Utilising Tree-Based Ensemble Learning for Speaker Segmentation

Mohamed Abou-Zleikha[1,⋆], Zheng-Hua Tan[1], Mads Græsbøll Christensen[2], and Søren Holdt Jensen[1]

[1] Department of Electronic Systems, Aalborg University, Denmark
[2] Audio Analysis Lab, ad:mt, Aalborg University, Denmark
{moa,zt,shj}@es.aau.dk,mgc@create.aau.dk

Abstract. In audio and speech processing, accurate detection of the changing points between multiple speakers in speech segments is an important stage for several applications such as speaker identification and tracking. Bayesian Information Criteria (BIC)-based approaches are the most traditionally used ones as they proved to be very effective for such task. The main criticism levelled against BIC-based approaches is the use of a penalty parameter in the BIC function. The use of this parameters consequently means that a fine tuning is required for each variation of the acoustic conditions. When tuned for a certain condition, the model becomes biased to the data used for training limiting the model's generalisation ability.

In this paper, we propose a BIC-based tuning-free approach for speaker segmentation through the use of ensemble-based learning. A forest of segmentation trees is constructed in which each tree is trained using a sampled version of the speech segment. During the tree construction process, a set of randomly selected points in the input sequence is examined as potential segmentation points. The point that yields the highest ΔBIC is chosen and the same process is repeated for the resultant left and right segments. The tree is constructed where each node corresponds to the highest ΔBIC with the associated point index. After building the forest and using all trees, the accumulated ΔBIC for each point is calculated and the positions of the local maximums are considered as speaker changing points. The proposed approach is tested on artificially created conversations from the TIMIT database. The approach proposed show very accurate results comparable to those achieved by the-state-of-the-art methods with a 9% (absolute) higher F_1 compared with the standard ΔBIC with optimally tuned penalty parameter.

1 Introduction

Speaker segmentation is the process of determining the speaker switching points in a speech signal leading to an accurate separation of the signal into speaker homogeneous subsegments. This is an essential initial stage in several speech and audio applications such as speaker tracking [5], audio classification [14] and speaker diarization

⋆ This work was supported in part by the Danish Council for Strategic Research of the Danish Agency for Science Technology and Innovation under the CoSound project, case number 11-115328. This publication only reflects the authors' views.

L. Iliadis et al. (Eds.): AIAI 2014, IFIP AICT 436, pp. 50–59, 2014.

systems [3]. Segmentation is usually performed without any prior knowledge about who or how many speakers are present in the speech segment, with an unknown maximum and minimum length of the speakers' segments and with no prior information about the acoustic conditions and the noise level and type. Having no or very limited prior information about these conditions makes speaker segmentation one of the very challenging task in the speech processing domain.

Several approaches have been proposed for speaker segmentation [7,13,4,10,8,15] (an extensive review can be find in [3]). Among these methods, ΔBIC metric is the most widely used for this task [7,13,4,10,8]. The main drawback of BIC-based approaches is that they require fine tuning of a penalty factor that highly affects the quality of segmentation. This penalty factor is related to the number of parameters those are used to model the speech segment. Traditionally, the penalty factor is tuned for each specific acoustic conditions which limits the generalisation capabilities of the system [16,21]. An alternative BIC function has been proposed which aims at eliminating the effect of the number of parameters in the BIC calculation [2]. This is accomplished by estimating the speech segment by two mixtures of the Gaussian mixture model and estimating the left and right segments around the segmentation point by one mixtures of the Gaussian mixture model.

In this paper, we introduce a novel approach based on the ensemble-based mechanism to estimate the speakers changing points in multi-speakers segments. The main advantage of the proposed approach is that it eliminates the need to tune the penalty factor originally employed in the BIC-based segmentation function. Through using the proposed method we were able to increase the robustness of the traditional BIC-based techniques when used in various acoustic conditions.

The approach works by building a set of segmentation trees called segmentation forest; each tree is built using an approach similar to the one proposed in [8]. Each node in each segmentation tree examines a set of randomly selected points as potential segmentation points. The gain function in each tree is the value of the BIC and the goal is to select the point that gives the highest BIC. The value of the penalty factor is randomly assigned for each BIC calculation. Each segmentation tree assigns a BIC value for the examined points. The final resultant BIC values for the sequence points are the accumulated values using all segmentation trees in the forest.

