However, the optimal value for ϵ may be different for other applications. To proof that the quality assessment approach is valid for other ϵ values we consider $\epsilon = 1\%$ and we analyze the detection task after hypothesis generation and selection. The value $\epsilon = 1\%$ has shown to be suitable for a speaker clustering task [15]. Results for NIST SRE08 and SRE10 are presented in Tables VIII and IX respectively.

Note that in this case the rate of correctly segmented recordings is significantly smaller. In fact, in both datasets the value of ξ is negative, which means that the whole dataset Ω is not useful for the application due to the large amount of incorrectly segmented recordings and the large DER values obtained for some of those recordings when compared to $\epsilon = 1\%$. Focusing on Table VIII, we can see that the proposed quality assessment approach is working very well: it is retrieving 68.10% of the recordings whose overall DER is as low as 0.38%, obtaining a Dataset Usefulness of $\xi = 42.52\%$ for Ω' compared to the $\xi = -11.15\%$ obtained for Ω. Note the significant increase in ξ and also that a positive value means that Ω' is now useful for the application.

A similar effect can be observed in Table IX. The quality assessment approach is selecting a subset Ω' with lower overall DER, but it selects a large amount of recordings 73.50%, much larger than the rate of recordings that are actually correctly segmented (38.60%). Although in the end an impressive increase in ξ is observed, it is not enough to obtain a positive value, which means that the proposed quality assessment methodology cannot retrieve a useful subset.

The problem of retrieving an amount of recordings larger than desired has to do with the calibration. In fact, for $\epsilon = 1\%$, the rate of correctly segmented recordings for the NIST SRE08 and for the NIST SRE10 datasets differ significantly (73.29% for SRE08 compared to 38.60% for SRE10). The weighted logistic regression is trained on the NIST SRE08, so it tends to accept a high number of recordings. This is a very common problem in detection tasks, and is usually solved assuming that the prior probability of correctly segmented recording is priorly known (for example, it can be extracted from a small dataset which is similar to the evaluation dataset). In fact, if we train the weighted logistic regression using NIST SRE08 data, but considering the NIST SRE10 prior, the results improve significantly, up to a point that the quality assessment approach enables us to retrieve a useful subset ($\Omega'_{p_{\mathrm{SRE10}}}$) with $\xi = 1.60\%$. However, the ξ is far from the maximum value we could achieve ($\xi = 20.66\%$ for Ω_c).

D. Discussion

The proposed strategy for hypothesis generation and selection enables us to improve the performance of speaker diarization on every recording and over the whole dataset, by selecting the most reliable segmentation among several hypotheses. This approach increases the usefulness of the dataset for an application, increasing the reliability of the recordings. However, the detection of reliable segmentation hypotheses is harder when using this strategy, because after selecting the hypothesis obtaining higher confidence measure, the distribution of the confidence measures for correctly and incorrectly segmented recordings are highly overlapped.

Nevertheless, using this approach and detecting those recordings correctly segmented enables us to obtain subsets for both datasets that are more useful than those obtained without hypothesis generation and selection and whose representativeness is very high (99.28% and 99.49% for NIST SRE08 and SRE10 respectively for $\epsilon = 10\%$). This is very interesting, since the process can be semi-supervised, in the sense that the remaining recordings of the dataset can be retrieved by means of manual inspection, since they are not many (16 for NIST SRE08 and 36 for NIST SRE10). This way, the dataset usefulness for each database can increase significantly, since the discarded recordings retrieved by manual inspection are mostly those that were not correctly segmented.

We have shown that the proposed approach for quality assessment for speaker diarization also work for an ϵ value different from $\epsilon = 10\%$. Therefore, it can be considered for any other application, as far as the threshold ϵ is set consequently. For example, the value of $\epsilon = 1\%$, which has been evaluated in this Section, has shown to be suitable for a speaker clustering task in [15].

VII. Conclusion

In this work we have presented a methodology to select and validate a representative and reliable subset from a huge dataset containing two-speaker conversations that need to be segmented for a speaker characterization application. For this purpose, we have proposed a figure of merit (ξ) to measure the usefulness of the dataset when considering a portion of the dataset. This measure takes into account the reliability of the subset considered, which is related to the accuracy of the speaker diarization hypotheses for the recordings in the subset, but also the representativeness of the subset, which is related to the amount of information of the dataset that is kept in the subset. The definitions of reliability and representativeness are not unique, as both depend on the application, so the accuracy of the proposed methodology will depend on how these measures are defined. Therefore, both reliability and representativeness should be carefully defined for every particular application. In general the reliability can be related to a quality or accuracy measure (DER in this work), while the representativeness has to do with the information that it is interesting to keep from the original dataset. Following the proposed methodology, it is interesting to use an accurate speaker diarization system to obtain reliable segmentation hypotheses, but also to detect those correctly segmented recordings and discard the incorrectly segmented ones, that may be destructive for the speaker characterization application.

In order to detect those correctly segmented recordings, we have presented a set of confidence measures along with a weighted logistic regression model that enables to choose the subset that maximizes the proposed figure of merit. We have tested it showing that the figure of merit can increase around 2.5% for the NIST SRE08 and SRE10 datasets, which means that in both cases the subset is representative enough and more reliable than the original dataset. In addition, we have presented results on a speaker verification task, showing that the proposed techniques for quality assessment enables us to discard a set of recordings that obtain much poorer accuracy than the selected ones.