The paper is organised as follows: In section 3, the standard BIC-based speaker changing point detection is presented, followed by a BIC-based speaker segmentation using divide-and-conquer strategy in Section 3. In Section 4, the proposed approach is explained. The experiments and evaluations conducted to validate the proposed approach are presented in Section 5 and finally a conclusion is presented in section 6.

2 Detecting Speaker Changing Points Using BIC

The BIC is a asymptotically optimal Bayesian-based model-selection criterion traditionally used to determine the parametric model that best fits a set of data samples $X = x_1, ..., x_N$, where $x_i \in \mathbb{R}^d$, where d is the dimension of the feature space [7].

According to the BIC approach, the model M_j that best fits a set of data samples is the model that maximises the function:

$$BIC_j = logP(X|M_j) - \lambda\frac{1}{2}k_j logN \tag{1}$$

where $logP(X|M_j)$ is the log likelihood of the data X for M_j, λ is the weight of the second term that relates to the number of parameters k_j in the model M_j and the number of samples N in the data.

Suppose we have two models representing the data X namely M_1 and M_2, the model M_1 is chosen over M_2 to fit the data X if $\Delta BIC = BIC_1 - BIC_2$ is positive.

For the speaker segmentation task, suppose a segment of speech X. To check if the speaker changes at a point index i, M_1 represents the model drawn from two full-covariance Gaussians using the separated segments at point i and M_2 is the model drawn from one full-covariance Gaussian. We calculate the ΔBIC between M_1 and M_2: the number of parameters k_1 in M_1 is twice the number of parameters in M_2, where $k_1 = d + \frac{d(d+1)}{2}$. ΔBIC is calculated as:

$$\Delta BIC = N{*}log|\Sigma| - i{*}log|\Sigma_{left}| - (N-i+1){*}log|\Sigma_{right}| - \frac{1}{2}\lambda(d + \frac{d(d+1)}{2})logN \tag{2}$$

where $|.|$ is the determinant of the covariance matrix, λ is the penalty factor that tunes the segmentation sensitivity [9], ideally this parameter is equal to 1. Point i is considered a speaker switching point, if $\Delta BIC > 0$.

Applying this approach on a speech segment detects a single changing point. In order to be able to detect multiple-switching points, a window growing mechanism is usually employed [20]. This approach starts by processing a small window N_{init} and tries to detect a changing point within this window. It then extends this window by a N_g until a changing point is detected or the window size reaches a maximum size N_{max}. If the search reaches N_{max} with no detection of a changing point, the window is shifted by N_s. Otherwise, the search process is repeated starting from the newly discovered switching point.

3 BIC-Based Speaker Segmentation Using Divide-and-Conquer

This approach uses the ΔBIC function to perform a hierarchical splitting of a speech sequence into its most two dissimilar parts [8]. This is done by scanning the whole segment and choosing the point that gives the highest ΔBIC. The resultant point is then considered as a potential speaker switching point (called i). The same approach is then repeated on the left and right segments of the point i. This process is applied recursively until the size of the segment becomes smaller than a threshold.

After processing the left and right segments, ΔBIC is checked, if it is positive, point i is considered as a switching point; otherwise, the leftmost sub-segment from the right segment and the rightmost sub-segment from the left segment are considered as one segment and the highest ΔBIC is calculated. If ΔBIC is positive, the new point is added to the set of changing points.

One can note that the process described can be seen as generating a tree-like structure, where each internal node in that tree is a segmentation point. It is also worth noticing that this approach still uses the standard ΔBIC where the tuning process of λ is still required.

The approach proposed in this paper employs a similar mechanism for generating a tree structure for the segmentation. These resultant trees are used to detect the changing points. In the next section, the approach proposed for speaker segmentation using a segmentation forest is explained.

4 BIC-Based Speaker Segmentation Using Segmentation Forest

The ensemble-based learning approaches such as random forest and density forest [6] have been successfully employed for several tasks in the audio domain such as emotion recognition [19,18], paralinguistic event detection [1] and audio event detection [11]. In this work, a segmentation forest is utilised as a special case of the random forest approach. Random forest is a tree-based non-parametric classification and regression approach. The principle is to grow an assemble of trees on a random selection of samples in a training set. Each tree is a classification and regression tree (CART). While constructing the trees and at each tree node, a randomly selected set of features is considered and the features are investigated as potential predictors that decide the split of the data. The splitting robustness is calculated as the information gain resulted from the splitting.

The proposed approach applies a similar mechanism to the one used by the random forest. However, instead of building decision trees, we build a set of segmentation trees, where the randomly selected features to be examined at each node are the potential segmentation positions, the gain function in our case is the ΔBIC.

Formally, a segmentation forest SF is a set of segmentation trees

$$SF = \{t_i\} : i = [1...T] \tag{3}$$

where t_i is the i^{th} individual tree and T is the total number of trees. Each tree t_i, in the forest is trained independently in a similar manner to the one used for building the tree in the divide-and-conquer approach explained in Section 3. The main differences between the two processes are:

- Instead of scanning the full set of positions in the sequence, a randomly selected set of points is examined.
- The value of λ in the ΔBIC function is randomly chosen between a min and a max values.
- We do not recheck internal points if the value of ΔBIC is negative, instead we assign the negative value obtained for ΔBIC to the corresponding point.
- Each tree is trained using a sampling from the original sequence without replacement.

A detailed description of the tree building process is presented in Algorithm 1.

The result of the pervious process is a set of segmentation trees where each node in each tree is associated with a position and a ΔBIC value. The final ΔBIC for

Algorithm 1. Building a segmentation tree $n \leftarrow train(X)$

- X : the speech segment
- n : the node that represents the segmentation of the input segment

if ($length(X) < threshold$) **then**
  return empty
end if
- $pp \leftarrow$ get round of $\sqrt{(length(X))}$ randomly chosen points from X as potential segmentation points to examine
- Find the point i from pp that gives the highest ΔBIC (called $mBIC$), where λ is randomly chosen for each point
- Split X at position i into X_l and X_r
- call the same process on X_l to get $n.child_l$
- call the same process on X_r to get $n.child_r$
- store the position i in $n.position$
- - store the value $mBIC$ in $n.\Delta BIC$

return n

each point is the accumulated ΔBIC from all trees. As a result, positive accumulated ΔBIC values are produced on and around the speaker segmentation points. To detect the exact changing point positions, the resultant ΔBIC value sequence is grouped forming a set of local regions according to the distance between the points. The position of the maxima of each local region represents a speaker changing point as illustrated in Figure 1.

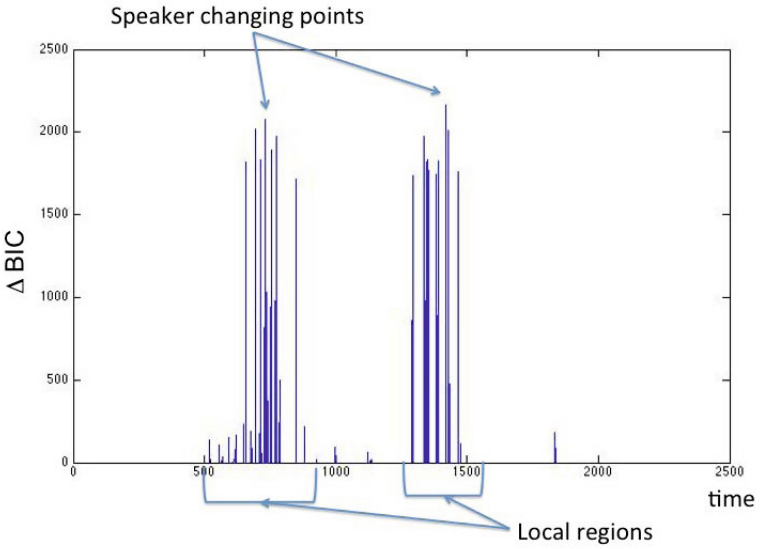

Fig. 1. An illustration of a accumulated ΔBIC and the local regions with their maxima

5 Experiment and Evaluation

The purpose of the experiments conducted is to validate the proposed approach and to check its efficiency for the speaker segmentation task. For this purpose, The performance and execution time of the proposed method is compared with several other BIC-based approaches reported in the literature.