Finally, we have presented a hypothesis generation and selection strategy that uses the confidence measures and logistic regression to obtain reliable initializations for the segmentation system and to select the best segmentation hypothesis among several hypotheses. This makes possible to obtain an even more representative and reliable subset for the speaker characterization application. We have validated this strategy on the NIST SRE10, obtaining a comparable improvement in performance in both development (NIST SRE08) and testing (NIST SRE10) datasets. We have seen that the final performance on

NIST SRE10 is far below that obtained for NIST SRE08. We have detected that, among other causes, this is due to a higher overlapped speech rate in NIST SRE10, which degrades the performance of our speaker diarization system. Therefore, it seems interesting to work towards robust overlapped speech detection, in order to improve the performance of the proposed speaker diarization system.

The proposed method for the assessment of diarization quality can be used in any speaker characterization application, as speaker diarization, speaker verification or speaker clustering, as far as the definition of correctly diarized recordings is adapted to the application (i.e., the threshold is carefully selected).

ACKNOWLEDGMENT

The authors would like to thank Dr. L. Buera for his interesting suggestions to this work.

REFERENCES

[1] NIST Speech Group, Nist Speaker Recognition Evaluation [Online]. Available: http://www.itl.nist.gov/iad/mig/tests/sre/

[2] S. S. Chen and P. S. Gopalakrishnan, "Speaker, environment and channel change detection and clustering via the bayesian information criterion," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP)*, 1998, pp. 127–132.

[3] D. Reynolds and P. Torres-Carrasquillo, "Approaches and applications of audio diarization," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP)*, Philadelphia, PA, Mar. 2005, vol. V, pp. 953–956.

[4] P. Kenny, P. Ouellet, N. Dehak, V. Gupta, and P. Dumouchel, "A study of interspeaker variability in speaker verification," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 16, no. 5, pp. 980–988, Jul. 2008.

[5] F. Castaldo, D. Colibro, E. Dalmasso, P. Laface, and C. Vair, "Stream based speaker segmentation using speaker factors and eigenvoices," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP)*, Las Vegas, NV, Apr. 2008, pp. 4133–4136.

[6] C. Vaquero, A. Ortega, J. Villalba, A. Miguel, and E. Lleida, "Confidence measures for speaker segmentation and their relation to speaker verification," in *Proc. INTERSPEECH, Makuhari*, Chiba, Japan, Sep. 2010, pp. 2310–2313.

[7] D. Reynolds, P. Kenny, and F. Castaldo, "A study of new approaches to speaker diarization," in *Proc. INTERSPEECH*, Brighton, U.K., Sep. 2009, pp. 1047–1050.

[8] P. Kenny, D. Reynolds, and F. Castaldo, "Diarization of telephone conversations using factor analysis," *IEEE J. Sel. Topics Signal Process.*, vol. 4, no. 6, pp. 1059–1070, Dec. 2010.

[9] C. Vaquero, A. Ortega, and E. Lleida, "Intra-session variability compensation and a hypothesis generation and selection strategy for speaker segmentation," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP)*, Prague, Czech Republic, May 2011, pp. 4532–4535.

[10] R. Kuhn, J.-C. Junqua, P. Nguyen, and N. Niedzielski, "Rapid speaker adaptation in eigenvoice space," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 8, no. 6, pp. 695–707, Nov. 2000.

[11] N. Dehak, P. Kenny, R. Dehak, P. Dumouchel, and P. Ouellet, "Front-end factor analysis for speaker verification," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 19, no. 4, pp. 788–798, Aug. 2010.

[12] A. O. Hatch and A. Stolcke, "Generalized linear kernels for one-versus-all classification: Application to speaker recognition," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP)*, 2006, pp. 585–588.

[13] P. Kenny, "Bayesian speaker verification with heavy-tailed priors," in *Proc. Oddyssey Speaker and Language Recognition Workshop*, Brno, Czech Republic, Jun. 2010.

[14] J. Cornfield, T. Gordon, and W. Smith, "Quantal response curves for experimentally uncontrolled variables," *Bull. Int. Statist. Inst.*, no. 38, pp. 97–115, 1961.

[15] C. Vaquero, A. Ortega, and E. Lleida, "Partitioning of two-speaker conversation datasets," in *Proc. INTERSPEECH*, 2011, pp. 385–388.

Carlos Vaquero was born in Zaragoza, Spain, in 1982. He received the M.Sc. degree in telecommunication engineering and the Ph.D. degree in signal processing from the University of Zaragoza (UZ), Zaragoza, Spain, in 2006 and 2011, respectively. From 2006 to 2011, he was with the Communications Technologies Group, Department of Electronic Engineering and Communications, UZ, under a research grant. Currently he is in Agnitio S.L. in Madrid, Spain, as research engineer. His research interest lies in the field of speaker recognition and diarization.

Alfonso Ortega 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 Ph.D. Extraordinary Award and the Teleónica Chair Award to the best technological Ph.D. He has participated in more than 40 research projects funded by national or international public institutions and 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.

Antonio Miguel was born in Zaragoza, Spain, in 1977. He received the M.Sc. degree in telecommunication engineering and the Ph.D. degree from the University of Zaragoza (UZ), Zaragoza, Spain, in 2001 and 2008. From 2000 to 2006, he was with the Communication Technologies Group, Department of Electronic Engineering and Communications, UZ, under a research grant. Since 2006, he has been an Assistant Professor in the same department. Currently, his research interest lies in the field of acoustic modeling.

Eduardo Lleida (M'89) 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 managed more than 50 speech-related projects and 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 dialog systems.