5.1 Experiment Setup

In order to evaluate the proposed approach, a comparison between its performance and several other BIC-based approaches is performed. A artificially created conversations by concatenating speech from the TIMIT database are used for evaluation [12]. TIMIT is an acoustic-phonetic databased which consists of 6300 utterances for 630 english speakers. Two conversation sets (A and B) are generated, the first set is for penalty parameter tuning in the standard and divide-and-conquer BIC-based approaches. It consists of 20 conversation, each contains 2 to 6 speakers and the length of each speaker segment varies between 2 and 6 seconds. The testing dataset, B, consists of 100 conversation. The number of speakers per conversation changes from 2 to 6 and the length of each speaker segment is between 2 and 6 seconds. The testing set contains 338 switching points. The feature vectors used were 23-dimensional mel-frequency cepstral coefficients (MFCC) and log energy extracted every 10 ms, with a window size of 25 ms.

In the process of detecting the changing points in the speaker segmentation modules, two types of errors occur: the first is due to missing a true speaker changing point. This type of errors can be measured using the precision (PRC), which is calculated as:

$$PRC = \frac{number\ of\ correctly\ found\ changes}{total\ number\ of\ changes\ found} \tag{4}$$

and the other measurement is the Missed Detection Rate (MDR), which is calculated as:

$$MDR = 100 * \frac{MD}{RC} \tag{5}$$

where MD is the number of missed changing points and RC is the number of true changing points. The second error type occurs when a false changing point is detected. This type can be measured using the recall (RCL), which is defined as:

$$RCL = \frac{number\ of\ correctly\ found\ changes}{total\ number\ of\ correct\ changes} \tag{6}$$

and the other measurement is False Alarm Rate (FAR), which is calculated as:

$$FAR = 100 * (1 - RCL) \tag{7}$$

Another measurement that combines the PRC and RCL in one value is the F_1-measure. This measurement is defined as:

$$F_1 = 2 * \frac{PRC * RCL}{PRC + RCL} \tag{8}$$

Table 1. Average and standard deviation values for the results obtained from each approach using the five error measurements

	BIC	$DICBIC$	$GMMBIC$	$SF-BIC$
$PRC(mean)$	0.77	0.89	0.59	0.83
$PRC(std)$	0.30	0.27	0.28	0.24
$RCL(mean)$	0.74	0.75	0.75	0.84
$RCL(std)$	0.31	0.29	0.31	0.25
$F_1(mean)$	0.74	0.79	0.65	0.83
$F_1(std)$	0.29	0.27	0.23	0.23
$FAR(mean)$	13.31%	3.97%	50.67%	10.78 %
$FAR(std)$	16.94	11.00	23.88	14.22
$MDR(mean)$	26.03%	25.33%	35.03%	16.47%
$MDR(std)$	30.48	28.82	30.61	24.85

The value of F_1 is between 0 and 1, the closer the value to 1 is, the better the system. For MDR and FAR measurements, the values are between 0 and 100, the smaller the value is, the better the system.

Four approaches are evaluated using the conversation set. The first approach is the standard window-growing approach (referred to as BIC) as described in Section 2. The λ value is tuned using the set A and the result value is set to $\lambda = 2.8$. The second approach is the divide-and-conquer approach (referred to as DACBIC) as described in Section 3 and the λ value is also tuned using the set A and it gives $\lambda = 3.2$. The third one is the window-growing approach using one mixture per GMM for separated segments and two mixture per GMM for combined segments to estimate the covariance matrices [2]. This removes the effect of the penalty parameter (referred to as GMM-BIC). Diagonal covariance GMM is used for GMMBIC implementation. The fourth approach is the proposed approach (referred to as SF-BIC). The number of trees in our approach is set to 50 and the stopping criterion threshold (minimum segment length) is 500. The number of points to examine at each node is $\sqrt{\frac{length(segment)}{N_{min}}}$ where N_{min} is the shifting factor. λ value was generated randomly between $[0..6]$. The tolerance value for the detected points is 0.5 second, i.e. if a detected point is positioned within 0.5 distance around a reference point, it is classified as a correctly detected point.

5.2 Evaluation

The five accuracy measures discussed are calculated for each examined approach. We used the same evaluation protocol proposed in [17]. According to this protocol, the previously discussed error metrics are first calculated followed by applying ANOVA and Tukey test.

Table 1 presents the mean and standard divination values of each measure for each of the examined system.

The results show that a better performance is obtained by the proposed method compared with the other approaches with respect to the overall performance of the system as depicted by the F_1 measure. The results also indicate that our approach is able to more

accurately predict the true changing points according to the RCL and MDR measures. The DICBIC detects less false alarm points compared with the other approaches as the results of PRC and FAR show.

To check if the performance obtained is significantly different among the compared approaches, a One-way ANOVA is applied for a 95% confidence level. The null hypothesis tested is that the groups means are equal, i.e. the approaches are not significantly different. The alternative hypothesis states that the groups means are unequal, which consequently means that at least one of the systems differs from the rest with respect to the examined measure. The $F - statistic$ values and the $p - values$ of all approaches are calculated and presented in Table 2. The results show that the performance is significantly different among all five measures.

Table 2. $F - statistic$ and $p - value$ results obtained from ANOVA for each approach using the five measurements

	$F - statistic$	$p - value$
PRC	66.64	1.42e-34
RCL	6.59	2.36e-04
F_1	44.85	8.45e-25
FAR	142.39	1.22e-61
MDR	6.59	2.36e-04

Since ANOVA test does not provide any information about which system is different from the other, the Tukey range test, or honestly significant differences method, is employed. Tukeys method provides a pair-wise comparison of the means while maintaining the confidence level at a predefined value. If the confidence interval includes zero, the differences are not significant, otherwise, the differences are significant.

The Tukey test is applied between the proposed approach and the three other approaches and the results obtained are presented in Table 3.

Table 3. The Tukey test results between the proposed approach and the other examined approaches

	$SF - BIC$ vs $GMMBIC$	$SF - BIC$ vs $DICBIC$	$SF - BIC$ vs BIC
PRC	0.374,0.448	-0.126,-0.0540	-0.0130 , -0.0635
RCL	0.105 , 0.1856	0.01201, 0.0886	0.01655 , 0.09561
F_1	0.3151 , 0.3805	-0.0401, 0.0302	0.0077 , 0.0820
FAR	-45.477 , -39.888	3.197 , 6.8120	-6.977 , -2.530
MDR	-26.489 , -18.561	-16.512, -8.860	-17.468 , -9.561

The results show that the differences in the performance between all approaches are significant except for the F_1 measure between the proposed approach and the DICBIC approach.

Table 4. Average and standard deviation values for the execution time of each approach

	BIC	$DICBIC$	$GMMBIC$	$SF-BIC$
$time(mean)(s)$	50.16	19.82	91.81	124.77
$time(std)$	25.76	12.26	50.24	54.18

The results obtained from the statistical analysis performed indicate that the performance of our approach is comparable to those achieved by the-state-of-the-art method as demonstrated by the insignificant difference between the accuracies obtained.

The execution time for each approach (shown in Table 4) shows that the proposed approach has the highest execution time. This is due to the time required to build a set of segmentation models instead of one. However, since each tree is built independently, this time consumption issue can be solved by constructing the trees in parallel.

6 Conclusion

In this paper, a tree-based ensemble method for speaker changing point detection is proposed. The approach trains a set of trees using a sampled version of the speech segment. To build a node in the tree, a randomly selected points are examined as potential segmentation points. The point that gives the highest ΔBIC is chosen. This process is recursively applied on the left and right subsegments until a stopping criterion is reached. As a result, a tree is constructed where each node stores the highest ΔBIC with the associated point index. Once the model is built, the accumulated ΔBIC for each point is calculated using the all trees in the forest. The final changing points are then calculated as the positions of the local maximums after grouping the points those have a positive ΔBIC into groups according to the distance between them.

We conduct a set of experiments to test the proposed approach and analyse its performance. For this purpose, a comparison is performed with three other state-of-the-art methods. The comparison shows that the proposed approach achieves better average results an insignificant performance difference from the best models reported in the literature. Our model, however, has the advantage of eliminating the need of parameter tuning and thereafter demonstrates more robustness and generalisation capability over changes in the acoustic conditions.

The future work includes building an interactive speaker changing point detection, where the user can modify, add or delete a set of changing points and the model uses this information to adapt itself.

References

1. Abou-Zleikha, M., , Tan, Z.H., Christensen, M.G., Jensen, S.H.: Non-linguistic vocal event detection and localisation using online random forest. In: Proceedings of 37th International Convention of Information and Communication Technology (MIPRO). IEEE (2014)
2. Ajmera, J., McCowan, I., Bourlard, H.: Robust speaker change detection. IEEE Signal Processing Letters 11(8), 649–651 (2004)

3. Anguera Miro, X., Bozonnet, S., Evans, N., Fredouille, C., Friedland, G., Vinyals, O.: Speaker diarization: A review of recent research. IEEE Transactions on Audio, Speech, and Language Processing 20(2), 356–370 (2012)
4. Ben, M., Betser, M., Bimbot, F., Gravier, G.: Speaker diarization using bottom-up clustering based on a parameter-derived distance between adapted gmms. In: Proceedings of ICSLP (2004)
5. Bonastre, J.F., Delacourt, P., Fredouille, C., Merlin, T., Wellekens, C.: A speaker tracking system based on speaker turn detection for nist evaluation. In: Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing, vol. 2, pp. II1177–II1180. IEEE (2000)
6. Breiman, L.: Random forests. Machine learning 45(1), 5–32 (2001)
7. Chen, S., Gopalakrishnan, P.: Speaker, environment and channel change detection and clustering via the bayesian information criterion. In: Proceedings of DARPA Broadcast News Transcription and Understanding Workshop (1998)
8. Cheng, S.S., Wang, H.M., Fu, H.C.: Bic-based speaker segmentation using divide-and-conquer strategies with application to speaker diarization. IEEE Transactions on Audio, Speech, and Language Processing 18(1), 141–157 (2010)
9. Grašič, M., Kos, M., Kačič, Z.: Online speaker segmentation and clustering using cross-likelihood ratio calculation with reference criterion selection. IET signal processing 4(6), 673–685 (2010)
10. Kotti, M., Benetos, E., Kotropoulos, C.: Automatic speaker change detection with the bayesian information criterion using mpeg-7 features and a fusion scheme. In: IEEE International Symposium on Circuits and Systems, p. 4. IEEE (2006)
11. Kumar, A., Dighe, P., Singh, R., Chaudhuri, S., Raj, B.: Audio event detection from acoustic unit occurrence patterns. In: Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing, pp. 489–492 (2012)
12. Lamel, L.F., Kassel, R.H., Seneff, S.: Speech database development: Design and analysis of the acoustic-phonetic corpus. In: Speech Input/Output Assessment and Speech Databases (1989)
13. Li, R., Schultz, T., Jin, Q.: Improving speaker segmentation via speaker identification and text segmentation. In: Proceedings of INTERSPEECH 2009 (2009)
14. Meinedo, H., Neto, J.: Audio segmentation, classification and clustering in a broadcast news task. In: Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing, vol. 2, pp. II–5. IEEE (2003)
15. Mohammadi, S.H., Sameti, H., Langarani, M.S.E., Tavanaei, A.: Kndist: A non-parametric distance measure for speaker segmentation. In: Proceedings of INTERSPEECH (2012)
16. Mori, K., Nakagawa, S.: Speaker change detection and speaker clustering using vq distortion for broadcast news speech recognition. In: Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing, vol. 1, pp. 413–416. IEEE (2001)
17. Moschou, V., Kotti, M., Benetos, E., Kotropoulos, C.: Systematic comparison of bic-based speaker segmentation systems. In: Proceedings of IEEE 9th Workshop on Multimedia Signal Processing, pp. 66–69. IEEE (2007)
18. Rong, J., Li, G., Chen, Y.P.P.: Acoustic feature selection for automatic emotion recognition from speech. Information processing & management 45(3), 315–328 (2009)
19. Schuller, B., Batliner, A., Seppi, D., Steidl, S., Vogt, T., Wagner, J., Devillers, L., Vidrascu, L., Amir, N., Kessous, L., et al.: The relevance of feature type for the automatic classification of emotional user states: low level descriptors and functionals. In: Proceedings of INTERSPEECH, vol. 2007, pp. 1–4 (2007)
20. Tritschler, A., Gopinath, R.A.: Improved speaker segmentation and segments clustering using the bayesian information criterion. In: Proceedings of Eurospeech, vol. 99, pp. 679–682 (1999)
21. Vandecatseye, A., Martens, J.P., Neto, J.P., Meinedo, H., Garcia-Mateo, C., Dieguez-Tirado, J., Mihelic, F., Zibert, J., Nouza, J., David, P., et al.: The cost278 pan-european broadcast news database. In: Proceedings of LREC (2004)